# Answers - Statistical sampling and data

Year: Year 12

## Answer to Question 1

a. Stratified: sample proportional to each group. The fraction is $\frac{60}{600} = \frac{1}{10}$, so year 10: **15**, year 11: **13**, year 12: **19**, year 13: **13** (total 60). Select each group randomly (e.g. random numbers).

b. The interval is $\frac{600}{60} = 10$. Choose a random start between 1 and 10 (say 7) and then take every 10th student: 7, 17, 27, \dots

c. It is an opportunity sample: early arrivals may be unrepresentative (e.g. only those who use buses or who are keen), so the sample is biased and not random.

## Answer to Question 2

a. Mean $= \frac{143}{7} =$ **20.4**. Median (4th value) $=$ **18**. Mode $=$ **15**.

b. New data: 12, 15, 15, 18, 20, 23. Mean $= \frac{103}{6} =$ **17.2**. Median $= \frac{15+18}{2} =$ **16.5**.

c. The mean changes by 3.2 and the median by 1.5. The mean uses every value so it is pulled towards an extreme value; the median depends only on the middle.

## Answer to Question 3

a. Multiply each value $x$ by its frequency $f$, and each $x^2$ by $f$, then add. $\sum fx = 4 + 14 + 30 + 24 + 15 =$ **87**. $\sum fx^2 = 4 + 28 + 90 + 96 + 75 =$ **293**.

b. Use $\bar{x} = \frac{\sum fx}{\sum f}$ and $\sigma^2 = \frac{\sum fx^2}{\sum f} - \bar{x}^2$. Mean $= \frac{87}{30} =$ **2.9**. Variance $= \frac{293}{30} - 2.9^2 = 9.767 - 8.41 = 1.357$. Standard deviation $= \sqrt{1.357} =$ **1.16**. On average a household has 2.9 pets, with a standard deviation of about 1.16.

c. With 30 values, the median is the mean of the 15th and 16th values; use cumulative frequencies. The 15th and 16th values lie in the group $x = 3$ (cumulative frequencies 4, 11, 21), so the median is **3**.

## Answer to Question 4

a. Find the mean first, then use $\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2$. Mean $= 9$.

$\sum x^2 = 16 + 49 + 64 + 121 + 225 = 475$. Variance $= \frac{475}{5} - 81 = 14$. Standard deviation $= \sqrt{14} =$ **3.74**. The values lie on average about 3.74 from the mean of 9.

b. Coding with $y = x - 50$ shifts every value by 50; find $\bar{y}$ and $\sigma_y$ first, then decode. Let $y = x - 50$. $\bar{y} = \frac{30}{10} = 3$, so $\bar{x} =$ **53**. Variance of $y = \frac{250}{10} - 9 = 16$, so the standard deviation is $\sqrt{16} =$ **4** (subtracting a constant does not change the spread).

## Answer to Question 5

a. Median $= \frac{9 + 10}{2} =$ **9.5**. Lower half 3, 5, 6, 8, 9 gives $Q_1 =$ **6**; upper half 10, 12, 13, 14, 31 gives $Q_3 =$ **13**. IQR $= 13 - 6 =$ **7**.

b. $Q_3 + 1.5 \times 7 = 23.5$ and $Q_1 - 1.5 \times 7 = -4.5$.

c. 31 is above 23.5, so it is an **outlier**. On a box plot it is marked with a cross beyond the whisker, and the whisker stops at the largest non-outlier (14).

## Answer to Question 6

a. In a histogram the height of each bar is the frequency density. Frequency density $= \frac{\text{frequency}}{\text{class width}} = \frac{24}{20} = 1.2$. So the bar for 30-50 has height 1.2.

b. Use linear interpolation: 25-30 is half of the class 20-30 and 30-40 is half of the class 30-50. Assume frequencies are evenly spread within each class. 25-30: half of $20 = 10$. 30-40: half of $24 = 12$. Total $\approx$ **22**. About 22 students took between 25 and 40 minutes.

c. In a histogram the area (not the height) represents frequency; a wide class with few students has a low frequency density.

## Answer to Question 7

a. Look at the pattern of the points: as $x$ increases, $y$ increases and the points lie close to a straight line. Strong **positive** linear correlation.

b. The gradient is the change in $y$ for each increase of 1 in $x$. Each extra hour of revision is associated with an increase of about 1.03 marks.

c. Substitute $x = 4.5$ into the regression line; 4.5 hours is within the range of the data. $y = 2.26 + 1.026 \times 4.5 =$ **6.88** (interpolation, reliable). The estimated mark is about 6.9 out of 10.

d. 12 hours is outside the data range (extrapolation): the line gives 14.6, impossible

for a test marked out of 10, and the linear relationship may not continue.

