Pearson Edexcel

Additional Assessment Materials

Summer 2021

Pearson Edexcel GCE in Mathematics

9MA0 (Applied) (Public release version)

Resource Set 1: Topic 2

Data Presentation and Interpretation

Additional Assessment Materials, Summer 2021

General guidance to Additional Assessment Materials for use in 2021

Context

- Additional Assessment Materials are being produced for GCSE, AS and A levels (with the exception of Art and Design).
- The Additional Assessment Materials presented in this booklet are an optional part of the range of evidence teachers may use when deciding on a candidate’s grade.
- 2021 Additional Assessment Materials have been drawn from previous examination materials, namely past papers.
- Additional Assessment Materials have come from past papers both published (those materials available publicly) and unpublished (those currently under padlock to our centres) presented in a different format to allow teachers to adapt them for use with candidate.

Purpose

- The purpose of this resource to provide qualification-specific sets/groups of questions covering the knowledge, skills and understanding relevant to this Pearson qualification.
- This document should be used in conjunction with the mapping guidance which will map content and/or skills covered within each set of questions.
- These materials are only intended to support the summer 2021 series.

1 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.

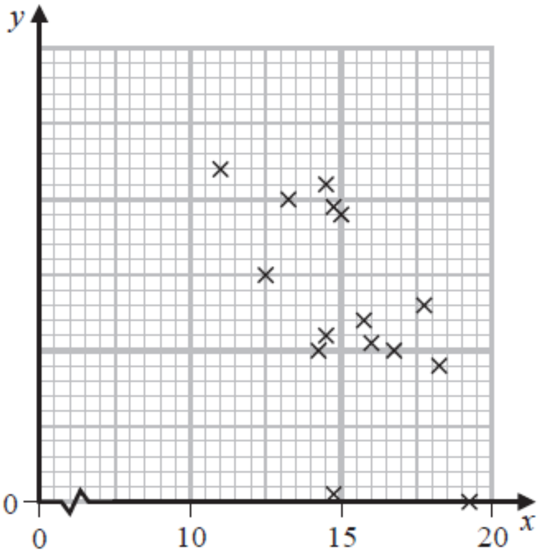


Figure 1

(*a*) Describe the correlation.

(1)

As the Values on the x-axis increase, the Values on the y-axis decrease, Which means that there is a negative Correlation between the two Variables.

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)

i) We Know that the y-axis value drops as the Daily Mean temperature increases; therefore the Y-axis Variable could be Daily Total Rainfall. (i.e. as the mean temperature increases, the total rainfall drops).

ii) Daily Total Rainfall is measured in millimetres (mm).

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)

c) we first define our test hypothesis (for Heathrow):

H_0 : Daily Total Sunshine and Daily Maximum Relative Humidity are Not significantly correlated.

V.S. H_1 : Daily Total Sunshine and Daily Maximum Relative Humidity are Significantly correlated.

or we can say that : $H_0 : \rho = 0$ v.s. $H_1 : \rho \neq 0$.

We have a Sample size of $n = 30$, and we are carrying our test out at $\alpha = 0.05$, we use the Product Moment coefficient table to find our critical Value, hich is 0.3610.

We have a negative correlation coefficient, So we use critical value -0.3610.

$r = -0.377 < -\overset{\text{critical Value}}{0.3610}$ which means that r is significant, therefore we reject H_0 and conclude that there is a significant correlation between the two variables.

Note that we carry out a two-tailed test and that is why we read our critical values from the $\frac{0.05}{2} = 0.025$ column.

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.

Humidity is high and there is evidence of correlation $(r > 0)$.

$\therefore$ expected sunshine lower on average for Heathrow

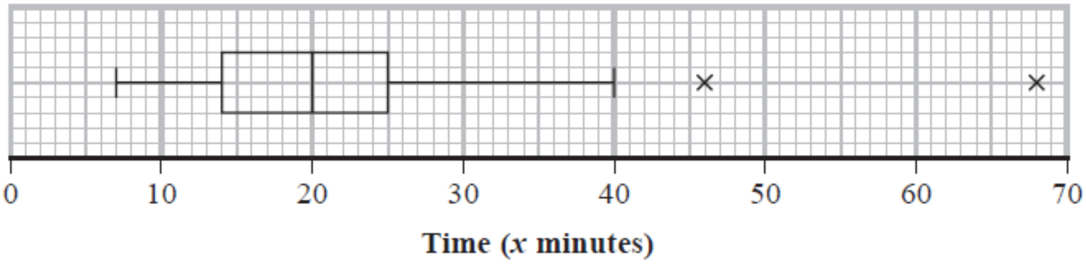
(1)

(Total for Question 1 is 7 marks)

2 Each member of a group of 27 people was timed when completing a puzzle.

The time taken, x minutes, for each member of the group was recorded.

