2.

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)

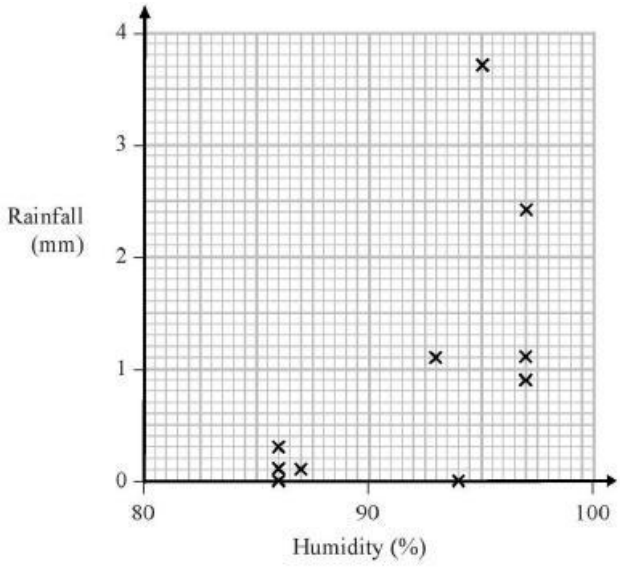
(b) Give a reason why Sara might

(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)

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)

2

Question	Scheme	Marks	AOs
(a)	IQR = 2.3 and 20.6 $\gg$ 2.4 + 1.5×2.3 (= 5.85) (Compare correct values)	B1	1.1b
		(1)	
(b)(i)	e.g. it is a piece of data and we should consider all the data (o.e.)	B1	2.4
(ii)	e.g. it is an extreme value and could unduly influence the analysis or it could be a mistake	B1	2.4
		(2)	
(c)	e.g. "as humidity increases rainfall increases"	B1	2.2b
		(1)	
(d)	e.g. a 10% increase in humidity gives rise to a 1.5 mm increase in rainfall or represents 0.15mm of rainfall per percentage of humidity	B1	3.4
		(1)	
(e)(i)	Not a good method since only uses 11 days from one location in one month.	B1	2.4
(ii)	e.g. She should use data from more of the UK locations and more of the months or using a spreadsheet or computer package she could use all of the available UK data	B1	2.4
		(2)	
		(7 marks)	

Part	Notes
(a)	B1 for sight of the correct calculation and suitable comparison with 20.6
(b)(i)	B1 for a suitable reason for including the data point
(ii)	B1 for a suitable reason for excluding the data point
(c)	B1 for a suitable interpretation of positive correlation mentioning humidity and rainfall
(d)	B1 for a suitable description of the rate: rainfall per percentage of humidity including reference to values.
(e)(i)	B1 for a comment that supports the idea that her sampling method was not a good one
(ii)	B1 for some sensible suggestions that would give a better representation of the data across the UK. Must show some awareness of the fact that LDS has different locations and more months of data available but must be clear they are NOT using any overseas locations. NB B0 for a comment that says use more than one location without specifying that only UK locations are required.

3

A random sample of 15 days is taken from the large data set for Perth in June and July 1987.

The scatter diagram in Figure 1 displays the values of two of the variables for these 15 days.

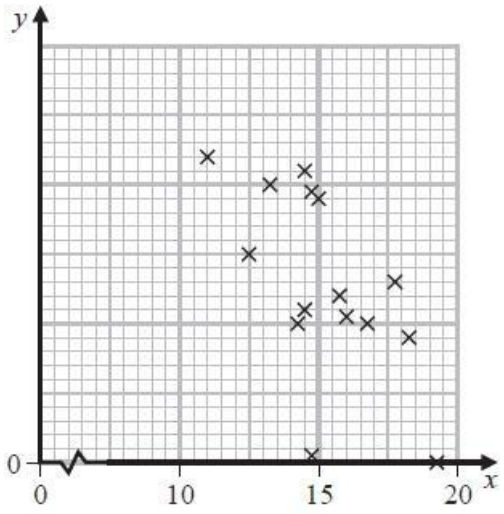


Figure 1

(a) Describe the correlation.

(1)

The variable on the *x*-axis is Daily Mean Temperature measured in °C.

(b) Using your knowledge of the large data set,

(i) suggest which variable is on the *y*-axis,
(ii) state the units that are used in the large data set for this variable.

