Measures of location / central tendency

In analysing data, one of the most important questions is *'what does a typical piece of data look like?'*.

What is this data?

142	175	152	202	96
201	154	195	148	139
178	142	161	179	136
146	124	133	160	132
153	169	130	125	169

Table 1: _____

What is the typical _____ without doing any calculations?

Fact — A central tendency of a dataset is a central or typical value. There are many possible central tendencies or **measures of location**.

For example:

Definition. The **(arithmetic) mean** of a set of data is the sum of all the data points divided by the number of data points

Fact — The mean can be expressed as:

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum x_i \\ &= \frac{\sum f_i x_i}{\sum f_i}\end{aligned}$$

Where x_i is either the i^{th} data point and n is the number of values or x_i is the i^{th} value of grouped data with frequency f_i.

Definition. The **median** of a set of data is the middle value once the data set has been arranged in order

Definition. The **mode** of a set of data is the most common value.

Household size	Number of households ('000s)
One person	985
Two people	1,013
Three people	642
Four people	596
Five people	242
Six people	65
Seven or more people	36
All households	3,579

Table 2: Source: Labour Force Survey (LFS), Office for National Statistics

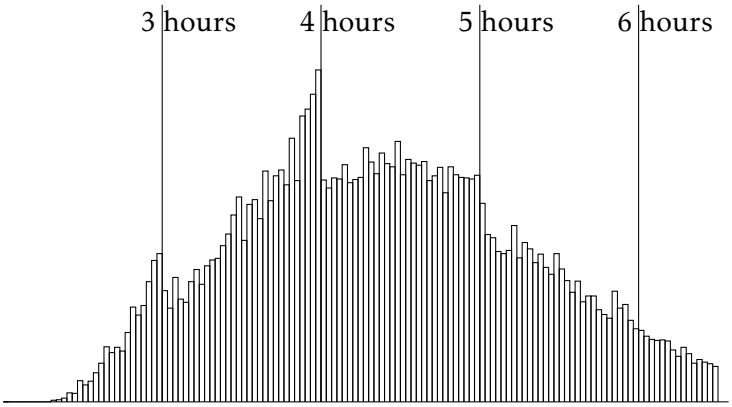
Finish time (t)	Finishers
2 hours $\leq t <$ 3 hours	2,112
3 hours $\leq t <$ 4 hours	10,565
4 hours $\leq t <$ 5 hours	13,084
5 hours $\leq t <$ 6 hours	7,226
6 hours $\leq t <$ 7 hours	2,255
7 hours $\leq t <$ 8 hours	510
8 hours $\leq t <$ 9 hours	130
9 hours $\leq t <$ 10 hours	5
	35,887

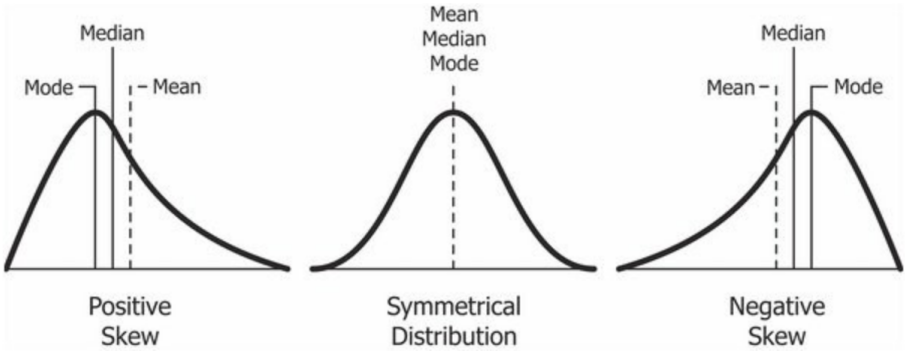
Table 3: 2022 London Marathon Finish Times

Definition. The *modal class* is the class interval which contains the modal value

Year	Name	Time
2025	Bilbo Sluggins	2:11
2024	Jeff	4:05
2023	Lettie	7:24
2019	Sammy	2:38
2018	Hosta	3:10
2017	Larry	2:47
2016	Herbie 2	3:25
2015	George	2:45
2014	Wells	3:19
2013	Racer II	2:47
2012	Racer	3:20
2011	Zoomer	3:23
2010	Sidney	3:41
2009	Terri	2:49
2008	Heikki	3:02

Table 4: Snail Racing World Championships


3 hours
4 hours
5 hours
6 hours


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

Measures of spread

Consider the following two data sets:

Show 1	60	59	60	60	60	60	61	58	60	62	60	60
Show 2	61	63	70	50	47	66	59	59	67	55	68	55

Table 5: Runtime for two Netflix shows

Both have **mean** _____ and **median** _____, although clearly the data-sets are fairly different.

Another thing we might be interested in aside from *'what does a typical piece of data look like?'* is *'how similar are the data to each other?'*

> Fact — A **measure of spread** is a number which describes how different or spread out a data set is.

Some examples might be:

Definition. The **range** of a data set is the largest value minus the smallest value

Definition. The **interquartile range (IQR)** of a data set is the upper quartile minus the lower quartile

Fact — The **variance** of a data set is:

$$\begin{aligned}\text{variance} &= \frac{1}{n}\sum (x_i - \overline{x})^2 \\ &= \frac{1}{n}\left(\sum x_i^2\right) - \overline{x}^2\end{aligned}$$

Fact — The **standard deviation** of a data set is:

$$\begin{aligned}\sigma &= \sqrt{\frac{1}{n}\sum (x_i - \overline{x})^2} \\ &= \sqrt{\frac{1}{n}\left(\sum x_i^2\right) - \overline{x}^2}\end{aligned}$$

Household size	Number of households ('000s)
One person	985
Two people	1,013
Three people	642
Four people	596
Five people	242
Six people	65
Seven or more people	36
All households	3,579

Table 6: Source: Labour Force Survey (LFS), Office for National Statistics

Finish time (t)	Finishers
2 hours $\leq t <$ 3 hours	2,112
3 hours $\leq t <$ 4 hours	10,565
4 hours $\leq t <$ 5 hours	13,084
5 hours $\leq t <$ 6 hours	7,226
6 hours $\leq t <$ 7 hours	2,255
7 hours $\leq t <$ 8 hours	510
8 hours $\leq t <$ 9 hours	130
9 hours $\leq t <$ 10 hours	5
	35,887

Table 7: 2022 London Marathon Finish Times

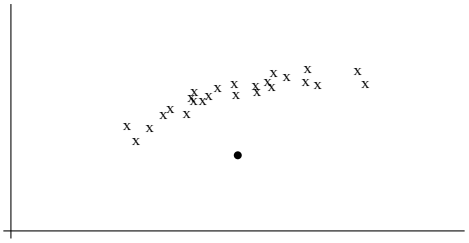
Coding of Data

Suppose you had to find the mean and standard deviation of the numbers $907, 908, 898, 902, 897$.

Fact — If $\{x_i\}$ has mean $\overline{x}$ and variance σ_x^2, and if $y_i = ax_i + b$, then

$$\overline{y} = a\overline{x} + b \text{ and } \sigma_y^2 = a^2\sigma_x^2$$

Outliers



Outliers are points which do not fit the general pattern of your data. This might be for legitimate reasons, eg _____ or illegitimate reasons eg _____.

For the purposes of OCR exam board, outliers are:

Definition. An **outlier** is any point which is $1.5 \times IQR$ away from the nearest quantile *or* 2 standard deviations away from the mean.

> Fact — Outliers **should not** be removed from data unless there is a very good reason to do so, for example negative height.