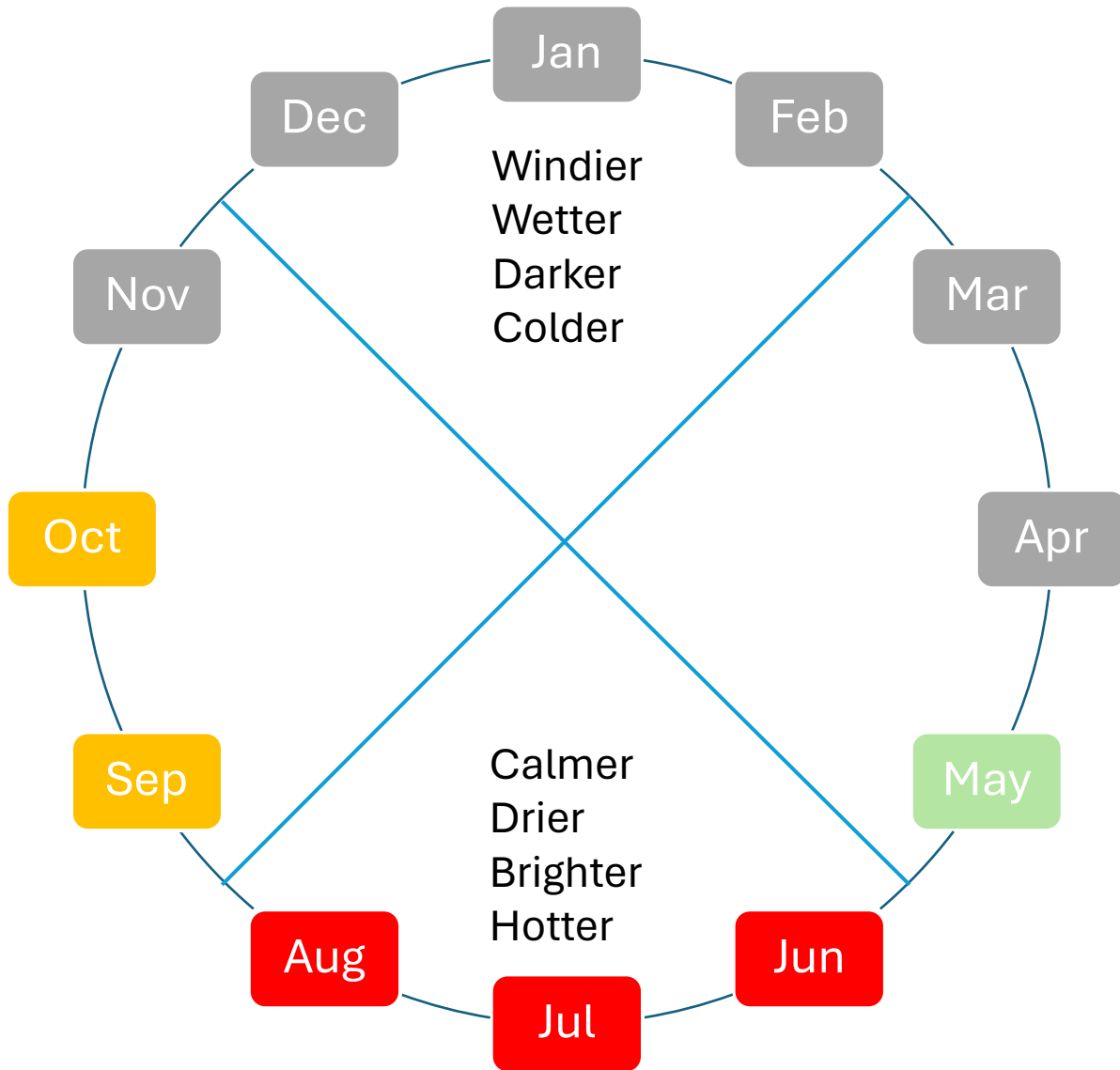


The Large Data Set (Edexcel)

- In AS it has accounted for 6.7% of all the stats marks, and only 1.25% of the overall AS marks
- In A2 it has accounted for a total of 5.3% of all the stats marks, and only 0.9% of the overall A2 marks
- I'm going to give you the best chance of being successful with the questions!



We only have data for May to October.

We have data from 1987 and 2015, so can compare the same month in two different years.



From **South** to **North**:

(alphabetical order except H and H switch)

Camborne (coastal – windier)

Hurn

Heathrow

Leeming

Leuchars (coastal – windier)

Common sense box:

If the temperature increases...

... the amount of sunshine _____

... the amount of rainfall _____

... the windspeed _____

As we move further north, during May to October...

