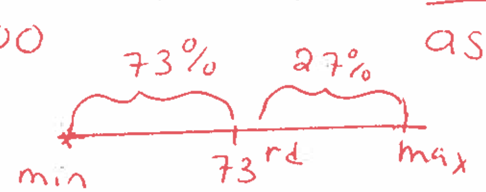
c) Interquartile Range (IQR)

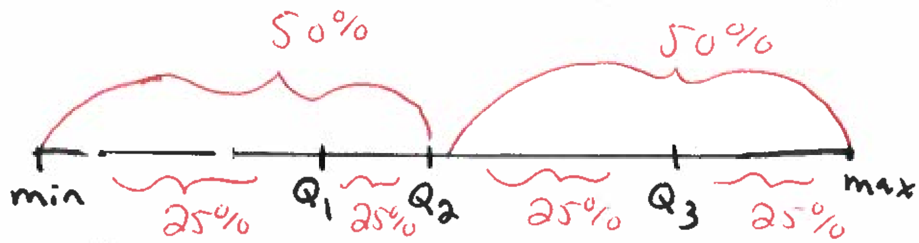
The p^{th} **percentile** of a data set is a number such that $p\%$ of the data lies at or below that value and $(100 - p)\%$ of the data lies at or above that value. $0 \leq p \leq 100$ ascending order



Example: The median is the 50^{th} percentile. Q_2

Quartiles:

- The **lower quartile** of a data set, denoted Q_1, is the 25^{th} percentile.
- The **upper quartile** of a data set, denoted Q_3, is the 75^{th} percentile.



Note: 50% of the data lies between Q_1 and Q_3.

To compute these two quartiles:

- Order the observations from smallest to largest.
- Compute the median of the data set. Call this the "overall median".
  - Q_1 is the median of those observations that lie below the overall median.
  - Q_3 is the median of those observations that lie above the overall median.

Note: The overall median of the data set is **not** included in the two halves used to compute the quartiles.

↳ matters for n odd

median

Example: $\{1, 1, 2, \cancel{3}, 5, 8, 13\}$ $n = 7$

$Q_1 = 1,\quad Q_3 = 8$

Example: $\{1, 1, 2, 3, 5, 8, 13, 21\}$ $n = 8$ median $\frac{3+5}{2} = 4$

$Q_1 = \frac{1+2}{2} = \frac{3}{2} = 1.5$ $\quad Q_3 = \frac{8+13}{2} = \frac{21}{2} = 10.5$

The **interquartile range** of a data set, denoted **IQR**, is the difference between the upper quartile and the lower quartile:

$$\mathrm{IQR} = Q_3 - Q_1$$

Example: $\{1, 1, 2, 3, 5, 8, 13\}$ $IQR = Q_3 - Q = 8 - 1 = 7$

Example: $\{1, 1, 2, 3, 5, 8, 13, 21\}$ $IQR = Q_3 - Q_1 = 10.5 - 1.5 = 9$

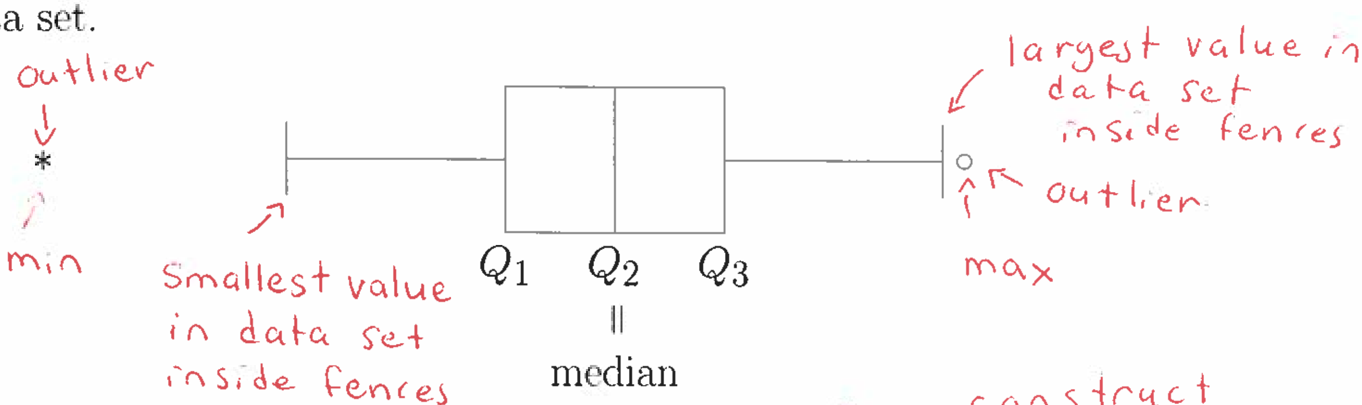
The **5-number summary** of a data set consists of:

