Measures of Spread

Lesson objectives

- I can describe the variability in a sample or population using measures of spread
- I can calculate the range
- I understand how to use quartiles and percentiles to analyse data

1.1

Lesson objectives | Teachers' notes | Lesson notes

MHR Page 275 #s 1 - 8

Definitions

Percentile

- The percent of all the data that are less than or equal to a specific value

Quartiles

- Three points that divide the data set into four equal groups
- The first quartile (Q1) is the middle number between the smallest number and the median; it is also the 25^{th} percentile
- The second quartile (Q2) is the median of the data set; it is also called the 50^{th} percentile
- The third quartile (Q3) is the middle number between the median and the largest number in a data set; it is also the 75^{th} percentile

Range

- The difference between the highest value and the lowest value of a data set
- Range = highest value - lowest value

Interquartile Range (IQR)

- The difference between the first and third quartiles
- IQR = Q3 - Q1

To help analyse the spread of data, you may need to identify the percentile rank or calculate percentiles.

Percentile Rank

$$R = \frac{p}{100}(n+1)$$

where p is the percentile, n is the size of the population, and R is the whole number rank of the data point. If R is not a whole number, round R down.

Percentile

$$p = 100\frac{(L+0.5E)}{n}$$

where p is the percentile, L is the number of data less than the data point, E is the number of data equal to the data point, and n is the size of the population.

Example 1

Percentiles

The list shows the marks for 25 students on a recent test out of 40.

31 28 28 30 20 25 38 40 26 28 15 21 28
36 25 16 21 34 37 30 23 24 36 32 25

a) Calculate the 80th percentile.
b) What percentile is a mark of 25?
c) What percentile is a mark of 40?

First, order the data

15, 16, 20, 21, 21, 23, 24, 25, 25, 25, 26, 28, 28, 28, 28, 30, 30, 31, 32, 34, 36, 36, 37, 38, 40

$$R = \frac{p}{100}(n+1)$$

$R = \frac{80}{100}(25+1)$

$R = 0.8(26)$

$R = 20.8$

Round down

Find midpoint of $20^{TH} + 21^{ST}$

80^{TH} percentile $= \frac{34+36}{2} = 35$

80% below 35

$$p = 100\frac{(L+0.5E)}{n}$$

$L = 7, E = 3$

$p = \frac{100(7+0.5(3))}{25}$

$p = 34$

A mark of 25 is in the 34^{TH} percentile

$\Rightarrow$ 34% of data below 25.

$$p = 100\frac{(L+0.5E)}{n}$$

$L = 24, E = 1$

$p = \frac{100(24+0.5(1))}{25}$

$p = 98$

A mark of 40 is in the 98^{TH} percentile

$\Rightarrow$ 98% of data below 40.

Your Turn

The mean playing times per game for the 22 hockey players on a team are given.

16.4, 18.3, 21.7, 18.5, 9.2, 17.9, 12.0, 15.2, 23.4, 20.5, 16.7, 13.4, 8.3, 17.9, 22.6, 18.1, 21.7, 14.6, 13.8, 24.3, 12.4, 17.4

a) Determine the 40th and 95th percentiles.

b) Determine the percentile rank of the player who averaged
   i) 9.2 min per game
   ii) 21.7 min per game
   iii) 18.1 min per game

First, order the data

8.3, 9.2, 12.0, 12.4, 13.4, 13.8, 14.6, 15.2, 16.4, 16.7, 17.4, 17.9, 17.9, 18.1, 18.3, 18.5, 20.5, 21.7, 21.7, 22.6, 23.4, 24.3

a) $R = \frac{p}{100}(n+1)$

$R = \frac{40}{100}(22+1)$

$R = 0.4(23)$

$R = 9.2$

Round down

Find midpoint of 9TH + 10TH

40^{TH} percentile $= \frac{16.4+16.7}{2}$

$= 16.55$

40% below 16.55

$R = \frac{p}{100}(n+1)$

$R = \frac{95}{100}(22+1)$

$R = 0.95(23)$

$R = 21.85$

Round down

Find midpoint of $21^{ST} + 22^{ND}$

95^{TH} percentile $= \frac{23.4+24.3}{2}$

$= 23.85$

95% below 23.85

Your Turn

The mean playing times per game for the 22 hockey players on a team are given.