## Answer to Question 8

a. Mia: **opportunity** (convenience) sample. Omar: **simple random** sample.
b. Advantage: quick and cheap. Disadvantage: biased, because it only includes people who already use a leisure centre on Saturday mornings.
c. Advantage: every adult on the register has an equal chance of selection, so it is unbiased. Disadvantage: the register excludes some residents (e.g. under 18s), and some people may not respond.
d. Omar\'s method, because it is representative of the whole population and removes the selection bias.

## Answer to Question 9

a. Mean $= \dfrac{140.6}{10} =$ **14.06**. $\sum x^2 = 2062.56$, so the standard deviation is **2.93**.
b. Ordered: 11.8, 12.1, 12.5, 12.9, 13.0, 13.3, 13.4, 14.0, 15.2, 22.4. Median $= \dfrac{13.0 + 13.3}{2} =$ **13.15**. $Q_1 = 12.5$ and $Q_3 = 14.0$ (medians of the halves).
c. IQR $= 14.0 - 12.5 = 1.5$. Upper limit $= 14.0 + 1.5 \times 1.5 = 16.25$. Since $22.4 > 16.25$, it **is an outlier**.
d. Because of the outlier the median (13.15) and IQR (1.5) are better than the mean and standard deviation, which are pulled upwards.

## Answer to Question 10

a. The modal class has the greatest frequency density: $\frac{12}{2} = 6$, $\frac{25}{2} = 12.5$, $\frac{30}{4} = 7.5$, $\frac{8}{8} = 1$. So **£10-£12**.
b. Mid-points 9, 11, 14, 20. $\sum fx = 108 + 275 + 420 + 160 = 963$. Mean $\approx \dfrac{963}{75} =$ **£12.84**.
c. $n = 75$, median at the 37.5th value. Cumulative: 12, 37, 67. It lies in 12-16: $12 + \dfrac{37.5 - 37}{30} \times 4 =$ **£12.07**.
d. The exact values within each class are not known, so we assume each value is at the mid-point (or evenly spread).

## Answer to Question 11

a. Frequency density is frequency divided by class width. $\dfrac{30}{30} = 1$. The value in the

table is correct.

b. Assume the values are spread evenly within each class and use proportions of the class widths. 25-30: half of 75 $= 37.5$. 30-40: $\frac{10}{15}$ of 45 $= 30$. Total $\approx$ **67.5** (about 68).
c. Use the mid-point of each class as a representative value. Mid-points 5, 15, 25, 37.5, 60: $\sum fx = 150 + 900 + 1875 + 1687.5 + 1800 = 6412.5$. Mean $\approx \frac{6412.5}{240} =$ **26.7 minutes**.
d. Positively skewed: a long tail to the right (a few very long journeys).

## Answer to Question 12

a. Class A: $62 - 35 = 27$. Class B: $58 - 40 = 18$.
b. The medians are similar (48 and 50), so a typical student scored about the same, but Class A has a larger spread (IQR 27 against 18, range 78 against 55).
c. Class B has the smaller spread, so it was **more consistent**.
d. Upper limit $= 62 + 1.5 \times 27 = 102.5$ and lower limit $= 35 - 40.5 = -5.5$. Max 90 and min 12 lie inside the limits.

## Answer to Question 13

a. Gradient: each extra hour of revision is associated with about 5 extra marks. Intercept: a student who does no revision would be expected to score about 34 marks (although this is outside the data range).
b. $34.4 + 4.96 \times 4 =$ **54.2 marks** (interpolation).
c. $34.4 + 4.96 \times 12 = 93.9$, but 12 hours is outside the range of the data (2 to 10 hours): extrapolation, and the relationship may level off.
d. $r = 0.996$ is very close to 1: very strong positive linear correlation. (It does not prove that revision causes higher scores.)

## Answer to Question 14

a. Use $\bar{y} = \frac{\sum y}{n}$ and $\sigma_y^2 = \frac{\sum y^2}{n} - \bar{y}^2$ with $n = 40$. $\bar{y} = \frac{64}{40} = 1.6$. Variance $= \frac{360}{40} - 1.6^2 = 9 - 2.56 = 6.44$. $\sigma_y = \sqrt{6.44} = 2.54$. So $\bar{y} = 1.6$ and $\sigma_y \approx 2.54$.
b. Rearrange the coding to $x = 10y + 100$. Adding 100 changes the mean but not the spread; multiplying by 10 multiplies both the mean and the standard deviation by 10. $x = 10y + 100$, so $\bar{x} = 10 \times 1.6 + 100 =$ **116 g** and $\sigma_x = 10 \times 2.54 =$ **25.4 g**. So the potatoes have a mean mass of 116 g and a standard deviation of about 25.4 g.
c. Find the value two standard deviations above the mean and compare. Mean

$+2\sigma = 116 + 50.8 = 166.8$ g. 175 g is greater, so **yes**. So this potato is unusually heavy compared with the rest of the sample.