(2)

Stav believes that there is a correlation between Daily Total Sunshine and Daily Maximum Relative Humidity at Heathrow.

He calculates the product moment correlation coefficient between these two variables for a random sample of 30 days and obtains r = −0.377

(c) Carry out a suitable test to investigate Stav's belief at a 5% level of significance.

State clearly

- your hypotheses
- your critical value

(3)

On a random day at Heathrow the Daily Maximum Relative Humidity was 97%

(d) Comment on the number of hours of sunshine you would expect on that day,

giving a reason for your answer.

(1)

3

	Scheme	Marks	AO
(a)	Negative	B1 **(1)**	1.2
(b)(i)	Rainfall / mm | or | Pressure / hPa or Pascals or hectopascals or mb or millibars	B1	2.2b
(ii)		B1ft **(2)**	1.1b
(c)	$H_0 : \rho = 0$ $H_1 : \rho \neq 0$	B1	2.5
	Critical value: $-0.361(0)$	M1	1.1b
	$r < -0.3610$ so significant result and there is evidence of a correlation between Daily Total Sunshine and Daily Maximum Relative Humidity	A1 **(3)**	2.2b
(d)	Humidity is high and there is evidence of correlation and $r < 0$ So expect amount of sunshine to be lower than the average for Heathrow(oe)	B1 **(1)**	2.2b
		(7 marks)	

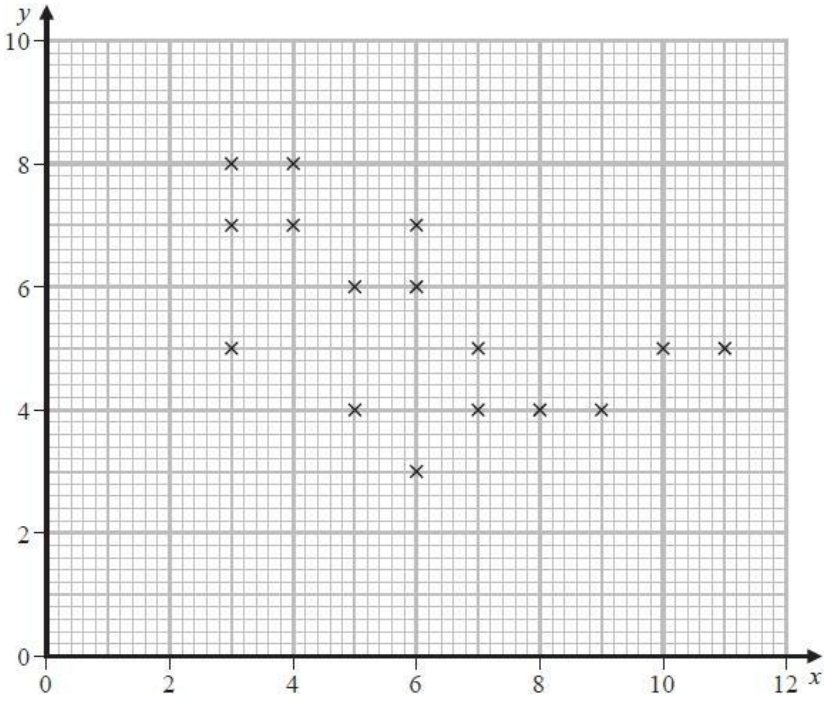
	Notes
(a)	B1 for stating negative. "Negative skew" is B0 though
(b)(i)	B1 for mentioning "rainfall" (allow "rain" or "precipitation") or "pressure" (if more than 1 answer both must be correct) NB the other quantitative variable for Perth is: Daily Mean Wind Speed and scores B0 [Not allowed "wind speed" since $r = +0.15$ and in winter might expect wind to raise temp]
(ii)	B1ft for giving the correct units. If Daily Mean Wind Speed (kn) or knots "Wind speed" and "knots" would score B0B1 but any other variable scores B0B0
(c)	B1 for both hypotheses correct in terms of ρ M1 for the correct critical value compatible with their H_1: allow $\pm 0.361(0)$ If the hypotheses are 1-tail then allow cv of ± 0.3061 e.g. Alternative hypothesis with $r < \pm 0.377$ implies a one-tail test or H_0 and H_1 in words saying "H_0: there is no correlation, H_1: there is correlation" is two-tail If there are no hypotheses (or they are nonsensical) assume 2-tail so M1 for $\pm 0.361(0)$ A1 for a correct conclusion in context based on comparing -0.377 with their cv. Condone incorrect inequality e.g. $-0.3610 < -0.377$ as long as they reject H_0 Do not accept contradictory statements such as "accept H_0 so there is evidence of …" Can say "support for Stav's belief"(o.e.e.g. "claim") or "evidence of a correlation between sunshine and humidity" condone "negative correlation" or comments such as "if humidity is high amount of sunshine will be low"
(d)	B1 for stating low amount of sunshine (o. e.) **and** some reference to $r < 0$ or fog Check for the following 2 features: (i) **low** sunshine: allow $\leqslant 5$ hrs (LDS mean for 2015 is 5.3, humidity 97% is 4.1, $\geqslant 97\%$ is 3.1) (ii) **negative** correlation may be described in words e.g. "high humidity gives low sunshine" or **fog** (LDS says >95% humidity is foggy) so less sunshine

