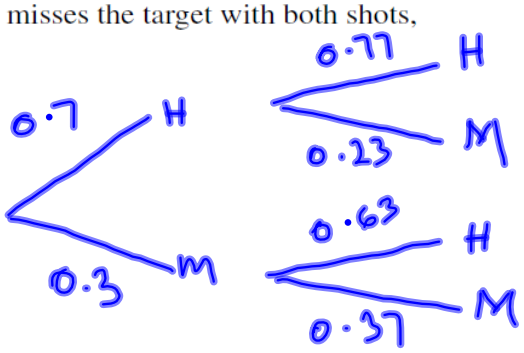
1 Siân shoots two arrows at a target. The probability that her first shot hits the target is 0.7. If her first shot hits the target then the probability that her second shot hits the target is 0.77. If her first shot misses the target then the probability that her second shot hits the target is 0.63. Find the probability that

(i) Siân misses the target with both shots, [2]



$P(\text{Miss, Miss}) = 0.3 \times 0.37 = 0.111$ ✓

(ii) Siân hits the target exactly once in her two shots. [2]

$P(1 \text{ hit}) = (0.7 \times 0.23) + (0.3 \times 0.63)$

$= \frac{7}{20}$ ✓

2 A hockey coach has to choose a team of 11 players from a group of 11 men and 11 women.

(i) If there is no restriction on the number of team members of each gender, find in how many different ways the coach can choose the team. [2]

22C11 = 705432 ✓

(ii) If the team is chosen at random from the group, find the probability that it consists of 6 men and 5 women. [4]

11C6 × 11C5

= 213444

$\frac{213441}{705432} = 0.303$ (3sf) ✓

3 A company employs a large number of people in its city office and records show that 35% of the employees live outside the city limits. The Finance Department employs 18 men and 23 women. The number of these men who live outside the city limits is denoted by X and the number of these women who live outside the city limits is denoted by Y.

(i) Assuming a binomial model, find $P(6 \leqslant X \leqslant 10)$. [3]

$X \sim B(18, 0.35)$

$P(6 \leq X \leq 10) = P(X \leq 10) - P(X \leq 5)$

$= 0.9788 - 0.3550$

$= 0.6238$ ✓

(ii) Assuming a binomial model, find $P(Y = 10)$. [3]

$Y \sim B(23, 0.35)$

$P(Y=10) = \binom{23}{10} \times 0.35^{10} \times 0.65^{13}$

$= 0.1167$

(iii) Give a reason why binomial models might not be suitable. [1]

It might be a biased sample.

Some employees may share houses so the probability that any employee lives outside the city may not be independent.

4 The table below shows the mean GCSE score and A Level mathematics grade of 6 randomly chosen students. The mean GCSE score for a student can take values from 0 to 8, with 8 representing the highest performance and 0 representing the lowest performance. The A Level grades are A, B, C, D, E and U, with A representing the highest performance and U the lowest performance.

Student	1	2	3	4	5	6
Mean GCSE score	4.1	5.2	6.4	5.3	7.4	3.6
A Level grade	E	C	D	B	A	U
RANK GCSE	5	4	2	3	1	6

(i) Calculate Spearman's rank correlation coefficient for the data. [4]

RANK ALEVEL	5	3	4	2	1	6
d	0	1	-2	1	0	0
d^2	0	1	4	1	0	0

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{(6\times 6)}{6(35)}$$

$$= 0.8285714$$

$$= 0.8286 \checkmark \quad (4)$$

For the same 6 students Spearman's rank correlation coefficient between their GCSE mathematics grade and their A Level mathematics grade was 0.943. The head teacher of a school wishes to use **either** the mean GCSE score **or** the GCSE mathematics grade as a predictor of students' A Level mathematics grades.