- the minimum value
- the lower quartile Q_1
- the median
- the upper quartile Q_3
- the maximum value

Example: 5-number summaries

$\{1, 1, 2, 3, 5, 8, 13\}$	$\{1, 1, 2, 3, 5, 8, 13, 21\}$
min = 1	min = 1
Q_1 = 1	Q_1 = 1.5
median = 3	median = 4
Q_3 = 8	Q_3 = 10.5
max = 13	max = 21

Boxplot: a graphical display of a data set which uses the 5-number summary. Boxplots give us information about the center, spread, and shape (symmetry vs. skewness) of the data, and the presence of outliers in the data set.



To construct a (horizontal) boxplot: → can construct vertically

1) Calculate the median, the quartiles (Q_1 and Q_3), and the IQR for the data set.

2) Draw a horizontal line which represents the scale of measurement.

3) Above this line, draw a box with the left end at Q_1 and the right end at Q_3. Draw a vertical line through the box at the median.

4) Compute

multiplication

   i) the **lower fence** $Q_1 - 1.5(IQR)$, and

   ii) the **upper fence** $Q_3 + 1.5(IQR)$.

   The fences are not part of the boxplot display and are only used in its construction.

5) Draw vertical dotted lines on the display to mark the fences. Draw a horizontal line from the left end of the box to the smallest data value between the fences and draw a horizontal line from the right end of the box to the largest data value between the fences. These horizontal lines are called **whiskers**.

6) Mark any data value outside of the fences with a special symbol. These data values are **outliers**.

7) Remove the fences from the display.

Note: Any data value which lies

- below $Q_1 - 3(\text{IQR})$ or
- above $Q_3 + 3(\text{IQR})$

is called an **extreme outlier**. We often use different symbols to distinguish extreme outliers from the other outliers.

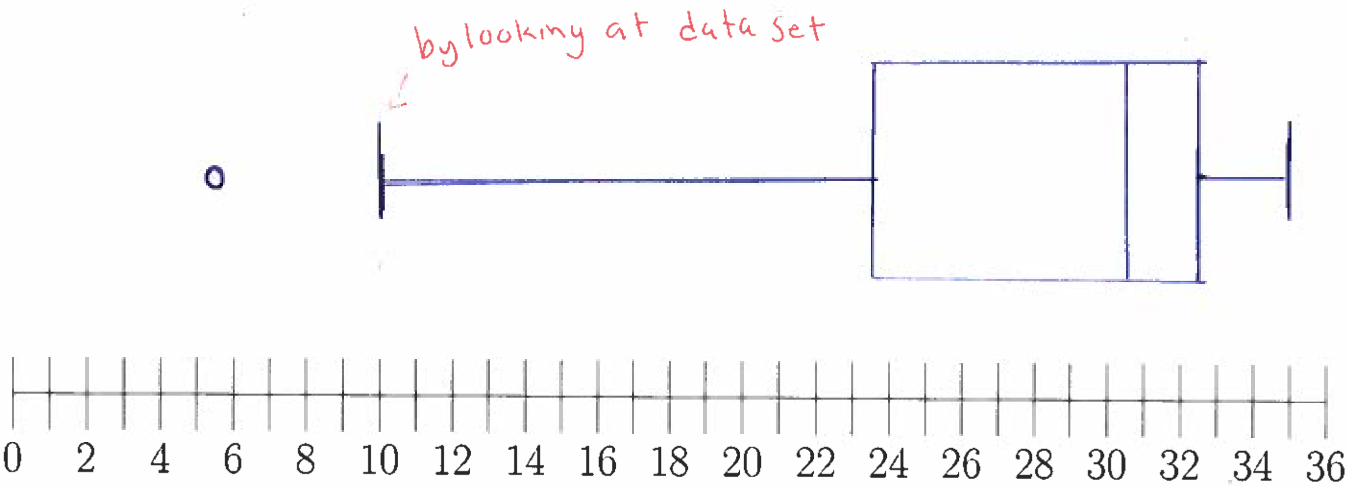
Example: A midterm exam for a U of A math class was written by 139 students and was out of 35 points. The 5-number summary for this exam was

