Worksheet - Statistical sampling and data

Year: Year 12 · 21

Show your working.

Question 1

Question 1 – Sampling methods *(Section A · Skills practice)*

A sixth-form college wants to survey 60 of its 600 students about parking. The students are in four year groups.

Year group	10	11	12	13
Number of students	150	130	190	130

a. Describe how to take a stratified sample of 60 and work out how many students come from each year group. [4 marks]
b. Describe how to take a systematic sample of 60 from an alphabetical list of the 600 students. [3 marks]
c. Give one disadvantage of asking the first 60 students who arrive in the morning. [2 marks]

Question 2

Question 2 – Mean, median and mode *(Section A · Skills practice)*

The number of emails received on seven days was 12, 15, 15, 18, 20, 23, 40.

a. Find the mean, median and mode. [4 marks]
b. The value 40 is removed. Find the new mean and median. [3 marks]
c. Which average is most affected by the extreme value, and why? [2 marks]

Question 3

Question 3 – Mean and standard deviation from a frequency table *(Section A · Skills practice)*

A survey records the number of pets in 30 households.

Pets x	1	2	3	4	5
Frequency f	4	7	10	6	3

a. Find $\sum fx$ and $\sum fx^2$. [3 marks]
b. Find the mean and the standard deviation. [4 marks]
c. State the median. [2 marks]

Question 4

Question 4 – Variance and coding *(Section A · Skills practice)*

Use the formulas for variance and standard deviation.

a. Find the mean and standard deviation of 4, 7, 8, 11, 15. [4 marks]
b. For 10 values of x it is given that $\sum(x-50)=30$ and $\sum(x-50)^2=250$. Find the mean and standard deviation of x. [4 marks]

Question 5

Question 5 – Quartiles, box plots and outliers *(Section A · Skills practice)*

Ten delivery times (in minutes) are: 3, 5, 6, 8, 9, 10, 12, 13, 14, 31.

a. Find the median, lower quartile, upper quartile and IQR. [4 marks]
b. A value is an outlier if it is more than 1.5 IQR above Q_3 or below Q_1. Find the limits. [3 marks]
c. Identify any outliers and describe how to show them on a box plot. [2 marks]

Question 6

Question 6 – Reading a histogram *(Section A · Skills practice)*

The times taken by students to complete a puzzle are summarised.

Time (minutes)	0-10	10-20	20-30	30-50	50-80
Frequency	5	12	20	24	9
Frequency density	0.5	1.2	2.0	1.2	0.3

a. Show how the frequency density for 30-50 is calculated. [1 mark]
b. Estimate the number of students taking between 25 and 40 minutes. [3 marks]
c. Explain why the bar for 50-80 is low even though it is the widest. [2 marks]

Question 7

Question 7 – Scatter diagrams and regression *(Section A · Skills practice)*

Six students recorded hours of revision x and a test mark y (out of 10).

x	1	2	3	4	5	6
y	3.1	4.5	5.2	6.8	7.1	8.4

a. Describe the correlation. [1 mark]
b. The regression line is $y=2.26+1.026x$. Interpret the gradient. [2 marks]
c. Use it to estimate the mark for 4.5 hours of revision. [2 marks]
d. Explain why a prediction for 12 hours is unreliable. [2 marks]

Question 8

Question 8 – Choosing a sample for a survey *(Section B · Real-life practice)*

A town council wants views on a new leisure centre. Mia suggests asking people leaving the existing leisure centre one Saturday morning. Omar suggests choosing 200 names at random from the electoral register.

a. Name each sampling method. [2 marks]
b. Give one advantage and one disadvantage of Mia\'s method. [3 marks]
c. Give one advantage and one disadvantage of Omar\'s method. [3 marks]
d. Which method would you recommend? Explain. [2 marks]

Question 9

Question 9 – Temperatures at a weather station *(Section B · Real-life practice)*

Daily maximum temperatures (°C) at a weather station on ten days in June: 12.1, 13.4, 11.8, 14.0, 12.9, 15.2, 13.3, 12.5, 22.4, 13.0. (Invented data in the style of the large data set.)

a. Find the mean and standard deviation. [4 marks]
b. Find the median and quartiles. [4 marks]
c. Test whether 22.4 is an outlier using the 1.5 IQR rule. [3 marks]
d. Which measures of location and spread are best for these data? [2 marks]

Question 10

Question 10 – Hourly pay *(Section B · Real-life practice)*

The hourly pay of 75 employees is summarised in the table.

Pay (£)	8-10	10-12	12-16	16-24
Frequency	12	25	30	8

a. Find the modal class. [1 mark]
b. Estimate the mean pay using mid-points. [4 marks]
c. Estimate the median using interpolation. [4 marks]
d. Explain why these are estimates only. [1 mark]

Question 11

Question 11 – Journey times to work *(Section B · Real-life practice)*

The time taken by 240 workers to reach work is recorded.

Time (min)	0-10	10-20	20-30	30-45	45-75
Frequency	30	60	75	45	30

Time (min)	0-10	10-20	20-30	30-45	45-75
Frequency density	3	6	7.5	3	1

a. Check the frequency density for 45-75. [1 mark]
b. Estimate the number taking between 25 and 40 minutes. [3 marks]
c. Estimate the mean journey time. [4 marks]
d. Describe the shape of the distribution. [1 mark]

Question 12

Question 12 – Comparing two classes *(Section B · Real-life practice)*

Two classes sat the same test. The five-number summaries are shown.