(ii) On the basis of the values of Spearman's rank correlation coefficient for these 6 students, state, giving a reason, which should be used. [2]

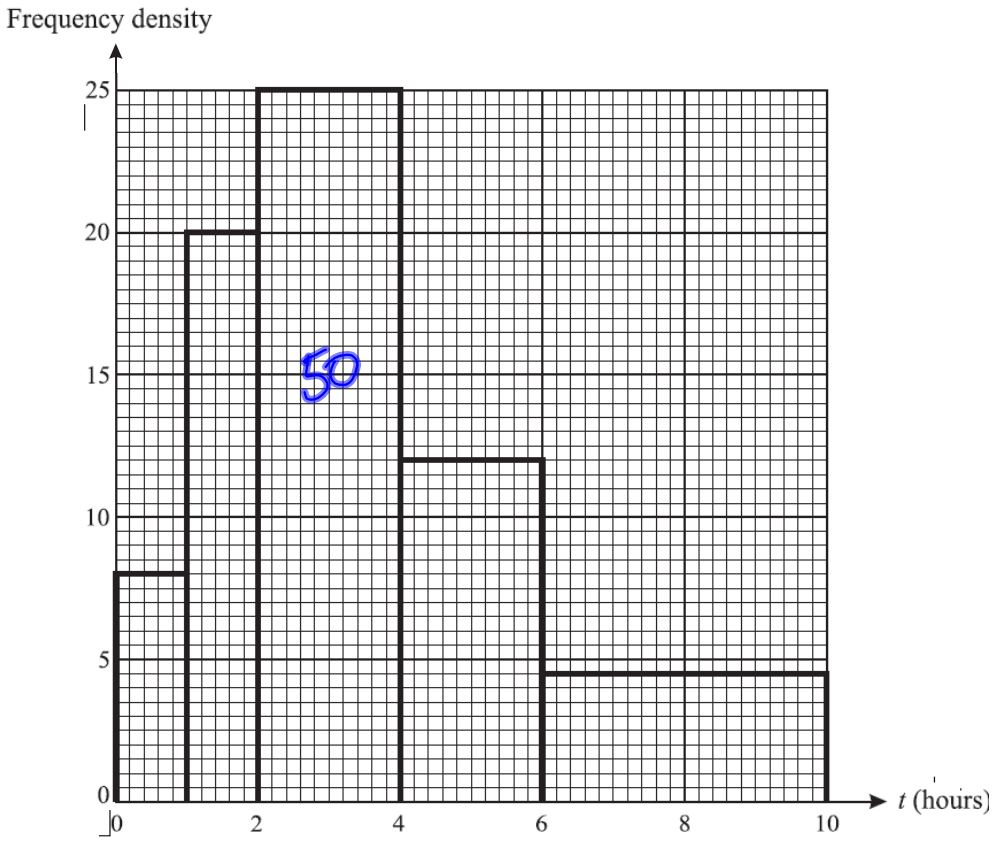
The GCSE maths grade ✓ is a better predictor of A-level performance because the $r_s = 0.943$ as compared to 0.8286, it has greater correlation ✓ (2)

(iii) Give a reason why the choice of predictor might be wrong. [1]

Correlation does not imply causation, although we should be reasonably confident that past performance in maths will predict future performance in maths.

MARKSCHEME: Sample too small to generalise

5 The owner of a car repair shop conducted a survey into the length of time that cars were in his repair shop undergoing repairs before being returned to their owners. He measured the time, t hours, that it took for each of a sample of 120 cars in his shop to be repaired and returned to their owners. The data are illustrated in the histogram below. 50 cars were repaired in the period $2 \leqslant t < 4$.



(i) Estimate the mean repair time for the sample of 120 cars. [4]

TIME	width	fd	f	mid	midxf	x^2	x^2f
$0 \le t < 1$	1	8	8	0.5	4	0.25	2
$1 \le t < 2$	1	20	20	1.5	30	2.25	45
$2 \le t < 4$	2	25	50	3	150	9	450
$4 \le t < 6$	2	12	24	5	120	25	600
$6 \le t < 10$	4	4.5	18	8	144	64	1152
			120		448		2249

$f = fd \times w$

$\bar{x} = \frac{448}{120} = 3.73$

$= 3.73$ ✓ (3sf)

4

(ii) Estimate the standard deviation of the repair times for the sample of 120 cars. [3]

$$\sigma^2 = \frac{\Sigma x^2 f}{\Sigma f} - \bar{x}^2$$

$$= \frac{2249}{120} - 3.73^2$$

$$= 4.803888889$$

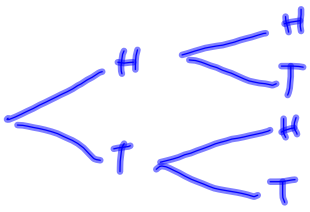
$$\sigma = \sqrt{4.803888889}$$

$= 2.19$ hrs (3sf) ✓

3

6 Alfie and Betty play rounds of a game by each tossing two unbiased coins. A round results in a 'matching' if **either** they both obtain two heads, **or** they both obtain two tails **or** they both obtain a head and a tail.