16.4, 18.3, 21.7, 18.5, 9.2, 17.9, 12.0, 15.2, 23.4, 20.5, 16.7, 13.4, 8.3, 17.9, 22.6, 18.1, 21.7, 14.6, 13.8, 24.3, 12.4, 17.4

a) Determine the 40th and 95th percentiles.

b) Determine the percentile rank of the player who averaged
   i) 9.2 min per game
   ii) 21.7 min per game
   iii) 18.1 min per game

First, order the data

8.3, 9.2, 12.0, 12.4, 13.4, 13.8, 14.6, 15.2, 16.4, 16.7, 17.4, 17.9, 17.9, 18.1, 18.3, 18.5, 20.5, 21.7, 21.7, 22.6, 23.4, 24.3

b) (i) $p = 100\frac{(L+0.5E)}{n}$

$L = 1, E = 1$

$p = 100\left(\frac{1+0.5(1)}{22}\right)$

$p = 6.8\overline{1}$

9.2 minutes is the 7th percentile

$\Rightarrow$ 7% of the times were below 9.2 mins

(ii) $p = 100\frac{(L+0.5E)}{n}$

$L = 17, E = 2$

$p = 100\left(\frac{17+0.5(2)}{22}\right)$

$p = 81.8\overline{1}$

21.7 minutes is the 82ND percentile

$\Rightarrow$ 82% of the times were below 21.7 mins

(iii) $p = 100\frac{(L+0.5E)}{n}$

$L = 13, E = 1$

$p = 100\left(\frac{13+0.5(1)}{22}\right)$

$p = 61.\overline{36}$

18.1 minutes is the 62ND percentile

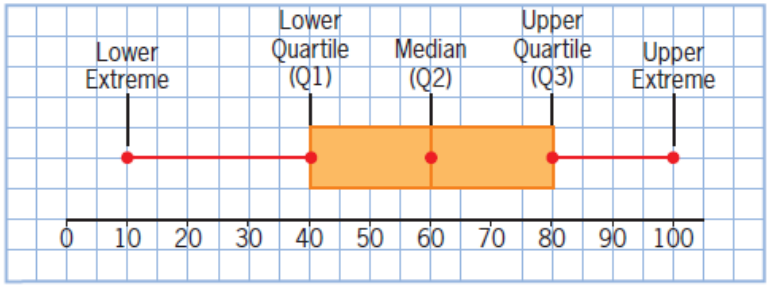
$\Rightarrow$ 62% of the times were below 18.1 minutes

To better understand the variability of a data set, you can use a variety of measures of spread. You can use a box and whisker plot to visually demonstrate the spread of a distribution along a number line.

To construct a box and whisker plot:

- Draw a rectangle whose ends are the first (lower) and third (upper) **quartiles**.
- Draw the median within the rectangle.
- Add "whiskers," which are horizontal line segments connecting the box to the extremes of the data, covering the entire **range**.

Each of the four zones illustrated by a box and whisker plot contains 25% of the data. The difference between the first and the third quartiles is known as the **interquartile range (IQR)**. The interquartile range represents the "middle half" of the data.



Example 2

Interquartile Range and Box and Whisker Plots

The table lists the heights of the 20 girls who signed up to try out for their school basketball team.