min = 5.5 $\quad Q_1 = 23.5 \quad$ median = 30.5 $\quad Q_3 = 32.5 \quad$ max = 35

IQR = $Q_3 - Q_1 = 32.5 - 23.5 = 9$

lower fence: $Q_1 - 1.5(\text{IQR}) = 23.5 - 1.5(9) = 10$

upper fence: $Q_3 + 1.5(\text{IQR}) = 32.5 + 1.5(9) = 46$



Boxplots and Distribution Shape: From a boxplot, we can describe the shape of a distribution by looking at the position of the median line in the box (compared to Q_1 and Q_3) and the lengths of the whiskers.

For a **symmetric** distribution:

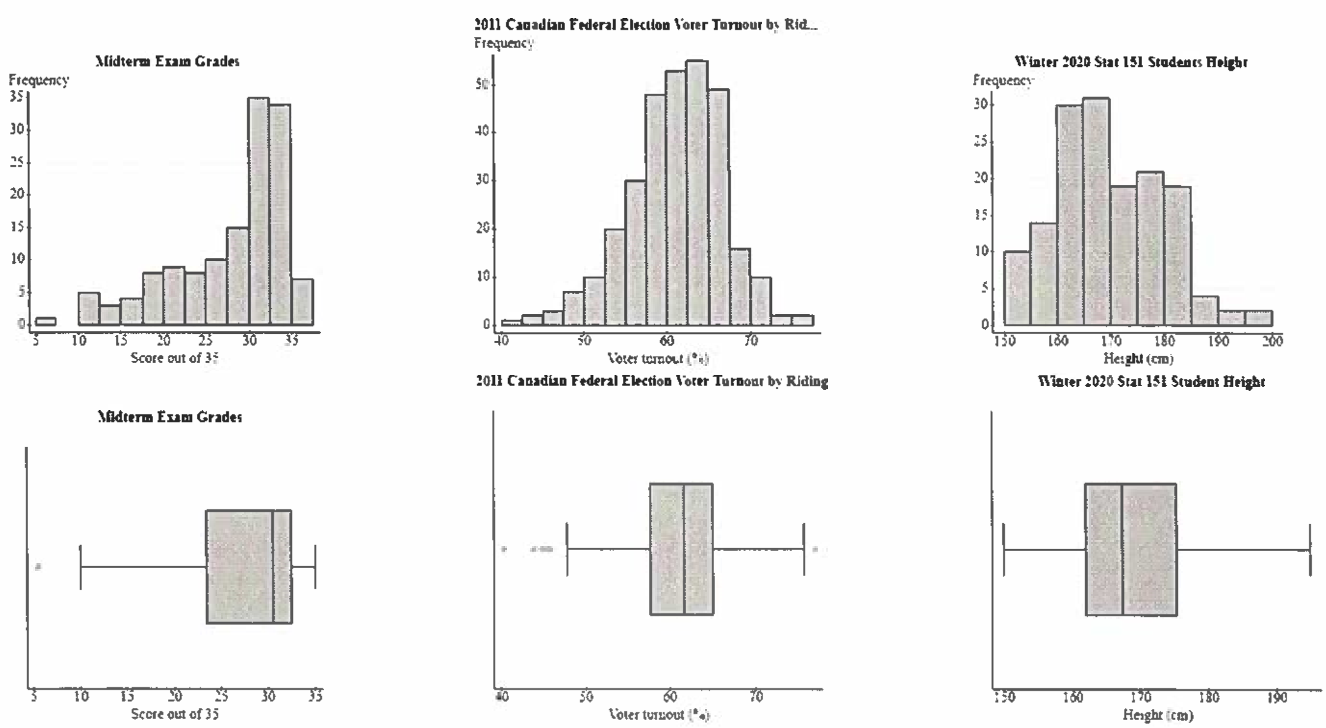
- the median line is in the centre of the box (half way between Q_1 and Q_3).
- the whiskers are the same length.