... the temperature _____

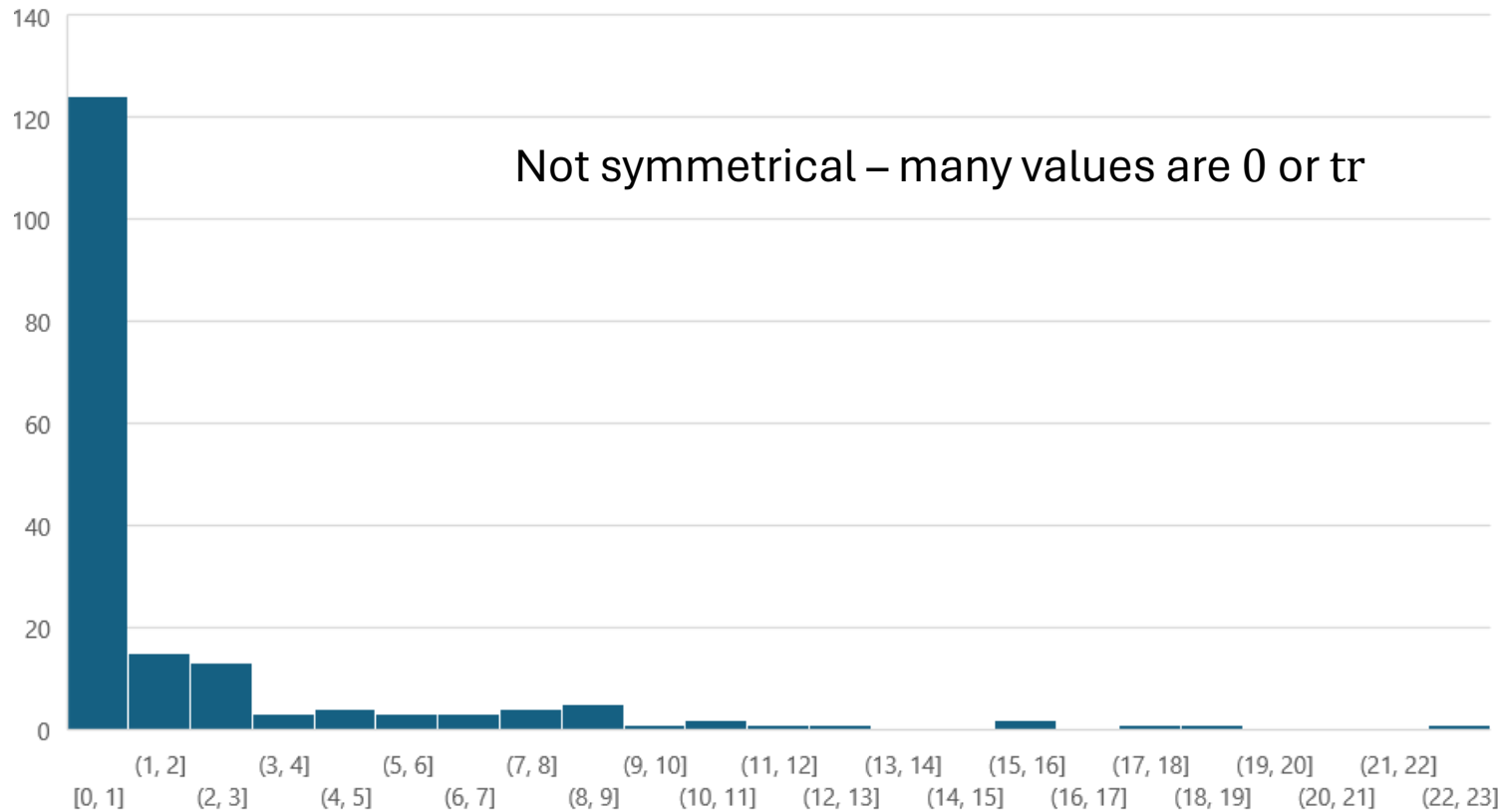
... the amount of rainfall _____

... the maximum hours of sunshine _____

UK Measurements pt. 1

Measurement		Units	Typical range	Examples	Details
Daily mean temperature	<i>How hot it is</i>	°C	~5°C to ~24°C	12.0°C 14.8°C	Warmer in summer
Daily total rainfall	<i>How much it rained</i>	mm	0 to ~20mm	0 mm 10.4 mm tr	Not symmetrical – many 0 and tr values
Daily total sunshine	<i>How many hours of sunshine</i>	hours	0 to ~14 hours	3.3 hrs 10.4 hrs	More sunshine in the summer
Cloud cover	<i>How much of the sky is covered in clouds</i>	oktas	0 to 8	3 4 0	Integers, measuring what fraction of the sky is covered
Humidity	<i>How much water vapour is in the air – above 95% is associated with fog</i>	%	~70% to 100%	100% 77% 95%	Integers
Daily mean visibility	<i>How far you can see</i>	Dm 1 Dm = 10 m 'decametre'	~200 Dm to ~4000 Dm	1300 Dm 2200 Dm 3100 Dm	Rounded to nearest 100
Daily mean pressure	<i>How much the atmosphere is pushing down</i>	hPa 'hectopascals'	~990 hPa to ~1040 hPa	1017 hPa 1006 hPa 997 hPa	Integers

Rainfall



‘Cleaning’ data

‘tr’ means there was a trace of water in the measuring instrument – it is used for values of rainfall, $0 < r \leq 0.05$
We can ‘clean’ data, which means replacing ‘tr’ with either 0 mm or 0.025 mm.

If we have n/a in any of our data, we cannot use it. Cleaning it means to remove that entry.

UK Measurements pt. 2 – wind

Measurement		Units	Typical range	Examples	Details
Daily mean windspeed	<i>How windy it is</i>	kn (knots)	~3 kn to ~10 kn	4 kn 11 kn	Integers only
Windspeed, Beaufort conversion		Qualitative	Light to Moderate	Light Moderate Fresh Strong	Qualitative. Most days are ‘light’
Daily maximum gust	<i>The strongest gust of wind that day</i>	kn (knots)	~8 kn to ~50 kn	17 kn 25 kn	Integers only
Wind/gust direction (bearings)	<i>Which direction the wind is blowing from</i>	°	10° to 360°	240° 70°	Multiples of 10 only
Wind/gust direction (cardinal)	<i>Which direction the wind is blowing from</i>	Compass direction	—	N SW ENE	Describes where the wind is blowing from , not to



A world map with a light blue background and green landmasses. Three red location pins are placed on the map: one in Jacksonville, Florida, USA; one in Beijing, China; and one in Perth, Australia. Each pin is associated with a text box containing climate information. The map also shows various country borders and names, as well as major bodies of water.

Jacksonville, Florida, USA

Hot summers

Hurricanes in October (but mean
windspeed still relatively low in our data)

Beijing, China

Hotter, wetter summers
Colder winters

Perth, Australia

Flipped seasons

International Measurements

Measurement		Units	Details
Daily mean temperature	How hot it is	°C	Warmer in summer... but Perth is colder
Daily total rainfall	How much it rained	mm	Beijing is rainy in the summer
Daily mean pressure	How much the atmosphere is pushing down	hPa ‘hectopascals’	
Daily mean windspeed	<i>How windy it is</i>	kn (knots)	Now are rounded to 1 dp
Windspeed, Beaufort conversion		Light to Moderate	Qualitative. Most days are ‘light’

AS 2020

2. 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.

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)

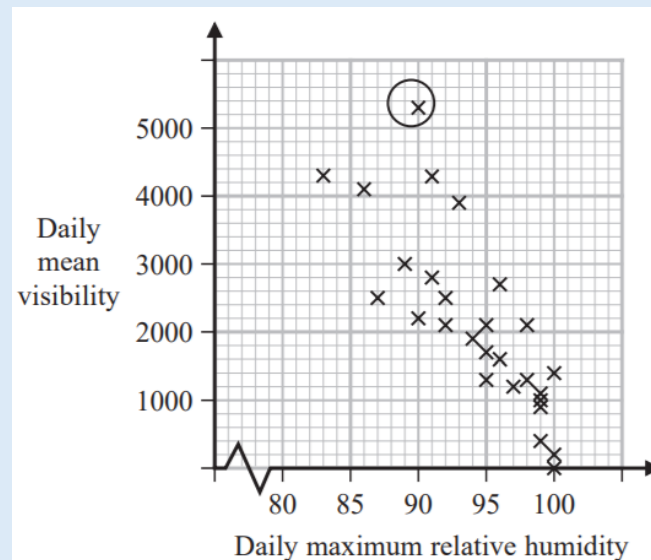


Figure 2

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

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.