These times are summarised in the following box and whisker plot.



(a) Find the range of the times.

(1)

We Can see the minimum value is at $x = 7$.

we can see the largest value is an outlier and is marked with an X at $x = 68$.

we therefore conclude that the range of the times is $68 - 7 = \underline{61}$.

(b) Find the interquartile range of the times.

(1)

We calculate the Interquartile range (IQR) by doing $Q_3 - Q_1$.

Reading of the plot we have that $Q_1 = 14$ and $Q_3 = 25$.

Therefore, $IQR = 25 - 14 = \underline{11}$. We can think of IQR as the length of the box in the box/whisker plot.

For these 27 people $\sum x = 607.5$ and $\sum x^2 = 17623.25$

(c) calculate the mean time taken to complete the puzzle,

(1)

The mean time is found by summing all the times and dividing by the number of People, and let's denote the mean by $\bar{x}$.

$$\Rightarrow \quad \bar{x} = \frac{\sum x}{n} = \frac{607.5}{27} = \underline{22.5} \text{ minutes.}$$

(*d*) calculate the standard deviation of the times taken to complete the puzzle.

(2)

Standard Deviation can be using the formula given in the formula Sheet and let's denote it by σ.

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{17623.25}{27} - (22.5)^2} = \underline{\underline{12.1}} \text{ minutes}$$

Taruni defines an outlier as a value more than 3 standard deviations above the mean.

(*e*) State how many outliers Taruni would say there are in these data, giving a reason for your answer.

(1)

Recall that the mean $\bar{x} = 22.5$ and $\sigma = 12.1$.

$\Rightarrow$ $22.5 + 3(12.21) = 59.1$; which means that any value over 59.1 will be an outlier. looking back at the data, we see that we have one outlier, that is $x = 68$.

Adam and Beth also completed the puzzle in *a* minutes and *b* minutes respectively, where *a* > *b*.

When their times are included with the data of the other 27 people

- the median time increases
- the mean time does not change

(*f*) Suggest a possible value for *a* and a possible value for *b*, explaining how your values satisfy the above conditions.

(3)

f) we know that the mean is the sum of all the values divided by the number of people, and the median is the 'middle value'.

we know that our mean $\bar{x}$ is 22.5 and our median is 20 (reading from box plot).

For the mean to remain unchanged we required : $\frac{607.5}{27} = \frac{607.5 + a+b}{29} \Rightarrow 29 \times 22.5 - 607.5 = a+b.$

$\Rightarrow a+b = 45$. let us choose a and b such that both a and b are greater than the current median. let $a = 23$ and $b = 22$ and we see now that our mean will remain the same but median will increase.

(*g*) Without carrying out any further calculations, explain why the standard deviation of all 29 times will be lower than your answer to part (d).

(1)

looking back at formula: $\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$, n increases but the mean $\bar{x}$ stays the Same, hence overall σ will be lower.

(Total for Question 2 is 10 marks)

3. The number of hours of sunshine each day, *y*, for the month of July at Heathrow are summarised in the table below.

Hours	$0 \le y < 5$	$5 \le y < 8$	$8 \le y < 11$	$11 \le y < 12$	$12 \le y < 14$
Frequency	12	6	8	3	2

A histogram was drawn to represent these data. The $8 \le y < 11$ group was represented by a bar of width 1.5 cm and height 8 cm.

(*a*) Find the width and the height of the $0 \le y < 5$ group.

(3)

For $8 \le y \le 11$, the width is 1.5cm and the height is 8cm. This means that the area can be found by doing $\text{Area}_1 = 1.5 \times 8 = 12$ and then we have that the area per 'frequeny observed' is $\frac{12}{8} = 1.5\text{cm}^2$ for each observation.

$\Rightarrow$ For $0 \le 5 \le 11$ we have $1.5 = \frac{\text{Area}_2}{12} \Rightarrow \text{Area}_2 = 18\text{cm}^2$.