For a **left-skewed** distribution:

- the median line is right of centre (closer to Q_3 than to Q_1).
- the whisker is longer on left side of box.

For a **right-skewed** distribution:

- the median line is left of centre (closer to Q_1 than to Q_3).
- the whisker is longer on right side of box.



left-skewed

Symmetric

right-skewed (slightly)

Comparing Approaches:
Mean/ Standard Deviation Vs. Median/IQR

Note: Report the mean and standard deviation together and report the median and IQR together.

Outliers

How far an outlier lies from the centre of a data set does not affect the median and IQR, but it does affect the mean and standard deviation.

quartiles

The median and IQR are said to be **resistant** to outliers since they only consider the order of the values and not the magnitude of the values.

range

The mean and standard deviation are **not resistant** to outliers.

Example: $\{1, 2, 3, 4, 5, 6\}$ $\{1, 2, 3, 4, 5, 100\}$

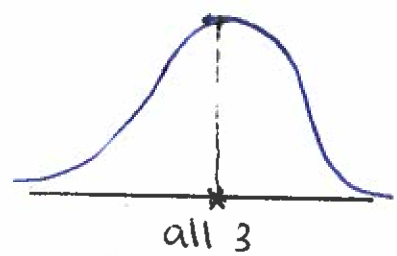
$\bar{y} = 3.5$ median $= 3.5$ $\bar{y} \approx 19.17$ median $= 3.5$

$s \approx 1.87$ IQR $= 3$ $s \approx 39.63$ IQR $= 3$

Shape vs. Centre

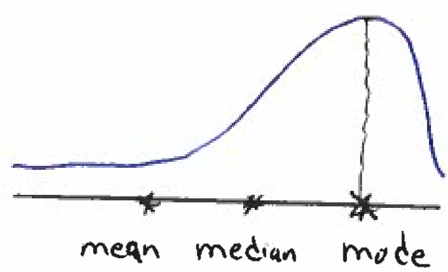
In a **symmetric** distribution

mean = median = mode



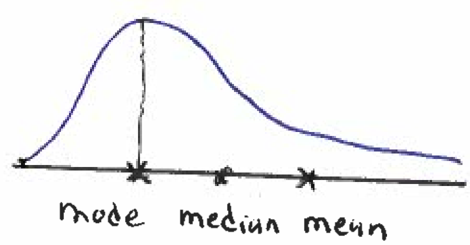
In a **left-skewed** distribution

mean < median < mode



In a **right-skewed** distribution

mode < median < mean



When summarizing a distribution, consider the shape, center, and spread of the data.

- Start by graphing the distribution and discussing its shape.