Height (cm)	Frequency	Cumulative Frequency
155–160	1	1
160–165	3	4
165–170	4	8
170–175	7	15
175–180	3	18
180–185	1	19
185–190	0	19
190–195	1	20

a) Determine the median, range, first and third quartiles, and interquartile range. Create a box and whisker plot of the data.
b) Describe the data in each zone of the plot.
c) Identify any outliers, if they exist.

a) Median position $= \frac{20+1}{2} = 10.5$
$\Rightarrow$ Between $10^{TH} + 11^{TH}$
$\Rightarrow$ 170 - 175 cm
Median is 172.5 cm

Range = Highest - Lowest
$= 192.5 - 157.5$
$= 35$ cm

$R = \frac{p}{100}(n+1)$

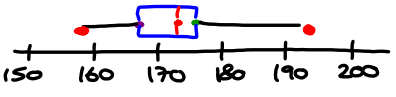
$R = \frac{25}{100}(20+1)$
$R = 5.25$
$\Rightarrow R = 5$
Midpoint of $5^{TH} + 6^{TH}$
$\Rightarrow$ 165 - 170
$Q_1 = 167.5$ cm

$R = \frac{p}{100}(n+1)$

$R = \frac{75}{100}(20+1)$
$R = 15.75$
$\Rightarrow R = 15$
Midpoint of $15^{TH} + 16^{TH}$
These happen to be on the boundaries of class intervals, so use the shared boundary. $Q_3 = 175$ cm

$IQR = Q_3 - Q_1$
$= 175 - 167.5$
$= 7.5$ cm

b) 25% of the data is in the intervals 155 - 167.5, 167.5 - 172.5, 172.5 - 175, and 175 - 195 cm



c) Outliers are values $Q_1 - 1.5(IQR)$ or $Q_3 + 1.5(IQR)$
Lower extreme
$167.5 - 1.5(7.5)$
$= 156.25$
Upper extreme
$175 + 1.5(7.5)$
$= 186.25$
No points less than 156.25
One point greater than 186.25 $\Rightarrow$ one outlier of approximately 192.5 cm

Your Turn

A summer camp activity involves measuring the distance travelled by 50 turtles in 15 min. The table shows the results.

Distance (m)	Frequency
0–5	1
5–10	0
10–15	6
15–20	12
20–25	15
25–30	5
30–35	7
35–40	1
40–45	3

1, 1, 7, 19, 34, 39, 46, 47, 50 — cumulative frequency (running total)

Median position $= \frac{50+1}{2} = 25.5$
$\Rightarrow$ Midpoint of $25^{TH}+26^{TH}$ values
$\Rightarrow$ 20–25
Median is 22.5cm
Range = Highest − Lowest
= 42.5 − 2.5
= 40m

a) Determine the median, range, first and third quartiles, and interquartile range. Make a box and whisker plot of the data.
b) Describe the data in each zone of the plot.
c) Identify any outliers, if they exist.

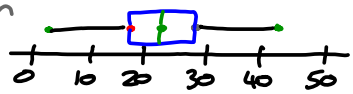
$R = \frac{p}{100}(n+1)$

$R = \frac{25}{100}(50+1)$
$R = 12.75$
$R = 12$
Midpoint of $12^{TH}+13^{TH}$
$\Rightarrow$ 15–20
$Q_1 = 17.5m$

$R = \frac{p}{100}(n+1)$

$R = \frac{75}{100}(50+1)$
$R = 38.25$
$R = 38$
Midpoint of $38^{TH}+39^{TH}$
$\Rightarrow$ 25–30
$Q_3 = 27.5m$

$IQR = Q_3 - Q_1$
$= 27.5 - 17.5$
$= 10m$



b) 25% of the data is between 2.5–17.5, 17.5–22.5, 22.5–27.5, and 27.5–42.5m

c) Lower extreme $= Q_1 - 1.5(IQR)$
$= 17.5 - 1.5(10)$
$= 2.5m$

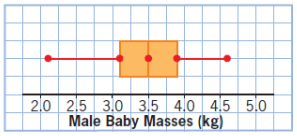
Upper extreme $= Q_3 + 1.5(IQR)$
$= 27.5 + 1.5(10)$
$= 42.5m$

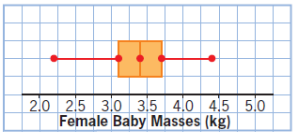
$\Rightarrow$ There are no outliers as there are no values outside these limits.