## Answer to Question 15

a. Mean $= \frac{278}{8} =$ **34.75**. Median $= \frac{27+28}{2} =$ **27.5**.

b. $Q_1 = \frac{25+26}{2} = 25.5$, $Q_3 = \frac{29+31}{2} = 30$. IQR $= 4.5$. Upper limit $= 30 + 6.75 = 36.75$. $90 > 36.75$, so it is an outlier.

c. Without 90: mean $= \frac{188}{7} =$ **26.9**. The outlier raised the mean by nearly 8 minutes but the median only by 0.5, so the median is the better measure of a typical delivery.

d. A traffic incident or a mistake in recording; investigate before deciding whether to remove it.

## Answer to Question 16

a. Frequency = density $\times$ width: $30$, $40$, $50$, $40$, $10$. Total $=$ **170**.

b. Median at the 85th value. Cumulative: 30, 70, 120. It lies in 30-40: $30 + \frac{85-70}{50} \times 10 =$ **33.0 years**.

c. Mid-points 10, 25, 35, 50, 70: $\sum fx = 300 + 1000 + 1750 + 2000 + 700 = 5750$. Mean $\approx \frac{5750}{170} =$ **33.8 years**.

d. 45-60: $\frac{15}{20} \times 40 = 30$; 60-80: 10. Total $\approx$ **40**.

## Answer to Question 17

a. Each extra £100 spent on advertising is associated with about £604 extra sales.

b. £450 means $x = 4.5$: $y = 1.52 + 0.604 \times 4.5 = 4.24$, i.e. **£4 240** (interpolation).

c. $10 = 1.52 + 0.604x$ gives $x = 14.05$, i.e. about £1 400 of advertising. This is far outside the range of data (£100 to £800), so the estimate is unreliable.

d. $r$ is very close to 1: very strong positive linear correlation. Correlation does not prove cause: other factors such as the season might affect both.

## Answer to Question 18

a. IQR $= Q_3 - Q_1$. May: $3.2 - 0 = 3.2$. August: $4.6 - 0 = 4.6$.

b. Find the upper limit $Q_3 + 1.5 \times \text{IQR}$ for each month and compare the largest

value. May: upper limit $= 3.2 + 1.5 \times 3.2 = 8.0$; $14.2 > 8.0$, outlier. August: upper limit $= 4.6 + 1.5 \times 4.6 = 11.5$; $12.4 > 11.5$, outlier.

c. Mean $>$ median means positive skew: many dry days and a few very wet days.

d. August: its IQR (4.6) is larger than May\'s (3.2). A larger IQR means the middle half of the data is more spread out.

## Answer to Question 19

a. $\frac{120}{1200} = \frac{1}{10}$.

b. Boys: 14, 13, 15, 18. Girls: 16, 17, 15, 12. Total $= 60 + 60 = 120$.

c. A simple random sample could by chance miss or over-represent a group; stratified sampling guarantees that each group is in proportion.

d. Number the 150 Year 12 girls and use random numbers (or a calculator) to select 15 different numbers.

## Answer to Question 20

a. Use the formulas $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$, $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$ and $S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$ with $n = 10$. $S_{xx} = 300 - \frac{50^2}{10} = 50$.

$S_{xy} = 640 - \frac{50 \times 120}{10} = 40$. $S_{yy} = 1480 - \frac{120^2}{10} = 40$. These summary values are used in the next two parts.

b. The gradient is $b = \frac{S_{xy}}{S_{xx}}$, and the line passes through the mean point $(\bar{x}, \bar{y})$.

$b = \frac{S_{xy}}{S_{xx}} = \frac{40}{50} = 0.8$. $\bar{x} = 5$, $\bar{y} = 12$, so $a = 12 - 0.8 \times 5 = 8$. Line: $y = 8 + 0.8x$. Check: when $x = 5$, $y = 8 + 4 = 12 = \bar{y}$ ✓.

c. Use the formula for the product-moment correlation coefficient.

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{40}{\sqrt{2000}} =$ **0.894**. This shows strong positive linear correlation between $x$ and $y$.

d. Substitute $x = 6$ into the regression line. $y = 8 + 0.8 \times 6 =$ **12.8**. Since $x = 6$ is close to the mean of the data, this estimate should be reliable.

## Answer to Question 21

a. The combined total is the sum of the two group totals, $n\bar{x}$ for each group. Total $= 20 \times 60 + 30 \times 70 = 3300$; mean $= \frac{3300}{50} =$ **66**. The combined mean is 66,

closer to 70 because group B is larger.

b. Rearrange $\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2$ to get $\sum x^2 = n(\sigma^2 + \bar{x}^2)$ for each group, then add.
Group A: $20(25 + 3600) = 72\,500$. Group B: $30(64 + 4900) = 148\,920$. Combined $\sum x^2 = 221\,420$. These are the sums of squares of all the scores.

c. Use $\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2$ for the combined group of 50 students. Variance $= \frac{221420}{50} - 66^2 = 4428.4 - 4356 = 72.4$. Standard deviation $= \sqrt{72.4} =$ **8.51**.
The combined standard deviation (8.51) is larger than either group's, because the two group means differ by 10.