- Decide which measures of center and spread to use:
  - If the distribution is roughly **symmetric**, use the **mean** and **standard deviation**. - unimodal, usually no outliers
  - If the distribution is **skewed**, use the **median** and **IQR**.

    or has outliers

    (5-number summary + boxplot)

- Identify any outliers.
  - If there are outliers, consider using the median and IQR.
  - If there are outliers, consider computing the mean and standard deviation with and without the outliers.

    compare results

- Identify any other unusual features of the distribution:
  - gaps
  - multiple modes -> might be 2 distinct groups in data set.
    * Investigate why. Consider splitting the data into two separate groups, if you can identify a reason for the separate modes.

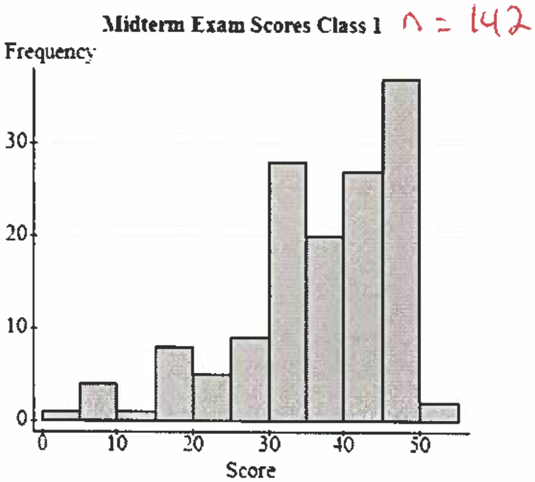
Chapter 4: Understanding and Comparing Distributions

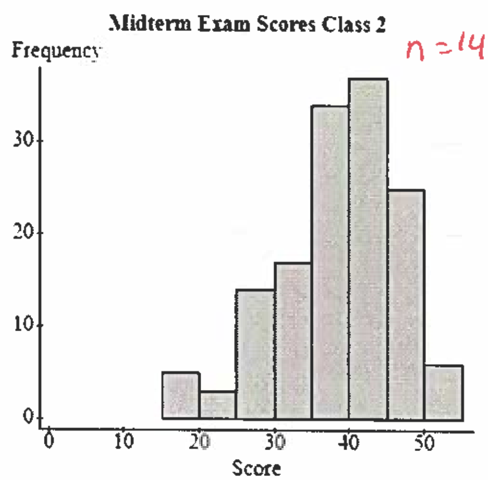
We can examine the relationship between two variables, one quantitative and one categorical, visually in several different ways: Compare groups

- Shape
- Centre
- Spread

Side-by-side-histograms: compare 2 or 3

Example: Midterm Exam Grades (out of 50)





- # of modes
- Symmetric vs. Skewness
- outliers / gaps

Back-to-back Stem-and-leaf Displays compare 2, 3

↳ 2

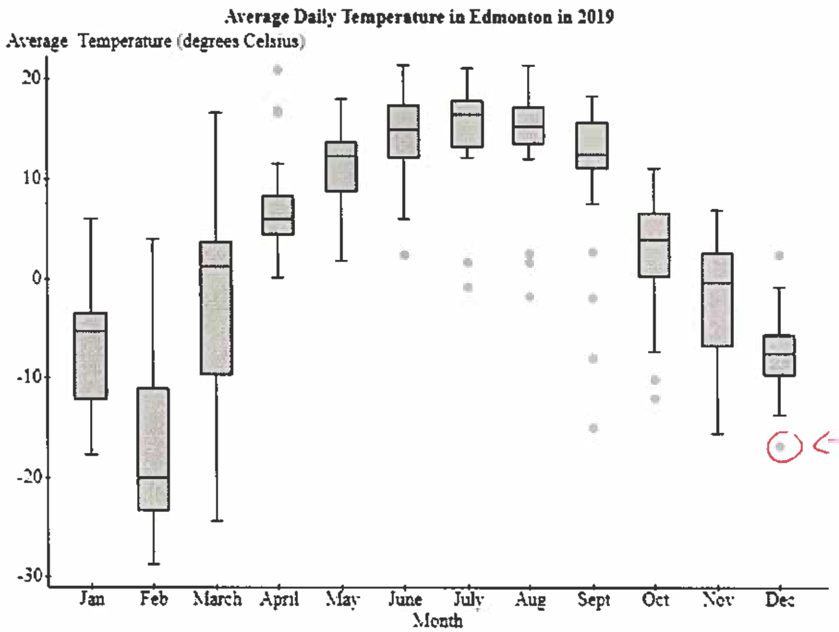
Example: 500 mile car race times (in minutes):