$\Rightarrow$ We can then find the width of $0 \le y \le 5$ by doing $\frac{1.5}{11-8} \times 5 = 2.5$, where 1.5 is the width of $8 \le y \le 11$ and $11-8$ is the range of values in $8 \le y \le 11$.

This means that width $0 \le y \le 5$ is 2.5cm.

Then the height will be $\frac{\text{Area}}{\text{width}} = \frac{18}{2.5} = \underline{\underline{7.2\text{cm}}}$.

(*b*) Use your calculator to estimate the mean and the standard deviation of the number of hours of sunshine each day, for the month of July at Heathrow. Give your answers to 3 significant figures.

(3)

b)

Hours	Frequency	Midpoint	Midpoint^2	freq × midpoint^2	freq × midpoint
0 → 5	12	2.5	6.25	75	30
5 → 8	6	6.5	42.25	253.5	39
8 → 11	8	9.5	90.25	722	76
11 → 12	3	11.5	132.25	396.75	34.5
12-14	2	13.0	169	338	26

Then Σ frequency $= 31$ and Σ freq × midpoint^2 $= 1785.25$, Σ

Then the mean can be found by doing Σ frequency × midpoint $= 6.629 \doteq 6.63$

Then $\sigma = \sqrt{\dfrac{1785.25 - 31 \times (6.629)^2}{31}} = 3.6921\ldots \ldots = 3.69$ (3 sig figs) hours

The mean and standard deviation for the number of hours of daily sunshine for the same month in Hurn are 5.98 hours and 4.12 hours respectably. Thomas believes that the further south you are the more consistent should be the number of hours of daily sunshine.

(*c*) State, giving a reason, whether or not the calculations in part (*b*) support Thomas' belief.

(2)

- The Standard Deviation of Heathrow (3.69) is less than that of Hurn (4.12)
- The mean of Heathrow (6.63) is greater than that of Hurn (5.98).

Standard Deviation is a measure of how consistent values are ; a lower standard deviation implies the values are closer together / more consistent / less spread out.

=> the number of hours of Sunshine at Heathrow is more consistent than the number of hours of Sunshine in Hurn.

However, Hurn is further South than Heathrow which means that Thomas's belief is not Supported by the calculations.

(*d*) Estimate the number of days in July at Heathrow where the number of hours of sunshine is more than 1 standard deviation above the mean.

(2)

Heathrow mean is 6.63 an standard deviation is 3.69.

=> Mean + 1 × Standard deviation is 10.32.

We can say for definite that all observations in $11 \leq y \leq 12$ and $12 \leq y \leq 14$ are greater than 10.32. Then we can say estimate how many observations in $8 \leq y \leq 11$ are greater than 10.32 by doing $\dfrac{11 - 10.32}{3} \times 8 = 1.8$

=> # days $= 1.8 + 3 + 2 = 6.81$ days

Helen models the number of hours of sunshine each day, for the month of July at Heathrow by N(6.6, 3.7^2).

(*e*) Use Helen’s model to predict the number of days in July at Heathrow when the number of hours of sunshine is more than 1 standard deviation above the mean.

(2)

I be a random variable, denoted X, which is normally distributed, i.e. $X \sim N(6.6, 3.7^2)$.

This means that the mean $\mu = 6.6$ and the standard deviation = Variance^2 => $\sigma = 3.7$.

Then $Z = \frac{X - \mu}{\sigma} = \frac{10.32 - 6.6}{3.7} = 1.0054 = 1$. This means that $\mathbb{P}(X > 10.32) = \mathbb{P}(X > 1)$ which is equivelant to $1 - \mathbb{P}(X \leq 1)$. We read $\mathbb{P}(X = 1)$ from the Normal Distribution tables => $1 - 0.84134 = 0.159$.

Then the predicted number of days will be $31 \times 0.15 = 4.93$ days, where 31 is the number of days in July.

(*f*) Use your answers to part (*d*) and part (*e*) to comment on the suitability of Helen’s model.

(1)

Part d = 6.8 days > 4.73 days = Part(e) => the model is not suitable due to the difference.

(Total for Question 3 is 13 marks)

4. The partially completed table below summarises the times taken by 120 job applicants to complete a task.

Time, *t* (minutes)	$5 < t \leq 7$	$7 < t \leq 10$	$10 < t \leq 14$	$14 < t \leq 18$	$18 < t \leq 30$
Frequency	10	23	51		

A histogram is drawn. The bar representing the $5 < t \leq 7$ has a width of 1 cm and a height of 5 cm.