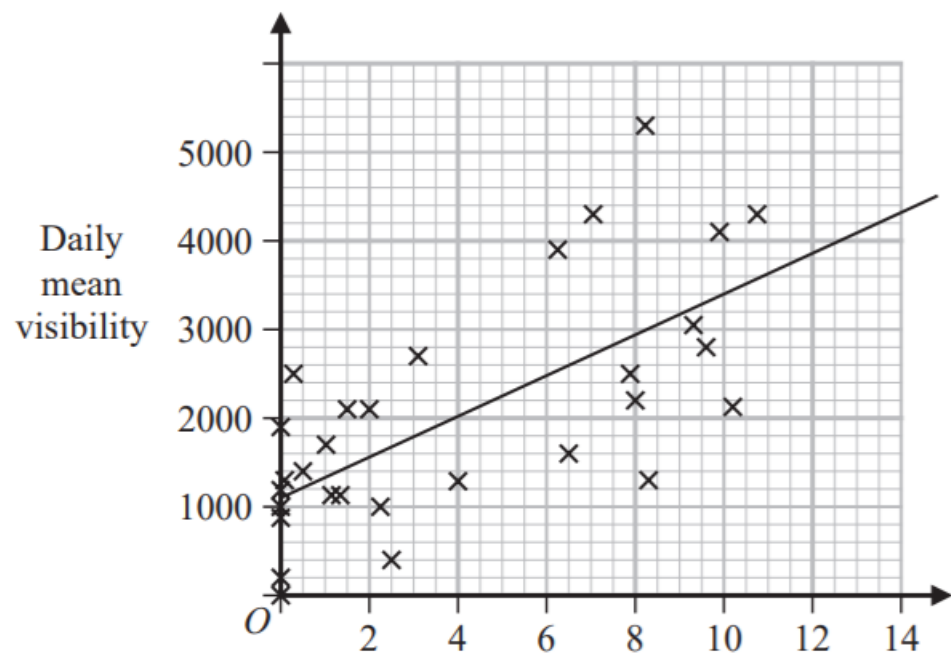


Figure 3

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

(1)

Question	Scheme	Marks	AOs
2(a)	0 to 500 m	B1	1.2
		(1)	
(b)	$1100 + 1600 + 1.5 \times 1600$ [= 5100]	M1	2.1
	5300 > 5100 therefore outlier	A1	1.1b
		(2)	
(c)	As the humidity increases the mean visibility decreases	B1	2.4
		(1)	
(d)	(Hours of) sunshine	B1	2.2b
		(1)	

(5 marks)

Notes			
(a)	B1:	For realising it is the maximum distance and distance given with correct units. Allow 0 to 50dm or < 500m or < 50dm	
(b)	M1:	Attempt to find Q_3 and the upper limit	
	A1:	5100, if a value for the point is stated it must be above 5100 otherwise it is A0. For a statement comparing and conclusion it is an outlier or it is above $Q_3 + 1.5IQR$. Allow accept the point circled is greater than 5100 oe	
(c)	B1:	For a suitable interpretation of a negative correlation mentioning humidity and visibility	
(d)	B1:	A correct deduction that the unlabelled variable is the hours of sunshine. Condone missing hours. Do not allow if more than one variable given. Must be quantative variable Not cloud cover since values bigger than 8 Not wind speed since values not integers Not daily mean temperature since mean temperature near to zero are unlikely in June	

3. Stav is studying the large data set for September 2015

He codes the variable Daily Mean Pressure, x , using the formula $y = x - 1010$

The data for all 30 days from Hurn are summarised by

$$\sum y = 214 \quad \sum y^2 = 5912$$

- (a) State the units of the variable x (1)
- (b) Find the mean Daily Mean Pressure for these 30 days. (2)
- (c) Find the standard deviation of Daily Mean Pressure for these 30 days. (3)

Stav knows that, in the UK, winds circulate

- in a **clockwise** direction around a region of **high** pressure
- in an **anticlockwise** direction around a region of **low** pressure

The table gives the Daily Mean Pressure for 3 locations from the large data set on 26/09/2015

Location	Heathrow	Hurn	Leuchars
Daily Mean Pressure	1029	1028	1028
Cardinal Wind Direction			

The Cardinal Wind Directions for these 3 locations on 26/09/2015 were, in random order,

W NE E

You may assume that these 3 locations were under a single region of pressure.

- (d) Using your knowledge of the large data set, place each of these Cardinal Wind Directions in the correct location in the table.
Give a reason for your answer. (2)

Qu 3	Scheme	Marks	AO
(a)	Hectopascal <u>or</u> hPa	B1	1.2
(b)	$\bar{x} = \bar{y} + 1010$ <u>or</u> $\frac{214}{30} + 1010$ = 1017.1333... awrt 1017	M1 A1	1.1b 1.1b
(c)	$\sigma_x = \sigma_y$ (or statement that standard deviation is not affected by this type of coding) $[\sigma_y =] \sqrt{\frac{5912}{30} - ("7.13[33...]")^2}$ <u>or</u> $\sqrt{146.1822...}$ = 12.0905... awrt 12.1	M1 A1	3.1b 1.1b 1.1b
(d)	High pressure (since approx. mean + sd) so clockwise Locations are (from North to South): Leuchars, Heathrow, Hurn Wind direction is direction wind blows <u>from</u> So: Heathrow (NE) Hurn (E) Leuchars (W)	B1 B1	2.4 2.2a
		(2)	
		(8 marks)	
	Notes		
FYI	1 hPa = 100 Pa; 10hPa = 1 kPa; 1Pa = 1 Nm ⁻²		
(a)	B1 for “hectopascal” <u>or</u> hPa (condone pascals, allow millibars <u>or</u> mb) o.e. Do NOT allow kPa <u>or</u> kilopascals <u>or</u> Pa on its own		
(b)	M1 for a strategy to find \bar{x} Allow an attempt to find $\sum x$ that gets as far as $\sum x = \sum y + 30 \times 1010 [= 30\,514]$ A1 for awrt 1017 (accept 1020) [Ignore incorrect units]		
(c)	1 st M1 for an overall strategy using the fact $\sigma_x = \sigma_y$ (can be implied by correct <u>final</u> ans) <u>or</u> for $\sum x = 30\,514$ and $\sum x^2 = 31\,041\,192$ (both seen and correct) 2 nd M1 for a correct expression (with $\sqrt{}$)(ft their \bar{y} to 3sf) allow awrt 146 for 146.1822.. <u>or</u> for correct expression in x can ft their $\sum x > 30\,000$ or their answer to (b) A1 (dep on 2 nd M1) for awrt 12.1 [Ignore incorrect units]		
Final answer	Final ans of awrt 12.1 scores 3/3 but if they then adjust for x e.g. add 1010 (M0M1A1)		
(d)	1 st B1 for at least one of these reasons (these 2 lines) clearly stated (may see diagram) Need “high pressure” and “clockwise” to score on 1 st line Contradictory statements B0 e.g. correct N~S list but say “anticlockwise” 2 nd B1 (indep of 1 st B1) for deducing the 3 correct directions either in the table or stated as above If the answers in table and text are different we take the table (as question says)		

AS Large Data Set Questions

AS 2018

4. Helen is studying the daily mean wind speed for Camborne using the large data set from 1987. The data for one month are summarised in Table 1 below.