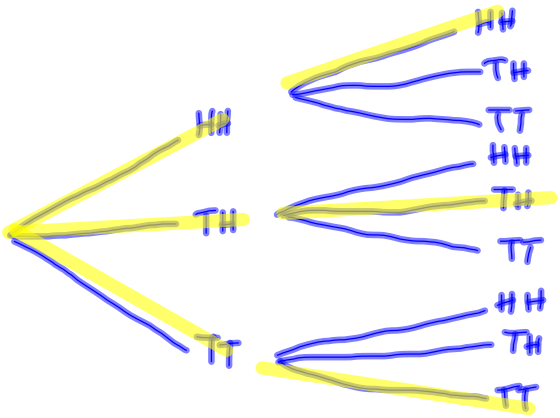
(i) Show that the probability that the first round results in a matching is $\frac{3}{8}$. [3]



$P(HH) = \frac{1}{4}$

$P(TH \text{ or } HT) = \frac{1}{2}$

$P(TT) = \frac{1}{4}$



(3)

$P(\text{matching}) = \left(\frac{1}{4} \times \frac{1}{4}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{4} \times \frac{1}{4}\right)$

$= \frac{1}{16} + \frac{4}{16} + \frac{1}{16} = \frac{6}{16} = \frac{3}{8}$ ✓

If the first round does not result in a matching, they continue to play rounds until a matching is obtained.

(ii) Find the probability that they play a total of 3 rounds or more. [4]

$X \sim Geo\left(\frac{3}{8}\right)$

(4)

$P(X > 2) = \left(\frac{5}{8}\right)^2 = 0.390625$

$= 0.391 \text{ (3sf)}$ ✓

(iii) Find the expectation of the total number of rounds played, and give its meaning in the context of the question. [2]

$E(X) = \frac{1}{3/8} = \frac{8}{3} = 2.\dot{6}$ ✓ (1)

The expected number of rounds is between 2 and 3.

MARK SCHEME : refer to average number of matchings

7 In a science lesson, Shivani measured the masses of 20 small objects. Their masses, in grams to the nearest 0.1 g, are given below, in numerical order.

1.6	2.2	3.0	3.1	3.4	4.1	4.2	4.7	4.7	4.8
5.2	5.3	5.3	5.4	5.5	5.8	5.8	5.8	5.9	6.1

(i) Find the median and interquartile range of this set of data. [3]

median = $\frac{4.8+5.2}{2} = 5.0$ ✓

LQ = $\frac{4.1+3.4}{2} = 3.75$ ✓

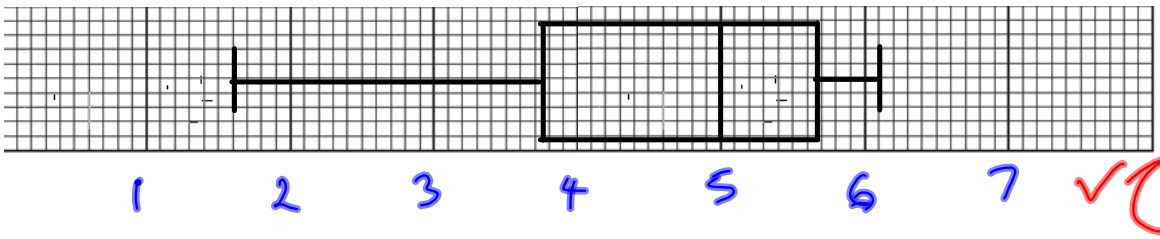
UQ = $\frac{5.5+5.8}{2} = 5.65$ ✓

IQR = $5.65 - 3.75$

$= 1.9$ ✓

3

(ii) Draw, on graph paper, a box-and-whisker plot of the data. [3]



✓ 3

During the same science lesson Emma also weighed 20 small objects, correct to the nearest 0.1 g, and obtained the following values:

lower quartile 4.5 g, median 5.1 g, upper quartile 6.0 g.

(iii) Compare Shivani's and Emma's data, referring both to the shapes of the distributions and to their variability. [3]

The medians are very close to each other.