(a) Given that the bar representing the group $14 < t \leq 18$ has a height of 4 cm, find the frequency of this group.

(2)

we have that the height is equal to the frequency divided by the interval width : $5 < t \leq 7$: $\frac{10}{7-5} = \frac{10}{2} = 5\text{cm}$ and we can use this to help us calculate the frequency of the $14 < t \leq 18$ group.

$$\Rightarrow \text{height} = \frac{\text{freq.}}{\text{interval width}} \Rightarrow 4 = \frac{\text{freq}}{18-14} \Rightarrow \text{frequency} = \underline{16}$$

(b) Showing your working, estimate the mean time taken by the 120 job applicants.

(3)

The frequencies must add to 120 which means that the frequency of $18 < t \leq 30$ can be calculated (let this be f_5), then $120 = 10 + 23 + 51 + 16 + f_5$, which means that $f_5 = 20$.

Then the mean can be worked out by calculating the mid-point of each interval and multiplying it by the frequency value, we then sum this and divide by n.

Time	Frequency	Midpoint	Frequency × Midpoint
$5 < t \leq 7$	10	~~7~~6	10
$7 < t \leq 10$	23	8.5	195.5
$10 < t \leq 14$	51	12	612
$14 < t \leq 18$	16	16	256
$18 < t \leq 30$	20	24	480

$$\Rightarrow \sum \text{frequency} \times \text{midpoint} = 1603.5 \Rightarrow \text{Mean} = \frac{1603.5}{120} = \underline{13.4} \text{ minutes.}$$

The lower quartile of the times is 9.6 minutes and the upper quartile of the times is 15.5 minutes.

For these data, an outlier is classified as any value greater than $Q_3 + 1.5 \times \text{IQR}$.

(c) Showing your working, explain whether or not any of the times taken by these 120 job applicants might be classified as outliers.

(2)

We have that $Q_1 = 9.6$ and $Q_3 = 15.5$ minutes.

This means that the $IQR = Q_3 - Q_1 = 15.5 - 9.6 = \underline{5.9}$

Then checking for outliers: $Q_3 + 1.5 \times IQR = 15.5 + 1.5 \times 5.9 = 24.35$.

This means that there could be outliers in the $18 < t \leq 30$ interval.

We can work out how many should be above 24.35 by doing

$\frac{30 - 24.35}{30-18} \times 18 = 8.475 = 8 \Rightarrow$ We have estimated that there will be 8 outliers in the dataset.

Candidates with the fastest 5% of times for the task are given interviews.

(d) Estimate the time taken by a job applicant, below which they might be given an interview.

(2)

We have 120 candidates, so 5% of 20 is 6.

We know that 10 people get between 5 and 7 minutes.

Then we know that $\frac{6}{10}$ of these times will be in the fastest 5%.

This equates to $\frac{6}{10} \times (7-5) = 1.2$. $\Rightarrow 5 + 1.2 = 6.2$ mins.

Therefore anyone who completes the task in less than 6.2 minutes will be given an interview.

(Total for Question 4 is 9 marks)

5.

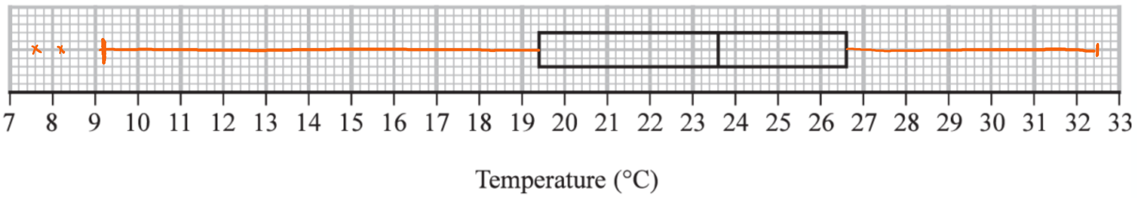


Figure 1

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 × *IQR* below Q_1 or
more than 1.5 x 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)**

$IQR = Q_3 - Q_1 = 26.6 - 19.4 = 7.2$ then the outliers are :

- $19.4 - 1.5 \times 7.2 = 8.6 \Rightarrow 7.6°C$ and $8.1°C$ are outliers(but $9.1°C$ is not)
- $26.6 + 1.5 \times 7.2 = 37.4 \Rightarrow 32.5°C$ is not an outlier.