Windspeed	n/a	6	7	8	9	11	12	13	14	16
Frequency	13	2	3	2	2	3	1	2	1	2

Table 1

- (a) Calculate the mean for these data. (1)
- (b) Calculate the standard deviation for these data and state the units. (2)

The means and standard deviations of the daily mean wind speed for the other months from the large data set for Camborne in 1987 are given in Table 2 below. The data are not in month order.

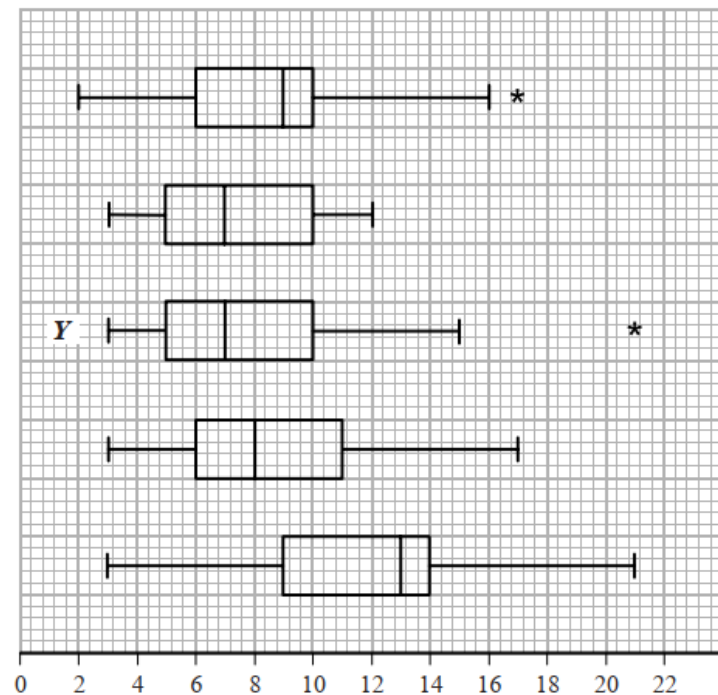
Month	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
Mean	7.58	8.26	8.57	8.57	11.57
Standard Deviation	2.93	3.89	3.46	3.87	4.64

Table 2

- (c) Using your knowledge of the large data set, suggest, giving a reason, which month had a mean of 11.57 (2)

The data for these months are summarised in the box plots on the opposite page. They are not in month order or the same order as in Table 2.

- (d) (i) State the meaning of the * symbol on some of the box plots.
- (ii) Suggest, giving your reasons, which of the months in Table 2 is most likely to be summarised in the box plot marked *Y*. (3)



Qu	Scheme	Marks	AO		
4 (a)	$\bar{x} = 10.2$ (2222...) awrt 10.2	B1	1.1b		
(b)	$\sigma_x = 3.17$ (20227...) awrt 3.17	(1) B1ft	1.1b		
	Sight of "knots" or "kn" (condone knots/s etc)	B1	1.2		
(c)	October since it is windier in the autumn or month of the hurricane or latest month in the year	(2) B1	2.2b		
		B1	2.4		
(d)(i)	They represent <u>outliers</u>	(2) B1	1.2		
(ii)	Y has low median so expect lowish mean (but outlier so > 7) <u>and</u> Y has big range/IQR or spread so expect larger st.dev Suggests B	M1	2.4		
		A1	2.2b		
		(3)			
		(8 marks)			
	Notes				
NB	$\bar{x} = \frac{184}{18}$ and $\sigma_x = \sqrt{\frac{2062}{18} - \bar{x}^2}$				
(a)	B1 for $\bar{x} = 10.2$ (allow exact fraction)	[This is an LDS mark]			
(b)	1 st B1ft allow 3.2 from a correct expr' accept $s = 3.26$ (3984...) [ft use of n/a] <u>Treating n/a as 0</u> May see $n = 31$ or $\bar{x} = 5.9354...$ which is B0 in (a) but here in (b) it gives $\sigma_x = 5.59$ (34...) or $s = 5.6858...$ (awrt 5.69) and scores 1 st B1 2 nd B1 accept kn accept in (a) or (b) (allow nautical miles/hour) [This is an LDS mark]				
(c)	1 st B1 choosing October but accept September. [This is an LDS mark] 2 nd B1 for stating that (Camborne) is windier in autumn/winter months "because it is winter/autumn/windier/colder in "month" " Sep \leq "month" \leq Mar scores B1B1 for "month" = Sep or Oct and B0B1 for other months in range				
(d)(i)	B1 for outlier or the idea of an extreme value allow "anomaly"				
(ii)	M1 for a comment relating to location that mentions both median and mean <u>and</u> a comment relating to <u>spread</u> that mentions both range/IQR and standard deviation and leads to choosing B , C or D Choosing A or E is M0 Incorrect/false statements score M0 e.g. $Q_3 = (\text{mean} + \sigma)$ or identify $Q_2 = \text{mean}$ or Y has small spread				
ALT	Use of outliers: outlier is $(\text{mean} + 3\sigma)$ ($B = 19.9$), ($C = 18.95$), ($D = 20.2$) Must <u>see</u> at least one of these values and compare to Y 's outlier[leads to D or B]				
	A1 for suitable inference i.e. B (accept D or B or D) M1 must be scored				

AS 2019

4. Joshua is investigating the daily total rainfall in Hurn for May to October 2015

Using the information from the large data set, Joshua wishes to calculate the mean of the daily total rainfall in Hurn for May to October 2015

- (a) Using your knowledge of the large data set, explain why Joshua needs to clean the data before calculating the mean.

(1)

Using the information from the large data set, he produces the grouped frequency table below.

- (b) Use linear interpolation to calculate an estimate for the upper quartile of the daily total rainfall.