	Min	Q_1	Median	Q_3	Max
Class A	12	35	48	62	90
Class B	20	40	50	58	75

a. Find the interquartile range for each class. [2 marks]
b. Compare the two classes. [3 marks]
c. Which class had the more consistent results? [1 mark]
d. Show that there are no outliers in Class A. [3 marks]

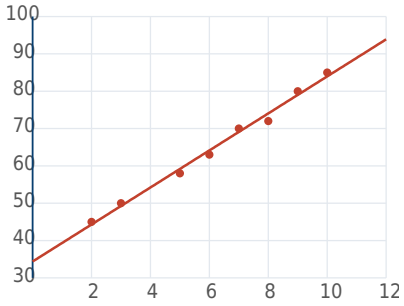
Question 13

Question 13 – Revision and results *(Section B · Real-life practice)*

Eight students recorded revision time x (hours) and test score y (marks out of 100).

x	2	3	5	6	7	8	9	10
y	45	50	58	63	70	72	80	85

a. The regression line is $y = 34.4 + 4.96x$. Interpret the gradient and the intercept. [3 marks]
b. Estimate the score for 4 hours of revision. [2 marks]
c. Explain why the estimate for 12 hours is unreliable. [2 marks]
d. The product-moment correlation coefficient is 0.996. Interpret. [2 marks]



Question 14

Question 14 – Coding potato weights *(Section B · Real-life practice)*

The masses x grams of 40 potatoes are coded using $y = \dfrac{x - 100}{10}$. It is found that $\sum y = 64$ and $\sum y^2 = 360$.

a. Find the mean and standard deviation of y. [4 marks]
b. Hence find the mean and standard deviation of x. [4 marks]
c. A potato weighing 175 g is found. Is it more than 2 standard deviations above the mean? [2 marks]

Question 15

Question 15 – The effect of an outlier *(Section B · Real-life practice)*

Delivery times (in minutes) for a pizza shop on 8 evenings were: 22, 25, 26, 27, 28, 29, 31, 90.

a. Find the mean and median. [3 marks]
b. Show that 90 is an outlier using the 1.5 IQR rule. [4 marks]
c. Find the mean without the outlier and comment. [3 marks]
d. Suggest a reason for the outlier and what the shop should do. [1 mark]

Question 16

Question 16 – Completing a histogram *(Section C · Problem solving)*

The ages of people visiting a museum were recorded. The histogram has the frequency densities below.

Age	0-20	20-30	30-40	40-60	60-80
Frequency density	1.5	4	5	2	0.5
Frequency	?	?	?	?	?

a. Complete the frequency row and state the total. [4 marks]
b. Find the median age by interpolation. [3 marks]
c. Estimate the mean age. [4 marks]
d. Estimate the number of visitors who were at least 45 years old. [2 marks]

Question 17

Question 17 – Advertising and sales *(Section C · Problem solving)*

A shop records advertising spend x (£100s) and sales y (£1 000s) for 8 months.

x	1	2	3	4	5	6	7	8

x	1	2	3	4	5	6	7	8
y	2.1	2.8	3.2	4.1	4.4	5.3	5.6	6.4

a. The regression line is $y = 1.52 + 0.604x$. Interpret the gradient in context. [2 marks]
b. Estimate sales for an advertising spend of £450. [2 marks]
c. The shop wants sales of £10 000. Use the line to find the advertising needed and say why this is risky. [3 marks]
d. The correlation coefficient is 0.996. What does it tell you, and does it mean advertising causes the sales? [3 marks]

Question 18

Question 18 – Two weather months *(Section C · Problem solving)*

Daily rainfall (mm) at a coastal town is summarised for May and August. (Invented values in the style of the large data set.)

	Q_1	Median	Q_3	Mean	Largest value
May	0.0	0.4	3.2	2.6	14.2
August	0.0	1.0	4.6	3.4	12.4

a. Find the IQR for each month. [2 marks]
b. Show that the largest value in each month is an outlier. [4 marks]
c. The mean is greater than the median in both months. What does this say about the shape? [2 marks]
d. Which month is more variable? [1 mark]

Question 19

Question 19 – Designing a stratified sample *(Section C · Problem solving)*

A school has 1 200 students. The numbers in each year group and gender are shown. A stratified sample of 120 is to be taken.

	Year 10	Year 11	Year 12	Year 13
Boys	140	130	150	180
Girls	160	170	150	120

a. Explain why the sampling fraction is $\frac{1}{10}$. [1 mark]
b. Complete a table of the sample sizes. [4 marks]
c. Why is stratified sampling better than simple random sampling here? [2 marks]
d. Describe how the students in the Year 12 girls\' stratum could be chosen. [2 marks]

Question 20

Question 20 – Regression from summary statistics *(Section C · Problem solving)*

For 10 pairs of data it is given that $\sum x = 50$, $\sum y = 120$, $\sum x^2 = 300$, $\sum xy = 640$ and $\sum y^2 = 1480$.

a. Find S_{xx}, S_{xy} and S_{yy}. [4 marks]
b. Find the regression line $y = a + bx$. [4 marks]
c. Find the product-moment correlation coefficient. [3 marks]
d. Estimate y when $x = 6$. [1 mark]

Question 21

Question 21 – Challenge: combining two groups *(Section D · Challenge)*

Group A has 20 students with mean score 60 and standard deviation 5. Group B has 30 students with mean score 70 and standard deviation 8.

a. Find the mean of the combined group. [2 marks]
b. Find $\sum x^2$ for each group using $\sum x^2 = n(\sigma^2 + \bar{x}^2)$. [4 marks]
c. Find the standard deviation of the combined group. [4 marks]