Emma's IQR = $6 - 4.5 = 1.5$ ✓

Emma's IQR is slightly less than Shivani's indicating that her data are more consistent. ✓

Emma's data are NEGATIVELY or LEFT SKEW ✓

Shivani's data are POSITIVELY or RIGHT SKEW ✓ 3

DEFINITIONS:

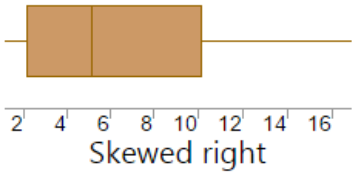
NEGATIVE SKEW - long tail left

POSITIVE SKEW - long tail right

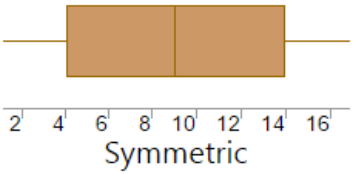
Positive Skew
RIGHT SKEW

Negative Skew
LEFT SKEW

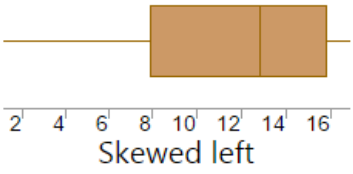
Boxplots often provide information about the shape of a data set. The examples below show some common patterns.



Skewed right



Symmetric



Skewed left

Each of the above boxplots illustrates a different skewness pattern. If most of the observations are concentrated on the low end of the scale, the distribution is skewed right; and vice versa. If a distribution is symmetric, the observations will be evenly split at the median, as shown above in the middle figure.

COURSE BOOK: Measures of Spread p.50-51

Another important feature of a set of data is its shape when represented as a frequency diagram. The three pictures in Fig. 3.9 show three different shapes which commonly occur when you draw frequency diagrams.

The distribution in Fig. 3.9a is symmetrical. If a distribution lacks symmetry it is said to be **skew**, or to have **skewness**. The distribution in Fig. 3.9a may therefore be said to have zero skewness. You were briefly introduced to the term 'skewness' in Section 2.9.

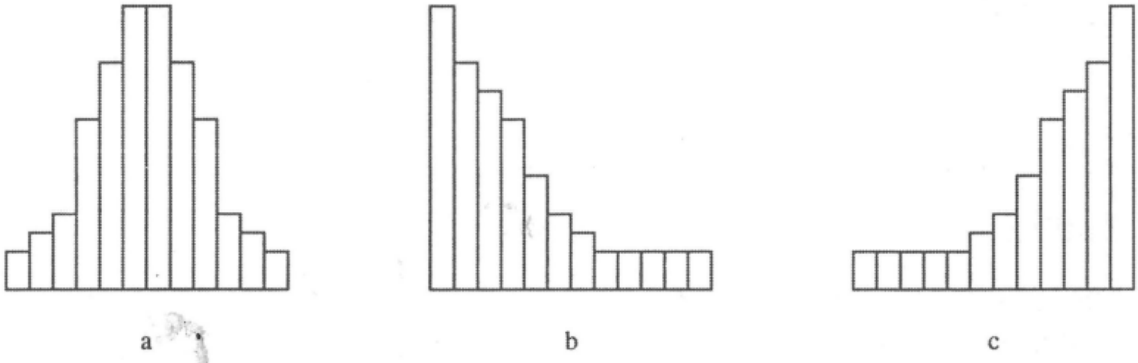


Fig. 3.9. Possible shapes of frequency distributions.

The distribution in Fig. 3.9b is certainly not symmetrical, and there is a 'tail' which stretches towards the higher values. This distribution is said to have **positive skew**, or to be **skewed positively**.

The distribution in Fig. 3.9c is also not symmetrical. There is a 'tail' which stretches towards the lower values. This distribution is said to have **negative skew**, or to be **skewed negatively**.

Another method of assessing the skewness of a distribution is to use the quartiles Q_1, Q_2 and Q_3. Remember that Q_2 denotes the median, and that Q_1 and Q_3 denote the lower and upper quartiles.

If $Q_3 - Q_2 \approx Q_2 - Q_1$ then the distribution is said to be (almost) symmetrical, and a box plot of such data, as in Fig. 3.10, would show a box in which the line corresponding to the median was in the centre of the box.

Fig. 3.10. Box plot for a set of data in which $Q_3 - Q_2 \approx Q_2 - Q_1$.

If $Q_3 - Q_2 > Q_2 - Q_1$, as in Fig. 3.11, then the data is said to have positive skew, and the line representing the median would be nearer to the left side of the box.