(2)

- (c) Calculate an estimate for the standard deviation of the daily total rainfall in Hurn for May to October 2015

(2)

- (d) (i) State the assumption involved with using class midpoints to calculate an estimate of a mean from a grouped frequency table.

- (ii) Using your knowledge of the large data set, explain why this assumption does not hold in this case.

- (iii) State, giving a reason, whether you would expect the actual mean daily total rainfall in Hurn for May to October 2015 to be larger than, smaller than or the same as an estimate based on the grouped frequency table.

(3)

Daily total rainfall (r mm)	Frequency	Midpoint (x mm)
$0 \leq r < 0.5$	121	0.25
$0.5 \leq r < 1.0$	10	0.75
$1.0 \leq r < 5.0$	24	3.0
$5.0 \leq r < 10.0$	12	7.5
$10.0 \leq r < 30.0$	17	20.0

You may use $\sum fx = 539.75$ and $\sum fx^2 = 7704.1875$

Part	Working or answer an examiner might expect to see	Mark	Notes
(a)	Trace data needs to be converted to numbers before the calculation can be carried out	B1	This mark is given for a valid explanation
(b)	$(1 +) \frac{138 - 131}{24} \times 4$	M1	This mark is given for a methods to find an estimate of the upper quartile
	$= 2.17$	A1	This mark is given for a correct estimate of the upper quartile
(c)	$\sigma = \sqrt{\frac{7704.1875}{184} - \left(\frac{539.75}{184}\right)^2}$	M1	This mark is given for using the formula for standard deviation to find an estimate for the standard deviation of the total daily rainfall
	$= 5.77$	A1	This mark is given for a correct estimate for the standard deviation of the total daily rainfall
(d)(i)	Using class midpoints to estimate the mean assumes that the values are uniformly distributed in each class	B1	This mark is given for an explanation that the data is assumed to be spread evenly across each class
(d)(ii)	The assumption does not hold since the majority of the data in the first class are 0	B1	This mark is given for a valid explanation by the assumption does not hold
(d)(iii)	The actual mean is likely to be smaller than the estimate; the first group has more values at 0 and close to 0	B1	This mark is given for a correct inference based on knowledge of the Large Data Set

AS 2020

2. 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.

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)

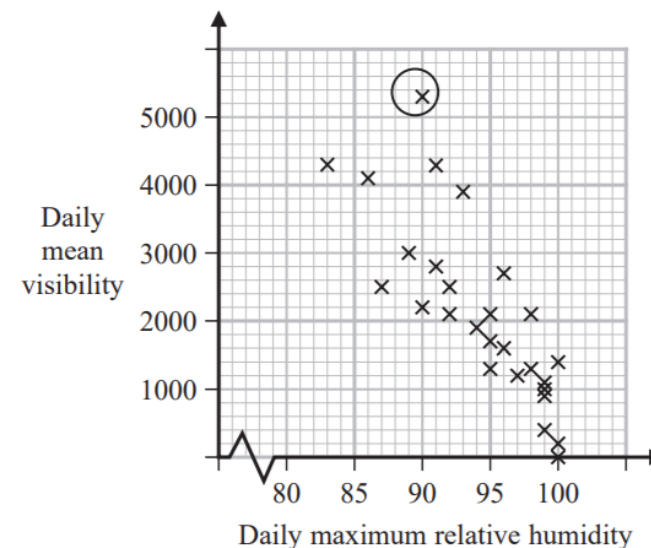


Figure 2

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

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.

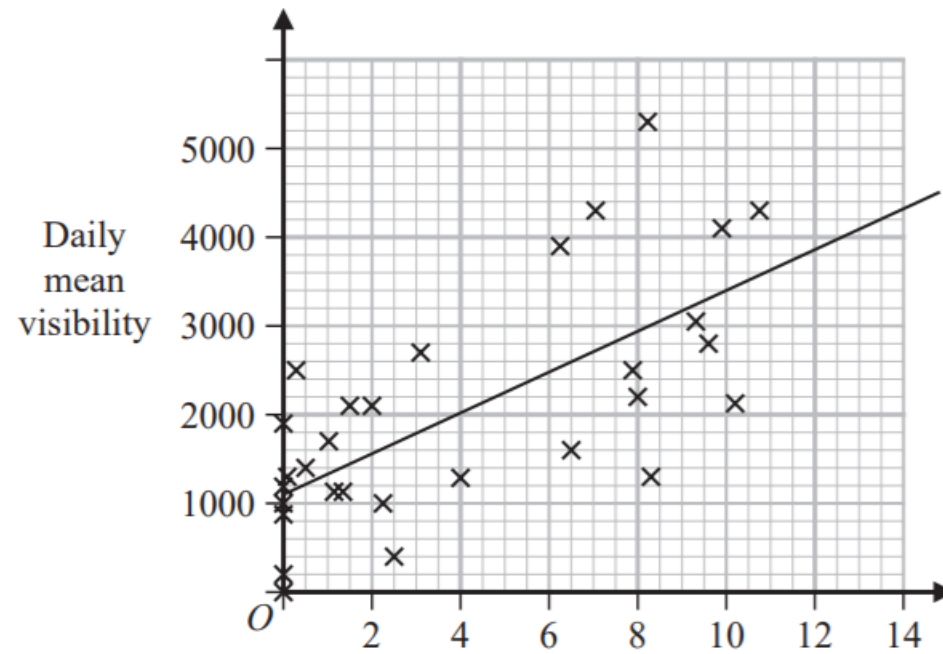


Figure 3

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

(1)

Question		Scheme	Marks	AOs
2(a)		0 to 500 m	B1	1.2
			(1)	
(b)		$1100 + 1600 + 1.5 \times 1600 [= 5100]$	M1	2.1
		$5300 > 5100$ therefore outlier	A1	1.1b
			(2)	
(c)		As the humidity increases the mean visibility decreases	B1	2.4
			(1)	
(d)		(Hours of) sunshine	B1	2.2b
			(1)	
(5 marks)				
Notes				
(a)	B1:	For realising it is the maximum distance and distance given with correct units. Allow 0 to 50dm or < 500m or < 50dm		
(b)	M1:	Attempt to find Q_3 and the upper limit		
	A1:	5100, if a value for the point is stated it must be above 5100 otherwise it is A0. For a statement comparing and conclusion it is an outlier or it is above $Q_3 + 1.5IQR$. Allow accept the point circled is greater than 5100 oe		
(c)	B1:	For a suitable interpretation of a negative correlation mentioning humidity and visibility		
(d)	B1:	A correct deduction that the unlabelled variable is the hours of sunshine. Condone missing hours. Do not allow if more than one variable given.		
		Must be quantative variable		
		Not cloud cover since values bigger than 8		
		Not wind speed since values not integers		
		Not daily mean temperature since mean temperature near to zero are unlikely in June		

AS 2021

3. Helen is studying one of the qualitative variables from the large data set for Heathrow from 2015.

She started with the data from 3rd May and then took every 10th reading.