4

Marc took a random sample of 16 students from a school and for each student recorded

- the number of letters, x, in their last name
- the number of letters, y, in their first name

His results are shown in the scatter diagram below.



(a) Describe the correlation between x and y.

(1)

Marc suggests that parents with long last names tend to give their children shorter first names.

(b) Using the scatter diagram comment on Marc's suggestion, giving a reason for your answer.

(1)

The results from Marc's random sample of 16 observations are given in the table below.

x	3	6	8	7	5	3	11	3	4	5	4	9	7	10	6	6
y	7	7	4	4	6	8	5	5	8	4	7	4	5	5	6	3

(c) Use your calculator to find the product moment correlation coefficient between x and y for these data.

(1)

(d) Test whether or not there is evidence of a negative correlation between the number of letters in the last name and the number of letters in the first name.

You should

- state your hypotheses clearly
- use a 5% level of significance

4

	Scheme	Marks	AO
(a)	Negative	B1 (1)	1.2
(b)	Marc's suggestion is compatible because it's negative correlation	B1 (1)	2.4
(c)	$(r =)$ $-0.54458266\ldots$ awrt -0.545	B1 (1)	1.1b
(d)	$H_0 : \rho = 0$ $H_1 : \rho < 0$	B1	2.5
	[5% 1-tail cv =] (±) 0.4259	M1	1.1a
	(significant result / reject H_0) There is evidence of negative correlation between the number of letters in (or length of) a student's last name and their first name	A1 (3)	2.2b
		(6 marks)	

	Notes
(a)	B1 for "negative" Allow "slight" or "weak" etc Allow a description e.g. "as x increases y decreases" or in context e.g. "people with longer last names tend to have shorter first names" A comment of "negative skew" is B0 **Need to see distinct or separate responses for (a) and (b)**
(b)	B1 for a comment that suggests data is compatible with the suggestion **and** a suitable reason such as "there is negative correlation" or a description in x and y or in context or the points lie close to a line with negative gradient or draw line $y = x$ and state that more points below the line so supports (or is compatible with) his suggestion A reason based on just a **single point** is B0 e.g. " 11 letters in last name has only 5 in first name"
(c)	B1 for awrt -0.545
(d)	B1 for both hypotheses correct in terms of ρ M1 for a critical value compatible with their H_1: 1-tail: awrt ± 0.426 (condone ± 0.425) **or** 2-tail (B0 scored for H_1) : awrt ± 0.497 If hypotheses are in words and can deduce whether one or two-tail then use their words. If no hypotheses or their H_1 is not clearly one or two tail assume one-tail A1 for compatible signs between cv and r **and** a correct conclusion in context mentioning correlation and number of letters or length and name (ft their value from (c)) Do NOT award this A mark if contradictory comments or working seen e.g. "accept H_0" or comparison of 0.426 with significance level of 0.05 etc
NB	The M1A1 can be scored independently of the hypotheses

5

Barbara is investigating the relationship between average income (GDP per capita), x US dollars, and average annual carbon dioxide (CO_2) emissions, y tonnes, for different countries.

She takes a random sample of 24 countries and finds the product moment correlation coefficient between average annual CO_2 emissions and average income to be 0.446

(a) Stating your hypotheses clearly, test, at the 5% level of significance, whether or not the product moment correlation coefficient for all countries is greater than zero.

