LESSON HANDOUT

Basic descriptive statistics

Descriptive statistics are measures that describe our data. We use them to make general observations and find direction for our analysis. We'll start by understanding four measures:

- The mean, the mode and the median, all of which are measures of central tendency and;
- The range, which is a measure of spread.

The mean

The mean is the average of a selection, hence why it's sometimes called the average. The mean gives us a broad measure of the central tendency of the dataset, i.e. it tells us what values looks like on average.

We calculate the mean by dividing the sum (or total) of the data set by the number of values in the dataset.

For example, to calculate the mean age for the below dataset, showing children in daycare, we would first add the numbers together (6 + 5 + 8 + 11 + 4 = 34) and then divide this total by the count of numbers in the data-set (5). Dividing 34 by 5 gives us **6.8, the average age of children in this dataset.**

Name	Age
John	6
Raivita	5
Cody	8
Sarah	11
Fred	4

The mode

The mode is the most commonly occurring figure in a data set. For example in the below data, the mode is 1 because it appears twice while all other figures appear only once:

1 1 7 8 9 0

The mode can also be applied to categorical data. For example, look at the below data set showing coffee orders during a cart's morning sales:

Order	Number of Shots
Large Flat White	2
Large Latte	2
Flat White	1
Flat White	1
Espresso	1
Flat White	1

The mode tells you that the most common order is a flat white, which appears three times.

This statistic would be far more useful for the coffee cart owner than, say, the average number of shots per order - in this case 1.3

The mean and mode in context

Taking too much stock in the average or the mode can be problematic and it would be wise to remember that: "The average breeds average results", As an example, take a look at this article from the Sydney morning herald about the average Australians in 2017.

This article gives us some interesting insights into the Australian population as a whole, in some cases mistaking the average for the mode. The article derives averages and modes from census data to paint the average Australian as a Female, named Rebecca, White, 38 years old, married with two kids, year 12 educated and working as a sales assistant.

Knowing this, should we be surprised that the next Australian we meet isn't a working, middle-aged mum?

Of course not, and this is a good way to think about how far an average can take you in general. They provide an insight into the nature of the whole but have real limitations, especially when applied to large data-sets.

The median

The median is the middle value, when the data set is ordered, normally from smallest to largest.

The median value is often used alongside the mean to understand the distribution of the data (or how the data is spread) and the median is also a good way to control for the effect of outliers.

To find the median, we first order values from smallest to largest and then select the middle value. In the case of the below dataset showing peoples heights, the median is Paulo.

Name	Height (cm)
Savannah	120
Hannah	125
Paulo	130
Astrid	140
John	225

In this case, the Median of 130 cm is far more representative of most peoples heights than the average of 148 cm.

All people sit within 10cm of the median except for John who is far tall taller and who would be considered an outlier - a value that lies outside of the other values.

On the other hand, only Astrid sits within 10 cm of the mean, making it a far less representative statistic.

Its the inclusion of John, the outlier, in this data-set which makes the average far less relevant, but it's not always possible to avoid these outliers. This is where the median can be particularly helpful in understanding the data set.

The range

The range is defined as the difference between the highest and lowest values in a dataset. The range gives us an indication of the spread of the data-set i.e the measure of how far apart the values in the data set exist from each other.

Let's look at this data set showing salaries:

Role	Salary$
Executive	$200,000
Manager	$130,000
Supervisor	$90,000
Assistant	$60,000
Custodian	$50,000

The range is calculated by subtracting the lowest value (50,000) from the highest value (200,000) to gain the value of $150 000.

The range, in this case, is telling us that there is a large difference in what the lowest and highest person is paid.

Key Terms:

- **Mean**: The Average. To calculate, **Sum** all values in the dataset, then **divide** by *total number of values*.
- **Mode**: The most commonly occurring value in the dataset.
- **Median**: The **middle** value in an *ordered dataset*.
- **Range**: The difference between the **highest** and **lowest** values in a dataset.

In the next lesson, you'll discover how to use excel functions to calculate these.