There were only 3 different outcomes with the following frequencies

Outcome	<i>A</i>	<i>B</i>	<i>C</i>
Frequency	16	2	1

- (a) State the sampling technique Helen used.

(1)

- (b) From your knowledge of the large data set

- (i) suggest which variable was being studied,
(ii) state the name of outcome *A*.

(2)

George is also studying the same variable from the large data set for Heathrow from 2015. He started with the data from 5th May and then took every 10th reading and obtained the following

Outcome	<i>A</i>	<i>B</i>	<i>C</i>
Frequency	16	1	1

Helen and George decided they should examine all of the data for this variable for Heathrow from 2015 and obtained the following

Outcome	<i>A</i>	<i>B</i>	<i>C</i>
Frequency	155	26	3

- (c) State what inference Helen and George could reliably make from their original samples about the outcomes of this variable at Heathrow, for the period covered by the large data set in 2015.

(1)

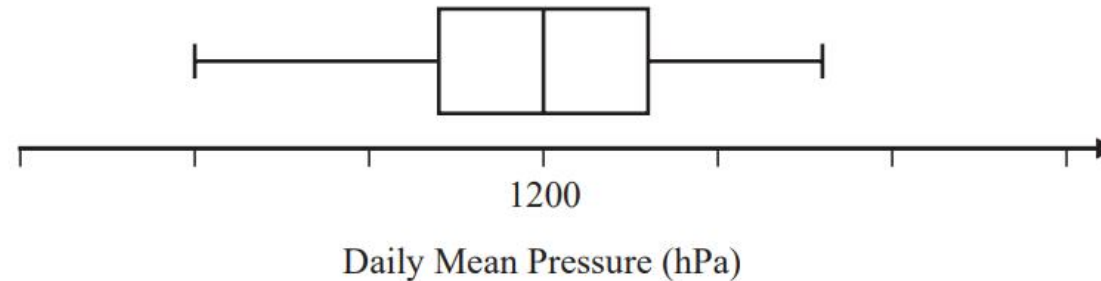
Qu	Scheme	Marks	AO
3. (a)	Systematic (sampling)	B1 (1)	1.2
(b)(i)	[Daily Mean] Wind Speed	B1	2.2a
(ii)	Light	B1 (2)	1.2
(c)	Variable A occurs most (around 80~90%) of the time	B1 (1) (4 marks)	2.2b
Notes			
(a)	B1 for identifying the correct sampling technique Allow slight misspelling e.g. “sysmatic”, “sytmatic” Do NOT allow “systemic”		
(b)(i)	B1 for identifying appropriate qualitative variable. {LDS mark} Allow “Wind speed” or “Wind strength” but NOT just “wind” or “wind direction”		
(ii)	B1 for realising that modal wind speed is “Light” {LDS mark} Allow just “light” or “most light”		
NB	These two B marks are independent so can score B0B1 for e.g. “rainfall” and “light”		
(c)	B1 for inferring that frequency of A can be estimated fairly reliably: {underestimates B and over estimates C } e.g. “ A is the most frequent” [can then ignore comments about B and C]		

AS 2022

4. Jiang is studying the variable Daily Mean Pressure from the large data set.

He drew the following box and whisker plot for these data for one of the months for one location using a linear scale but

- he failed to label all the values on the scale
- he gave an incorrect value for the median



Using your knowledge of the large data set, suggest a suitable value for

(a) the median,

(1)

(b) the range.

(1)

(You are not expected to have memorised values from the large data set. The question is simply looking for sensible answers.)

4. (a)	Accept 990 to 1030 inclusive	B1 (1)	1.1b
(b)	Any range between 10 and 50 inclusive	B1 (1)	1.1b
		(2 marks)	
Notes			
(a)	B1 (Median pressures usually around 1000~1020)	[LDS mark]	
(b)	B1 Any answer in this range Allow answers in the form $a \sim b$ where $ b - a $ is between 10 and 50 Also allow the case where <u>both</u> a and b are in $[10, 50]$	[LDS mark]	

A2 Large Data Set Questions

A2 2018

1. Helen believes that the random variable C , representing cloud cover from the large data set, can be modelled by a discrete uniform distribution.

(a) Write down the probability distribution for C .

(2)

(b) Using this model, find the probability that cloud cover is less than 50%

(1)

Helen used all the data from the large data set for Hurn in 2015 and found that the proportion of days with cloud cover of less than 50% was 0.315

(c) Comment on the suitability of Helen's model in the light of this information.

(1)

(d) Suggest an appropriate refinement to Helen's model.

(1)

A2 2019

2.

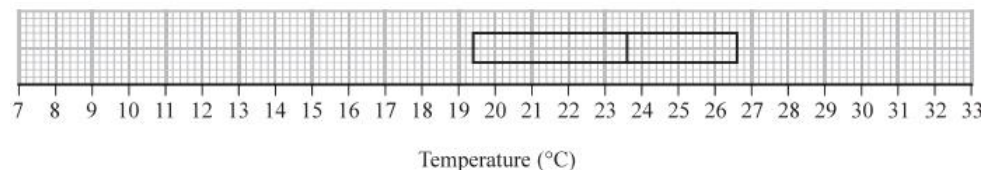


Figure 1

The partially completed box plot in Figure 1 shows the distribution of daily mean air temperatures using the data from the large data set for Beijing in 2015

An outlier is defined as a value
more than $1.5 \times \text{IQR}$ below Q_1 or
more than $1.5 \times \text{IQR}$ above Q_3

The three lowest air temperatures in the data set are 7.6°C , 8.1°C and 9.1°C
The highest air temperature in the data set is 32.5°C

(a) Complete the box plot in Figure 1 showing clearly any outliers.

(4)

(b) Using your knowledge of the large data set, suggest from which month the two outliers are likely to have come.

(1)

Using the data from the large data set, Simon produced the following summary statistics for the daily mean air temperature, $x^\circ\text{C}$, for Beijing in 2015

$$n = 184 \quad \sum x = 4153.6 \quad S_{xx} = 4952.906$$

(c) Show that, to 3 significant figures, the standard deviation is 5.19°C

(1)

Simon decides to model the air temperatures with the random variable

$$T \sim N(22.6, 5.19^2)$$

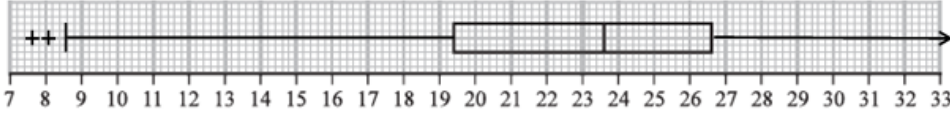
(d) Using Simon's model, calculate the 10th to 90th interpercentile range.

