Additional Assessment Materials

Summer 2021

Pearson Edexcel GCE in As Mathematics

8MA0_21 (Public release version)

Resource Set 1: Topic 1

Statistical sampling

Pearson: helping people progress, everywhere

Pearson aspires to be the world's leading learning company. Our aim is to help everyone progress in their lives through education. We believe in every kind of learning, for all kinds of people, wherever they are in the world. We've been involved in education for over 150 years, and by working across 70 countries, in 100 languages, we have built an international reputation for our commitment to high standards and raising achievement through innovation in education. Find out more about how we can help you and your students at: www.pearson.com/uk

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- Additional Assessment Materials have come from past papers both published (those materials available publicly) and unpublished (those currently under padlock to our centres) presented in a different format to allow teachers to adapt them for use with candidate.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. Sara is investigating the variation in daily maximum gust, t kn, for Camborne in June and July 1987.

She used the large data set to select a sample of size 20 from the June and July data for 1987. Sara selected the first value using a random number from 1 to 4 and then selected every third value after that.

(a) State the sampling technique Sara used.

(1)

a) Sara makes use of Systematic Sampling. This is where a random start point is chosen and then every nth value after that is chosen. In our case, $n = 3$.

(b) From your knowledge of the large data set, explain why this process may not generate a sample of size 20. **(1)**

b) There may be missing data due to it being a large data set.

The data Sara collected are summarised as follows

$$n = 20 \qquad \sum t = 374 \qquad \sum t^2 = 7600$$

(c) Calculate the standard deviation.

(2)

(Total for Question 1 is 4 marks)

c) We want to find the standard deviation of a sample, So we use the formula: $S = \sqrt{\frac{\sum t^2 - n\bar{t}^2}{n-1}}$. We know that $n = 20$, $\sum t^2 = 7600$, and $\bar{t} = \frac{374}{20} = 18.7$

$$\Rightarrow S = \sqrt{\frac{7600 - 20(18.7)^2}{20-1}} = 5.6484\ldots = \underline{5.65}$$

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)

a) A stratified sample involves breaking a whole population down into subgroups and we then take a certain number of samples from each subgroup.
We must proportionally pick the number of samples from each subgroup and we do this as follows:

$$\text{Sample size Year 12} = \frac{40}{84+56} \times 84 = 24$$

$$\text{Sample size Year 13} = \frac{40}{84+56} \times 56 = 16$$

Therefore, rounding to the nearest whole number, we should take 24 samples from year 12 and 16 samples from year 13.

(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)

b) We have the equation of the regression line to be $P = 26.1 + 5.60s$.

$P_1 = 26.1 + 5.60 = 31.7$
$P_2 = 26.1 + 5.60(1.5) = 34.5$
$P_3 = 26.1 + 5.60(2) = 37.3$
$P_4 = 26.1 + 5.60(2.5) = 40.1$

$P_2 - P_1 = 34.5 - 31.7 = 2.8$
$P_3 - P_2 = 37.3 - 34.5 = 2.8$
$P_4 - P_3 = 40.1 - 37.3 = 2.8$

Therefore, from a manual check we have that an extra 0.5 hours of sleep will increase student's performance by 2.8 marks on average.

(c) Describe one limitation of this regression model.

c) One limitation is that outliers may be present in the data and these could have an effect on the accuracy of the model.

(1)

(Total for Question 2 is 5 marks)

3. A lake contains three different types of carp.

There are an estimated 450 mirror carp, 300 leather carp and 850 common carp.

Tim wishes to investigate the health of the fish in the lake.

He decides to take a sample of 160 fish.

(a) Give a reason why stratified random sampling cannot be used.

a) If we use Stratified Sampling we would have to have a Sampling frame which contains each and every fish, but it would be impractical to do this due to the nature of the task. **(1)**

(b) Explain how a sample of size 160 could be taken to ensure that the estimated

populations of each type of carp are fairly represented.

You should state the name of the sampling method used.

(2)

b) Tim Should use Quota Sampling, where proportions will be taken into account, i.e. More Common Carp should be Sampled Compared to mirror Carp, and more mirror Carp Compared to leather Carp.

As part of the health check, Tim weighed the fish.

His results are given in the table below.

Weight (w kg)	**Frequency (f)**	**Midpoint (m kg)**
$2 \leq w < 3.5$	8	2.75
$3.5 \leq w < 4$	32	3.75
$4 \leq w < 4.5$	64	4.25
$4.5 \leq w < 5$	40	4.75
$5 \leq w < 6$	16	5.5

(You may use $\sum fm = 692$ and $\sum fm^2 = 3053$)

(*c*) Calculate an estimate for the standard deviation of the weight of the carp.

(2)

c) Here, we want to calculate the standard deviation of a whole population. So our formula is $\sigma = \sqrt{\frac{\sum fm^2 - n\bar{m}^2}{n}}$.