(3)

Barbara believes that a non-linear model would be a better fit to the data.
She codes the data using the coding $m = \log_{10} x$ and $c = \log_{10} y$ and obtains the model $c = -1.82 + 0.89m$

The product moment correlation coefficient between c and m is found to be 0.882

(b) Explain how this value supports Barbara's belief.

(1)

(c) Show that the relationship between y and x can be written in the form $y = ax^n$ where a and n are constants to be found.

(5)

5

Part	Working or answer an examiner might expect to see	Mark	Notes
(a)	$H_0 : \rho = 0$ $H_1 : \rho > 0$	B1	This mark is given for both hypotheses in terms of ρ found correctly
	For sample size 24 at the 5% level of significance, the critical value = 0.3438	M1	This mark is given for selecting a suitable critical value compatible with H_1
	$0.446 > 0.3438$, so reject H_0 There is evidence that the product moment correlation coefficient (pmcc) is greater than 0	A1	This mark is given for a correct conclusion stated
(b)	The value of the pmcc is close to 1 so there is a strong positive correlation	B1	This mark is given for a correct explanation about the strength of the correlation
(c)	$\log_{10} y = -1.82 + 0.89 \log_{10} x$	M1	This mark is given for a correct substitution of both c and m
	$y = 10^{-1.82 + 0.89 \log x}$	M1	This mark is given for dealing with logs to find an expression in terms of y
	$y = 10^{-1.82} \times 10^{0.89 \log x}$ $y = 10^{-1.82} \times 10^{(\log x)^{0.89}}$	M1	This mark is given for a method to find values for a and n
	$y = 0.015 \times x^{0.89}$	A1	This mark is given for find a correct value of $a = 0.015$
		A1	This mark is given for find a correct value of $n = 0.89$
			(Total 9 marks)

6

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.

6

Question	Scheme	Marks	AOs
(a)	Label **each** year group	B1	1.1b
	Use random numbers to select a ...	B1	1.1b
	Simple random sample of 24 Year 12s and 16 Year 13s.	B1	1.1b
		(3)	
(b)	Increase by 2.8 marks	B1	3.4
		(1)	
(c)	e.g. 'the best performance is predicted for the students who never wake up'	B1	3.5b
		(1)	
			(5 marks)

Notes	
(a)	**B1:** for a suitable numbered/labelled/ordered(o.e.) list/database/register(o.e.) for **each** year group. Condone poor numbering but if just one list, then the Year 12s must be distinguishable from the Year 13s
	B1: for use of random numbers/sample/selection to choose students
	B1: for 24 Year 12s, and 16 Year 13s
Note:	A description of a systematic sample: only allow access to the first mark and therefore may score maximum B1B0B0
(b)	**B1:** Using the gradient of the regression equation must include increase(o.e.) and 2.8 'Increase by approximately 3 marks' is B0 but isw if 2.8 is seen $5.6 \div 2$ is not sufficient
(c)	**B1:** for any suitable limitation of the model e.g. the idea that the longer you sleep the better performance in the test or only valid between 0 and 24 hours (within range of the data) or only applicable to the amount of sleep the night before the test or only takes sleep into consideration/does not include other variables (factors) or cannot score below 26.1 marks on the test or the model might not be linear over the entire range or the model might predict more than the maximum mark B0: e.g. might not be correlation between s and p or individual student performance may vary

7

Tessa owns a small clothes shop in a seaside town. She records the weekly sales figures, £ w, and the average weekly temperature, t °C, for 8 weeks during the summer.
The product moment correlation coefficient for these data is −0.915

(a) Stating your hypotheses clearly and using a 5% level of significance, test whether or not the correlation between sales figures and average weekly temperature is negative.

(3)

(b) Suggest a possible reason for this correlation.

(1)

Tessa suggests that a linear regression model could be used to model these data.

(c) State, giving a reason, whether or not the correlation coefficient is consistent with Tessa's suggestion.

(1)

(d) State, giving a reason, which variable would be the explanatory variable.

(1)

Tessa calculated the linear regression equation as $w = 10\,755 - 171t$

(e) Give an interpretation of the gradient of this regression equation.