Example 3

Interpreting Quartiles

The box and whisker plots illustrate the spread of Canadian full-term male and female baby masses, in kilograms, at birth.





	Boys (kg)	Girls (kg)
Median	3.5	3.4
Range	4.6 – 2.1 = 2.5	4.4 – 2.2 = 2.2
Q1	3.1	3.1
Q3	3.9	3.7
IQR	3.9 – 3.1 = 0.8	3.7 – 3.1 = 0.6

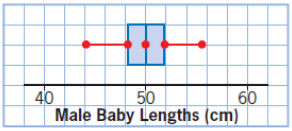
Compare the spreads of birth masses for boys and girls.

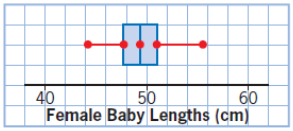
Comment on the medians, IQR, and range.

The males have a slightly higher median mass (3.5kg vs 3.4).
The males have a larger range (2.5kg vs 2.2kg).
The males have a larger IQR (0.8kg vs 0.6kg).
As the range and IQR for the males are larger, this means that their data is more spread out.

Your Turn

The box and whisker plots illustrate the spread of Canadian full-term male and female baby lengths, in centimetres, at birth. Compare the spreads of birth lengths for boys and girls.





	Boys (cm)	Girls (cm)
Median	50	49.5
Range	55.5 – 44 = 11.5	55.5 – 44.5 = 11
Q1	48.5	47.75
Q3	52	51
IQR	52 – 48.5 = 3.5	51 – 47.75 = 3.25

Comment on the medians, IQR, and range.

The median length for males was slightly longer (50cm vs 49.5 cm).

The range for males is slightly larger (11.5cm vs 11cm).

The IQR for males was again slightly larger (3.5cm vs 3.25cm)

As the range and IQR for the males is larger, this means their data is more spread out (but not by much in this particular case).

Key Concepts

- A measure of spread helps you understand how closely a set of data is clustered around its centre.
- The range is the difference between the maximum value and minimum value.
- A percentile is the percent of all the data that are less than or equal to the specific data point.
- Quartiles divide the data set into four equal parts. Q1 is the 25th percentile, Q2 is the median (or 50th) percentile, and Q3 is the 75th percentile.
- The interquartile range (IQR) is the distance between the first and third quartiles. To calculate, subtract the value for Q1 from the value for Q3. The interquartile range contains the middle 50% of the data.
- A box and whisker plot uses a rectangle to visually demonstrate the spread of the distribution along a number line by displaying the median, quartiles, and upper and lower extremes.
- An outlier exists if it is less than Q1 – 1.5 × IQR or greater than Q3 + 1.5 × IQR.

R1. Consider the information in the table regarding two data sets. Compare their spreads.

	Data Set 1	Data Set 2
Median	56.3	57.1
Min	32.1	24.2
Max	65.9	71.1
Q1	43.2	34.5
Q3	60.2	63.2

$Range_1 = 65.9 - 32.1 = 33.8$

$Range_2 = 71.1 - 24.2 = 46.9$

$IQR_1 = 60.2 - 43.2 = 17.0$

$IQR_2 = 63.2 - 34.5 = 28.7$

The median for data set 1 is 0.8 less than for data set 2.

The IQR for data set 1 is 17 and the IQR for data set 2 is 28.7.

We can see that the data is more spread out for data set 2 than it is for data set 1 as it has a larger range and IQR.

R2. What problems can occur if the range is used to measure the spread of a set of data?

The range only gives information about the extreme values, not how closely the data is clustered around its centre.

R3. What information does the interquartile range provide?

The interquartile range contains the middle 50% of the data. The smaller this value, the more closely the data is clustered around the centre.