We have that $n = 160$, $\sum fm^2 = 3053$ and $\bar{m} = \frac{692}{160} = 4.325$.

$$\Rightarrow \sigma = \sqrt{\frac{3053 - 160(4.325)^2}{160}} = 0.6128\ldots = \underline{0.61}$$

Tim realised that he had transposed the figures for 2 of the weights of the fish.

He had recorded in the table 2.3 instead of 3.2 and 4.6 instead of 6.4

(d) Without calculating a new estimate for the standard deviation, state what effect

(i) using the correct figure of 3.2 instead of 2.3

(ii) using the correct figure of 6.4 instead of 4.6

would have on your estimated standard deviation.

Give a reason for each of your answers.

d:) $2.3 \rightarrow 3.2$ would have no change in the standard deviation as 2.3 and 3.2 both fall under the same weight interval (column 1).

d::) We know that standard deviation is a measure of how spread out data are, and a lower standard deviation means that the data is more likely to be around the mean ($\bar{m}$).
4.6 is closer to $\bar{m} = 4.325$ which means that using 4.6 instead of 6.4 will result in a lower standard deviation.

(2)

(Total for Question 3 is 7 marks)

4. 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)**

a) IQR = $Q_3 - Q_1 = 2.4 - 0.1 = 2.3$

Then $1.5 \times IQR = 1.5 \times 2.3 = 3.45$.

Then $Q_3 + 3.45 = 2.4 + 3.45 = \underline{5.85}$

$r = 20.6 > 5.85 \Rightarrow r = 20.6$ is an outlier.

(b) Give a reason why Sara might

(i) include

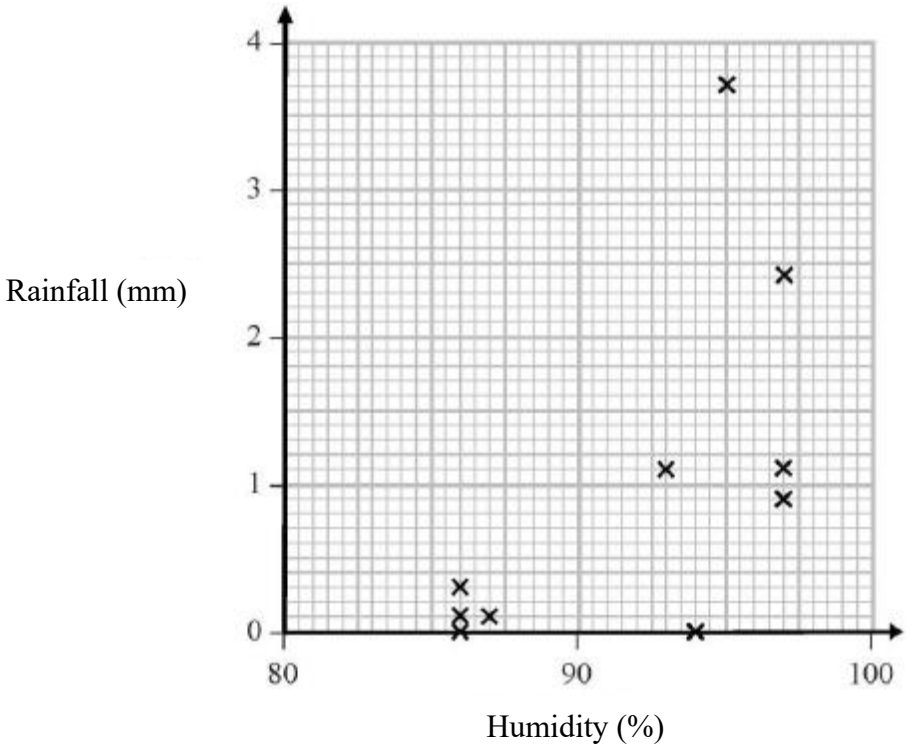
(ii) exclude

this day's reading. **(2)**

b) i) Sara might include this because there is nothing to suggest there was a recording error, i.e. it was just a very hum'd day.

ii) Sara might exclude this because it may have been recorded wrongly given it is so much larger than any other data point, and excluding it will allow for better conclusions to be drawn from the remaining results.

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)

c) From the graph, we can see that as the humidity level increases, the rainfall will increase, thus there is a positive Correlation.

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)**

d) $r = -12.8 + 0.15h$ => the gradient can be read off the equation from the coefficient of h. It is 0.15. Recall that gradient is $\frac{\Delta y}{\Delta x} = 0.15$ => $\Delta y = \Delta x \times 0.15$, So each time humidity increases by 1%, the rainfall will increase, on average, by 0.15 mm.

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

e) i) Sara used a combination of Quota Sampling (She specified data to be from a certain Month/Year) and then used random sampling within that Month. Hence this is Suitable.

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

e) ii) Sara should increase the Sample Size to better use the large data set. She could use Systematic Sampling across a whole year to randomly choose 'more days'.

(2)

(Total for Question 4 is 7 marks)