(1)

7

Qu	Scheme	Marks	AO
(a)	$H_0 : \rho = 0$ $H_1 : \rho < 0$	B1	2.5
	Critical value: -0.6215 (Allow any cv in range $0.5 < \lvert cv \rvert < 0.75$)	M1	1.1a
	$r < -0.6215$ so significant result and there is evidence of a negative correlation between w and t	A1	2.2b
		(3)	
(b)	e.g. As temperature increases people spend more time on the beach and less time shopping (o.e.)	B1	2.4
		(1)	
(c)	Since r is close to -1, it is consistent with the suggestion	B1	2.4
		(1)	
(d)	t will be the explanatory variable since sales are likely to depend on the temperature	B1	2.4
		(1)	
(e)	Every degree rise in temperature leads to a drop in weekly earnings of £171	B1	3.4
		(1)	
		(7 marks)	

Notes

(a)
B1 for both hypotheses in terms of ρ
M1 for the critical value: sight of ± 0.6215 or any cv such that $0.5 < \lvert cv \rvert < 0.75$
A1 must reject H_0 on basis of comparing -0.915 with -0.6215 (if $-0.915 < 0.6215$ is seen then A0 but may use $\lvert r \rvert$ o.e. which is fine)
and mention "negative", "correlation/relationship" and at least "w" and "t"

(b)
B1 for a suitable reason to explain negative correlation using the context given.
e.g. "As temperature drops people are more likely to go shopping (than to the beach)"
e.g. "As temperature increases people will be outside rather than in shops"
A mere description in context of negative correlation is B0
SO e.g. "As temperature increases people don't want to go shopping/buy clothes" is B0
e.g. "Less clothes needed as temp increases" is B0

(c)
B1 for a suitable reason e.g. "strong"/"significant"/"near perfect" "correlation", $\lvert r \rvert$ close to 1 and saying it is consistent with the suggestion. Allow "yes" followed by the reason.

(d)
B1 For identifying t and giving a suitable reason.
Need idea that "w depends on t" or "w responds to t" or "t affects w" (o.e.)
Allow t (temperature) affects the other variable etc
Just saying "t is the independent variable" or "t explains change in w" is B0
N. B. Suggesting causation is B0 e.g. "t causes w to decrease"

(e)
B1 for a description that conveys the idea of rate per degree Celsius.
Must have 171, condone missing "£" sign.

8

A meteorologist believes that there is a relationship between the daily mean windspeed, w kn, and the daily mean temperature, t °C. A random sample of 9 consecutive days is taken from past records from a town in the UK in July and the relevant data is given in the table below.

t	13.3	16.2	15.7	16.6	16.3	16.4	19.3	17.1	13.2
w	7	11	8	11	13	8	15	10	11

The meteorologist calculated the product moment correlation coefficient for the 9 days and obtained $r = 0.609$

(a) Explain why a linear regression model based on these data is unreliable on a day when the mean temperature is 24 °C

(1)

(b) State what is measured by the product moment correlation coefficient.

(1)

(c) Stating your hypotheses clearly test, at the 5% significance level, whether or not the product moment correlation coefficient for the population is greater than zero.

(3)

Using the same 9 days a location from the large data set gave $\overline{t} = 27.2$ and $\overline{w} = 3.5$

(d) Using your knowledge of the large data set, suggest, giving your reason, the location that gave rise to these statistics.

(1)

(Total for question = 6 marks)

8

Question	Scheme	Marks	AOs
(a)	e.g. It requires extrapolation so will be unreliable (o.e.)	B1	1.2
		(1)	
(b)	e.g. Linear association between w and t	B1	1.2
		(1)	
(c)	H_0: $\rho = 0$ H_1: $\rho > 0$	B1	2.5
	Critical value 0.5822	M1	1.1a
	Reject H_0		
	There is evidence that the product moment correlation coefficient is greater than 0	A1	2.2b
		(3)	
(d)	Higher $\bar{t}$ suggests overseas and not Perth…lower wind speed so perhaps not close to the sea so suggest **Beijing**	B1	2.4
		(1)	
			(6 marks)

Notes:

(a)
B1: for a correct statement (unreliable) with a suitable reason

(b)
B1: for a correct statement

(c)
B1: for both hypotheses in terms of ρ
M1: for selecting a suitable 5% critical value compatible with their H_1
A1: for a correct conclusion stated

(d)
B1: for suggesting Beijing with some supporting reason based on t or w
Allow Jacksonville with a reason based just on higher $\bar{t}$