(*b*) Using your knowledge of the large data set, suggest from which month the two outliers are likely to have come. October **(1)**

Using the data from the large data set, Simon produced the following summary statistics for the daily mean air temperature, *x* °C, for Beijing in 2015

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

(*c*) Show that, to 3 significant figures, the standard deviation is 5.19 °C **(1)**

$\sigma = \sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{4952.906}{184}} = 5.1882... = \underline{5.19°C}$ as required.

Simon decides to model the air temperatures with the random variable

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

(*d*) Using Simon's model, calculate the 10th to 90th interpercentile range. **(3)**

Mean: $\mu = 22.6$ and Standard deviation: $\sigma = 5.19$. from our normal distribution.

We first find the 10th percentile, we find this from normal distribution table.

$\Rightarrow z = -1.2816$ then $-1.2816 = \frac{x_1 - 22.6}{5.19} \Rightarrow x_1 = \underline{15.9}$

$\Rightarrow$ the 90th Percentile is then: $1.2816 = \frac{x_2 - 22.6}{5.19} \Rightarrow x_2 = \underline{29.3}$

$\Rightarrow$ the interpercentile range is $x_2 - x_1 = 29.3 - 15.9 = \underline{13.4}$

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)

(Total for Question 5 is 11 marks)

Rainfall since not symmetric (lots of days with 0 rainfall)

6. Charlie is studying the time it takes members of his company to travel to the office. He stands by the door to the office from 08 40 to 08 50 one morning and asks workers, as they arrive, how long their journey was.

(a) State the sampling method Charlie used.

(1)

Charlie uses opportunity Sampling since he makes use of people as and when they become available.

(b) State and briefly describe an alternative method of non-random sampling Charlie could have used to obtain a sample of 40 workers.

(2)

Charlie could use Quota Sampling where he would create subgroups of data. He should ask 20 men and 20 women.

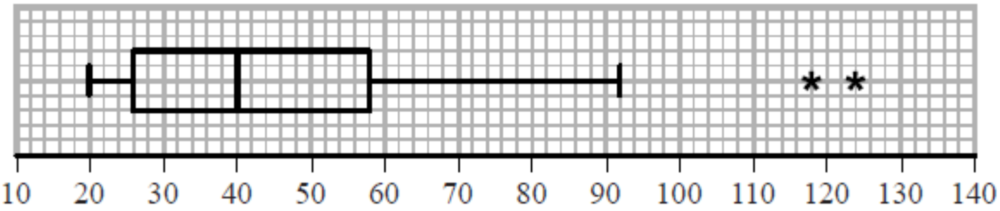
Taruni decided to ask every member of the company the time, *x* minutes, it takes them to travel to the office.

(c) State the data selection process Taruni used.

(1)

Every member of the company was asked - this is called a census

Taruni's results are summarised by the box plot and summary statistics below.



Journey time (minutes)

$n = 95$ $\Sigma x = 4133$ $\Sigma x^2 = 202\ 294$

(d) Write down the interquartile range for these data.

(1)

$IQR = Q_3 - Q_1 = 58 - 26 = \underline{32}$ minutes.

(e) Calculate the mean and the standard deviation for these data.

(3)

$\bar{x} = \frac{\Sigma x}{n} = \frac{4133}{95} = \underline{43.5}$ minutes and

$\frac{202\ 294}{95} - (43.5)^2$

$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}} = \sqrt{\frac{202\ 294}{95} - (43.5)^2} = \underline{15.4}$

(f) State, giving a reason, whether you would recommend using the mean and standard deviation or the median and interquartile range to describe these data.

(2)

The data is not Symmetric (shown in the box plot), so we should use the median and IQR to describe these data.

Rana and David both work for the company and have both moved house since Taruni collected her data. Rana's journey to work has changed from 75 minutes to 35 minutes and David's journey to work has changed from 60 minutes to 33 minutes.

Taruni drew her box plot again and only had to change two values.

(g) Explain which two values Taruni must have changed and whether each of these values has increased or decreased.

(3)

The median was 40 but will change since both 75 and 60 mins are replaced by 35 and 33 mins and they're both less than 40, which means that the median will be 'two-people' less than 40.

The upper quantile Q_3 will also decrease since 75 and 60 were between Q_3 and the maximum value.

(Total for Question 6 is 13 marks)