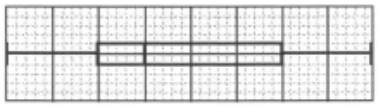


Fig. 3.11. Box plot for a set of data in which $Q_3 - Q_2 > Q_2 - Q_1$.

If $Q_3 - Q_2 < Q_2 - Q_1$, as in Fig. 3.12, then the data would be said to have negative skew, and the line representing the median would be nearer to the right side of the box.

For the data in Example 3.4.1, $Q_1 = 2$, $Q_2 = 4$ and $Q_3 = 8$, so $Q_3 - Q_2 = 8 - 4 = 4$ and $Q_2 - Q_1 = 4 - 2 = 2$. The set of data in Example 3.4.1 therefore has positive skew.

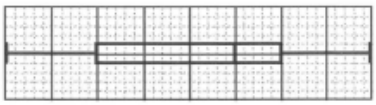


Fig. 3.12. Box plot for a set of data in which $Q_3 - Q_2 < Q_2 - Q_1$.

The length of the whiskers can also give some indication of skewness. If the left whisker is shorter than the right whisker then that would tend to indicate positive skew, whereas if the right whisker is shorter than the left whisker, negative skew would be implied.

8 A teacher decided to investigate the connection between students' performances in History and Geography. She selected 5 students at random and recorded their scores, x and y, in History and Geography tests respectively. She found that, for this set of data, the regression line of y on x had equation $y = 18.5 + 0.1x$ and the regression line of x on y had equation $x = 16.6 + 0.4y$.

(i) Using these equations, calculate the values of the mean score in the History test and the mean score in the Geography test for the 5 students. [3]

$a = \bar{y} - b\bar{x}$ $\qquad a' = \bar{x} - b'\bar{y}$

$18.5 = \bar{y} - 0.1\bar{x}$ ① $\qquad 16.6 = \bar{x} - 0.4\bar{y}$ ②

① × 10 $\qquad 185 = 10\bar{y} - \bar{x}$

② + $\qquad 16.6 = -0.4\bar{y} + \bar{x}$

$201.6 = 9.6\bar{y}$

$\bar{y} = \frac{201.6}{9.6} = 21$

substitute into ①

$18.5 = 21 - 0.1\bar{x}$

$0.1\bar{x} = 21 - 18.5$

$-0.1\bar{x} = 2.5$

$\bar{x} = 25$

check with ② $\qquad 16.6 = \bar{x} - 0.4\bar{y}$ $\qquad \bar{y} = 21$ ✓

$16.6 = 25 - (0.4 \times 21)$ $\qquad \bar{x} = 25$ ✓

$= 16.6$ ✓

NOTE: a quicker method is to use the fact that the 2 regression lines cross at (mean x, mean y)

so $y = 18.5 + 0.1(16.6 + 0.4y)$

$= 18.5 + 1.66 + 0.04y$

$0.96y = 20.16$

$y = \frac{20.16}{0.96} = 21$ etc ....

(ii) Hence show that $\Sigma x = 125$ and $\Sigma y = 105$. [1]

$\Sigma x = 25 \times 5 = 125$ ✓ $\Sigma y = 21 \times 5 = 105$ ✓ (1)

It is given that $\Sigma x^2 = 3215$, $\Sigma y^2 = 2227.5$ and $\Sigma xy = 2634$.

(iii) Obtain the product moment correlation coefficient for the data. [3]

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$= \frac{9}{\sqrt{90 \times 22.5}}$$

$$= 0.2$$

$r = 0.2$ ✓ (3)

$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$$

$$= 2634 - \frac{125 \times 105}{5}$$

$$= 9$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$= 3215 - \frac{125^2}{5}$$

$$= 90$$

$$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$= 2227.5 - \frac{105^2}{5}$$

$$= 22.5$$

(iv) The Geography score of a sixth student was mislaid but his History score was known to be 26. Use one of the equations to estimate this student's Geography score, and give a reason for the use of the chosen equation. [3]

$y = 18.5 + 0.1x$ ✓

$= 18.5 + (0.1 \times 26)$

$= 21.1$ ✓

(3)

In this system of equations the 'y' values represent the Geography score so this is what we need to calculate. ✓

(v) Comment on the reliability of your estimate. [1]

This is unlikely to be a good estimate because the product moment correlation coefficient ($r = 0.2$) indicates a very weak link between the two values. ✓

(1)