(3)

Simon wants to model another variable from the large data set for Beijing using a normal distribution.

(e) State two variables from the large data set for Beijing that are **not** suitable to be modelled by a normal distribution. Give a reason for each answer.

(2)

(a)	$IQR = 2.6 - 19.4 = 7.2$	B1	This mark is given for finding the interquartile range
	$19.4 - (1.5 \times 7.2) = 8.6$ $19.4 + (1.5 \times 7.2) = 37.4$	M1	This mark is given for a method find the values for the whiskers of the boxplot
			
		A1	This mark is given for plotting the correct whisker (8.6) on the boxplot
		A1	This mark is given for plotting the two correct outliers 7.6 °C and 8.1 °C
(b)	October (since it is the month with the coldest temperatures between May and October in Beijing)	B1	This mark is given for a correct suggestion with a supporting reason.
(c)	$\sigma = \sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{4952.906}{184}} = \sqrt{26.92} = 5.19$	B1	This mark is given for showing the calculation for the standard deviation to three significant figures
(d)	$z = (\pm) 1.2816$	B1	This mark is given for identifying the z-value for the 10th and 90th percentiles (from tables or calculator)
	$2 \times z \times 5.19$	M1	This mark is given for a method to find the interpercentile range between the 10th and 90th value
	$= 13.303$	A1	This mark is given for finding a correct interpercentile range between the 10th and 90th value
(e)	Daily wind speed (Beaufort) since it is qualitative data	B1	This mark si given for stating a correct variable with a supporting reason
	Rainfall (since it is not symmetric)	B1	This mark si given for stating a correct variable with a supporting reason

A2 2019

4. Magali is studying the mean total cloud cover, in oktas, for Leuchars in 1987 using data from the large data set. The daily mean total cloud cover for all 184 days from the large data set is summarised in the table below.

Daily mean total cloud cover (oktas)	0	1	2	3	4	5	6	7	8
Frequency (number of days)	0	1	4	7	10	30	52	52	28

One of the 184 days is selected at random.

- (a) Find the probability that it has a daily mean total cloud cover of 6 or greater.

(1)

Magali is investigating whether the daily mean total cloud cover can be modelled using a binomial distribution.

She uses the random variable X to denote the daily mean total cloud cover and believes that $X \sim B(8, 0.76)$

Using Magali's model,

- (b) (i) find $P(X \geq 6)$

(2)

- (ii) find, to 1 decimal place, the expected number of days in a sample of 184 days with a daily mean total cloud cover of 7

(2)

- (c) Explain whether or not your answers to part (b) support the use of Magali's model.

(1)

There were 28 days that had a daily mean total cloud cover of 8

For these 28 days the daily mean total cloud cover for the **following** day is shown in the table below.

Daily mean total cloud cover (oktas)	0	1	2	3	4	5	6	7	8
Frequency (number of days)	0	0	1	1	2	1	5	9	9

- (d) Find the proportion of these days when the daily mean total cloud cover was 6 or greater.

(1)

- (e) Comment on Magali's model in light of your answer to part (d).

(2)

Knowledge of the LDS is helpful,
but not essential for this question

Part	Working or answer an examiner might expect to see	Mark	Notes
(a)	$\frac{523 + 52 + 28}{184} = \frac{132}{184} = 0.717$	B1	This mark is given for a correct value for the probability for the cloud cover
(b)(i)	$P(X \geq 6) = 1 - P(X \leq 5)$	M1	This mark is given for using $1 - P(X \leq 5)$ with $B(8, 0.76)$
	$= 1 - 0.2967$ $= 0.703$	A1	This mark is given for finding as correct value for the probability
(b)(ii)	$184 \times P(X = 7)$ $= 184 \times 0.2811$	M1	This mark is given for using $184 \times P(X = 7)$ with $B(8, 0.76)$
	$= 51.7$	A1	This mark is given for finding as correct value for the probability
(c)	The answer to part (b)(i) of 0.703 is similar to 0.7127 in part (a) The answer to part (b)(ii) of 51.7 is very close to 52 found in the data set	B1	This mark is given for a correct evaluation of the outcomes from part (b) to determine the appropriateness of <u>Magali's</u> model
(d)	$\frac{5 + 9 + 9}{28} = \frac{23}{28} = 0.821$	B1	This mark is given for a correct value for the probability for the cloud cover
(e)	The answer to part (d) of 0.821 is greater than that in part (a) of 0.717 This shows that there is a higher chance of having high cloud cover if the previous day had high cloud cover	B1	This mark is given for a correct comparison for the answer to part (d) with the data set
	Thus independence does not hold so a binomial model might not be suitable	B1	This mark is given for a correct conclusion stated

A2 2020

2. 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.

- (a) Describe the correlation.

The variable on the x -axis is Daily Mean Temperature measured in $^{\circ}\text{C}$.

- (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.

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

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.

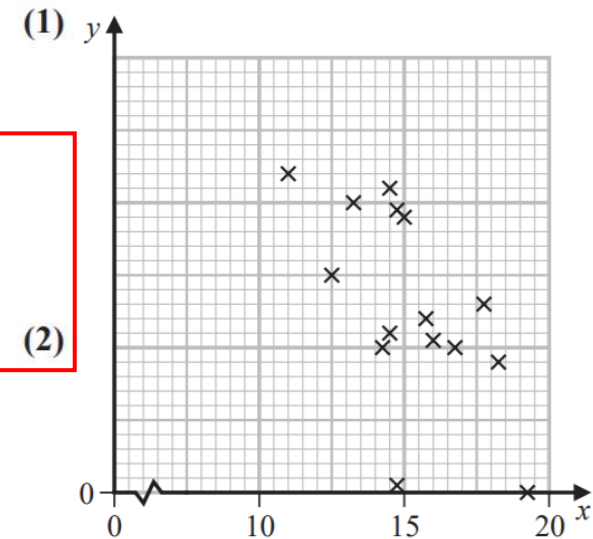


Figure 1

(3)

(1)