Make A		Make B
	180	67889
378	181	0256
2477	182	148
0	183	

(8|181|2 means 181.8 minutes for a car of Make A and 181.2 minutes for a car of Make B)

Side-by-side boxplots: compare several: shape
Centre
Spread
outliers

Example: Daily Average Temperature in Edmonton in 2019

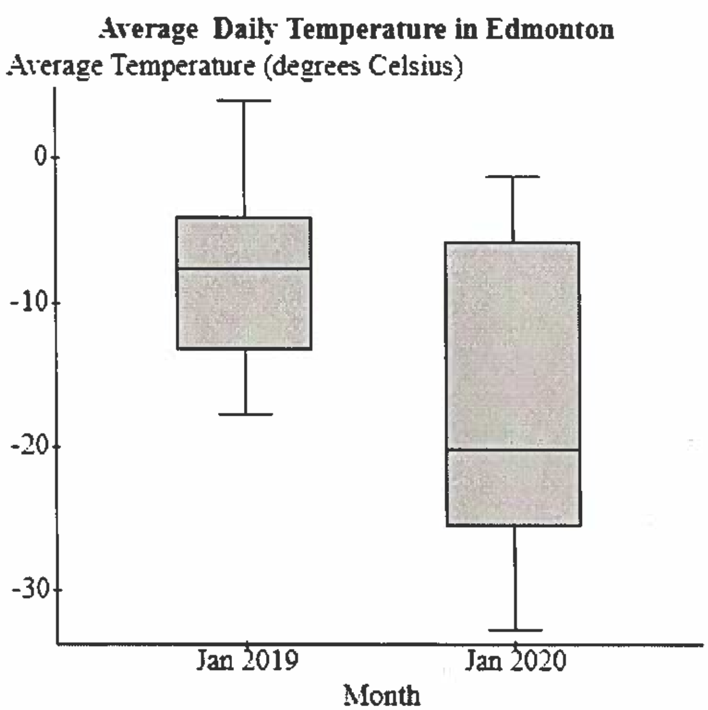


- mediuns: skewed vs symmetric

- IQR

- outliers

Example: Daily Average Temperature in Edmonton in Jan 2019/2020

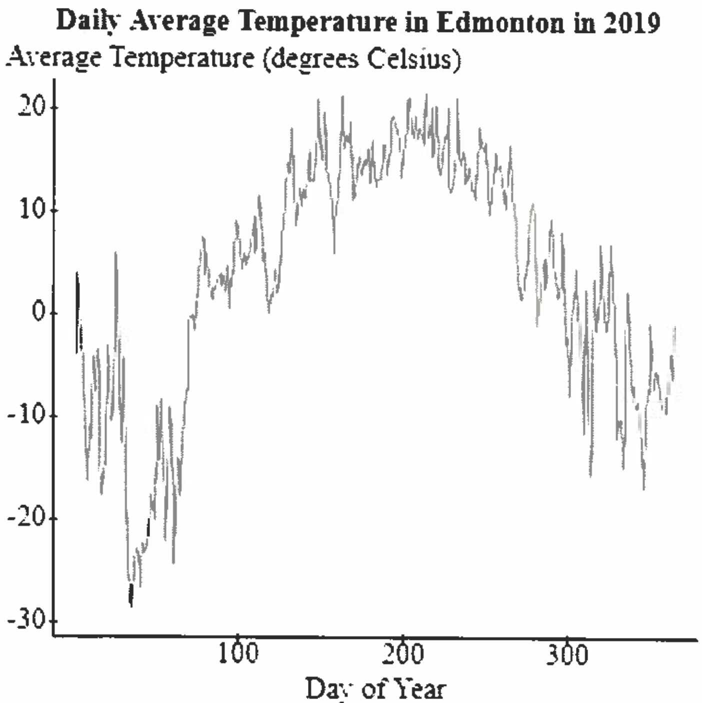


(Data from the Government of Canada Website)

Timeplot or Time Series Plot: a display of values against time.