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

	Notes
(a)	B1 for stating negative. “Negative skew” is B0 though
(b)(i)	<p>B1 for mentioning “rainfall” (allow “rain” <u>or</u> “precipitation”) <u>or</u> “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]</p>
(ii)	<p>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</p>
(c)	<p>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 <u>or</u> 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 $\pm 0.361(0)$</p> <p>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 <u>belief</u>”(o.e.e.g. “claim”) or “evidence of a correlation between <u>sunshine</u> and <u>humidity</u>” condone “negative correlation” or comments such as “if humidity is high amount of sunshine will be low”</p>
(d)	<p>B1 for stating <u>low</u> amount of sunshine (o. e.) and some reference to $r < 0$ or fog Check for the following 2 features:</p> <p>(i) low sunshine: allow ≤ 5 hrs (LDS mean for 2015 is 5.3, humidity 97% is 4.1, $\geq 97\%$ is 3.1) (ii) negative correlation may be described in words e.g. “high humidity gives low sunshine” <u>or</u> fog (LDS says $>95\%$ humidity is foggy) so less sunshine</p>

A2 2021

3. Stav is studying the large data set for September 2015

He codes the variable Daily Mean Pressure, x , using the formula $y = x - 1010$

The data for all 30 days from Hurn are summarised by

$$\sum y = 214 \quad \sum y^2 = 5912$$

(a) State the units of the variable x

(1)

(b) Find the mean Daily Mean Pressure for these 30 days.

(2)

(c) Find the standard deviation of Daily Mean Pressure for these 30 days.

(3)

Stav knows that, in the UK, winds circulate

- in a **clockwise** direction around a region of **high** pressure
- in an **anticlockwise** direction around a region of **low** pressure

The table gives the Daily Mean Pressure for 3 locations from the large data set on 26/09/2015

Location	Heathrow	Hurn	Leuchars
Daily Mean Pressure	1029	1028	1028
Cardinal Wind Direction			

The Cardinal Wind Directions for these 3 locations on 26/09/2015 were, in random order,

W NE E

You may assume that these 3 locations were under a single region of pressure.

(d) Using your knowledge of the large data set, place each of these Cardinal Wind Directions in the correct location in the table.
Give a reason for your answer.

(2)

Qu 3	Scheme	Marks	AO
(a)	Hectopascal <u>or</u> hPa	B1	1.2
(b)	$\bar{x} = \bar{y} + 1010$ <u>or</u> $\frac{214}{30} + 1010$ = 1017.1333... awrt 1017	M1 A1	1.1b 1.1b
(c)	$\sigma_x = \sigma_y$ (or statement that standard deviation is not affected by this type of coding) $[\sigma_y =] \sqrt{\frac{5912}{30} - ("7.13[33...]")^2}$ <u>or</u> $\sqrt{146.1822...}$ = 12.0905... awrt 12.1	M1 A1	3.1b 1.1b 1.1b
(d)	High pressure (since approx. mean + sd) so clockwise Locations are (from North to South): Leuchars, Heathrow, Hurn Wind direction is direction wind blows <u>from</u> So: Heathrow (NE) Hurn (E) Leuchars (W)	B1 B1	 2.4 2.2a
		(2)	
		(8 marks)	
	Notes		
FYI	1 hPa = 100 Pa; 10hPa = 1 kPa; 1Pa = 1 Nm ⁻²		
(a)	B1 for "hectopascal" <u>or</u> hPa (condone pascals, allow millibars <u>or</u> mb) o.e. Do NOT allow kPa <u>or</u> kilopascals <u>or</u> Pa on its own		
(b)	M1 for a strategy to find \bar{x} Allow an attempt to find $\sum x$ that gets as far as $\sum x = \sum y + 30 \times 1010 [= 30\,514]$ A1 for awrt 1017 (accept 1020) [Ignore incorrect units]		
(c)	1 st M1 for an overall strategy using the fact $\sigma_x = \sigma_y$ (can be implied by correct <u>final</u> ans) <u>or</u> for $\sum x = 30\,514$ and $\sum x^2 = 31\,041\,192$ (both seen and correct) 2 nd M1 for a correct expression (with $\sqrt{}$)(ft their \bar{y} to 3sf) allow awrt 146 for 146.1822.. <u>or</u> for correct expression in x can ft their $\sum x > 30\,000$ or their answer to (b) A1 (dep on 2 nd M1) for awrt 12.1 [Ignore incorrect units]		
Final answer	Final ans of awrt 12.1 scores 3/3 but if they then adjust for x e.g. add 1010 (M0M1A1)		
(d)	1 st B1 for at least one of these reasons (these 2 lines) clearly stated (may see diagram) Need "high pressure" and "clockwise" to score on 1 st line Contradictory statements B0 e.g. correct N~S list but say "anticlockwise" 2 nd B1 (indep of 1 st B1) for deducing the 3 correct directions either in the table or stated as above If the answers in table and text are different we take the table (as question says)		

A2 2022

3. Dian uses the large data set to investigate the Daily Total Rainfall, r mm, for Camborne.

(a) Write down how a value of $0 < r \leq 0.05$ is recorded in the large data set.

(1)

Dian uses the data for the 31 days of August 2015 for Camborne and calculates the following statistics

$$n = 31 \qquad \sum r = 174.9 \qquad \sum r^2 = 3523.283$$

(b) Use these statistics to calculate

- (i) the mean of the Daily Total Rainfall in Camborne for August 2015,
- (ii) the standard deviation of the Daily Total Rainfall in Camborne for August 2015.

(3)

Dian believes that the mean Daily Total Rainfall in August is less in the South of the UK than in the North of the UK.

The mean Daily Total Rainfall in Leuchars for August 2015 is 1.72 mm to 2 decimal places.

(c) State, giving a reason, whether this provides evidence to support Dian's belief.

(2)

Dian uses the large data set to estimate the proportion of days with no rain in Camborne for 1987 to be 0.27 to 2 decimal places.

(d) Explain why the distribution $B(14, 0.27)$ might **not** be a reasonable model for the number of days without rain for a 14-day summer event.

(1)

Question	Scheme	Marks	AOs
3(a)	tr	B1	1.2
		(1)	
(b)(i)	$\mu = \frac{174.9}{31} = 5.6419\dots$ awrt 5.64	B1	1.1b
(ii)	$\sigma_r = \sqrt{\frac{3523.283}{31} - \mu^2}$	M1	1.1b
	$= 9.04559\dots$ awrt 9.05	A1	1.1b
		(3)	
(c)	Leuchars is in the North and Camborne is in the South	M1	2.4
	The mean is smaller for Leuchars than Camborne therefore there is no evidence that Dian's belief is true	A1ft	2.2b
		(2)	
(d)	eg $p = 0.27$ is unlikely to be constant.	B1	2.4
		(1)	
(7 marks)			