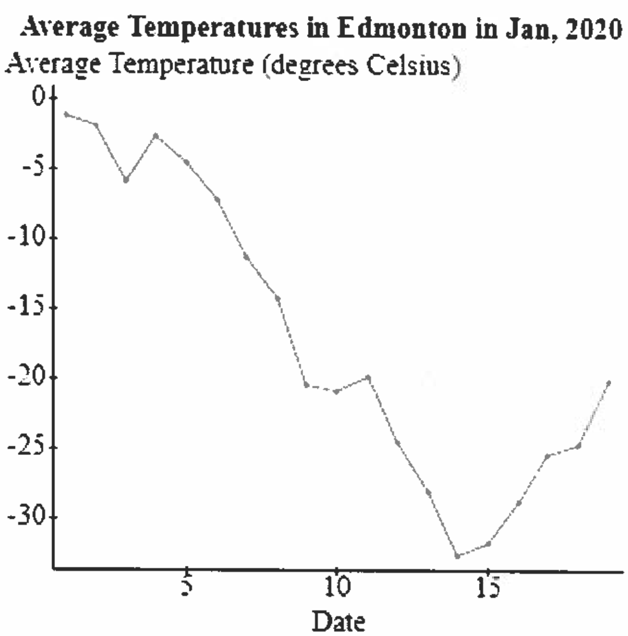
→ see how data behaves / changes over time.

Example: Daily Average Temperature in Edmonton in 2019



← time on horizontal axis

Example: Daily Average Temperature in Edmonton in Jan 2020



(Data from the Government of Canada Website)

Chapter 5: The Standard Deviation as a Ruler and the Normal Model

Section 5.2: Shifting and Scaling

the same constant for each data value.

Shifting Data: adding (or subtracting) a constant to every data value.

Example: $\{1, 2, 3, 4\}$ $\xrightarrow{y+4}$ $\{5, 6, 7, 8\}$

When we add (or subtract) a constant to every value in a data set,

a) the measures of **position** will increase (or decrease) by the value of the constant.

$c \in \mathbb{R}$

- mean
- median
- Q_1
- Q_3
- min
- max

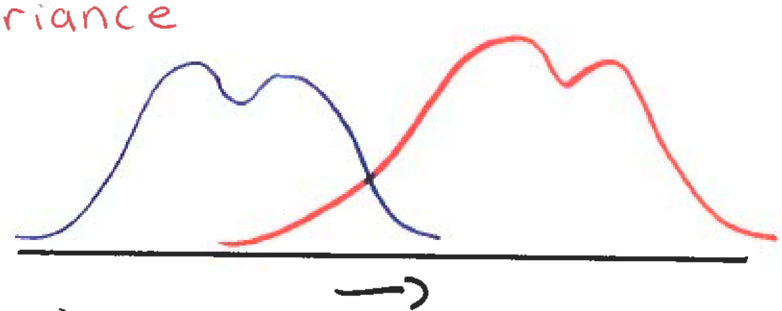
all percentiles

$+ c$

b) the measures of **spread** do **not** change.

- standard deviation, variance
- IQR
- range

c) the shape does **not** change.



Example: $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

Mean	Variance	Std. dev.	Median	Range	Min	Max	Q_1	Q_3	IQR	
5.5	9.17	3.03	5.5	9	1	10	3	8	5	y
12.5	9.17	3.03	12.5	9	8	17	10	15	5	$y+7$
0.5	9.17	3.03	0.5	9	-4	5	-2	3	5	$y-5$ $y+(-5)$