LESSON HANDOUT

Frequency Distributions

Frequency distributions provide a visualization of the distribution or spread of your data which will help you understand and determine which statistical methods to apply during your analysis.

What is a frequency distribution?

A frequency distribution shows the frequency at which values in a dataset occur. When applied to numerical data, a frequency distribution is called a **histogram**.

Setting class sizes for histograms

When creating a histogram (working with continuous numerical data) we have one additional decision to make - which groups to divide out data into before we count the frequency. We call these groups **classes**.

When setting classes, we need to consider:

- Setting classes too broad might mean we lose information and patterns about data points in our distribution.
- Setting classes too small, may result in 'too much noise' and make patterns hard to discern.

We also have to ensure that:

- The classes do not overlap, and;
- The range within and between the classes are even (where possible).

Thankfully, there is an empirical method we can use to determine class size.

To determine the **number** of classes, we square root the *total count of values*. Say there are 40 values, 40s square root is 6.32 - so we want approximately 6 or 7 classes. To determine the **range** of each class we divide the range of our dataset by the square root. For example, 58.24 / 6.3 gives us 9.4 - so we want each class' range to be about 9.

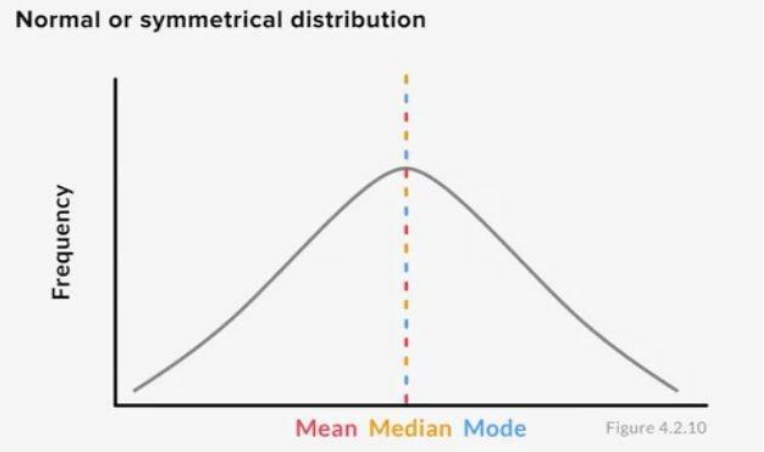
Interpreting Frequency Distributions

Distributions might take an irregular shape (from which we can only deduce that our values are randomly distributed), or they might take the shape of one of the four common patterns detailed below.

Normal or symmetrical distribution

In a normal distribution, values are evenly distributed on either side of the highest frequency.
The mean, median and mode will all be centrally located.

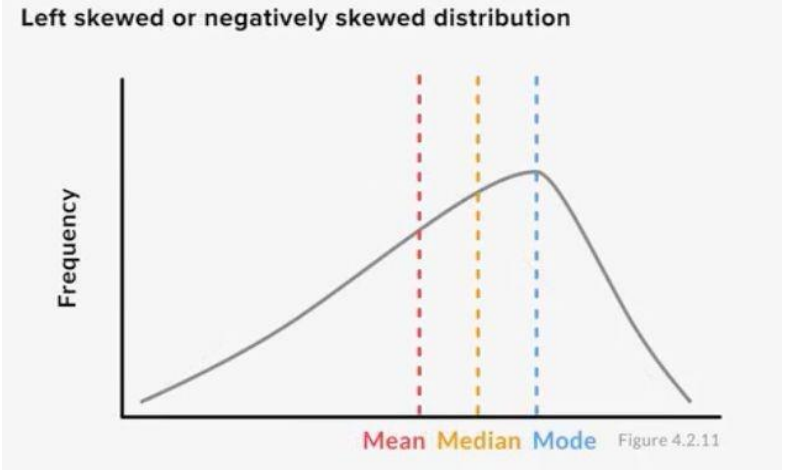
- **So what.** The main thing that this frequency distribution tells us is that our values are evenly distributed and that both the mean and median are accurate measures of central tendency. It also means the standard deviation is a useful way of segmenting our data.



Left skewed or negatively skewed distribution

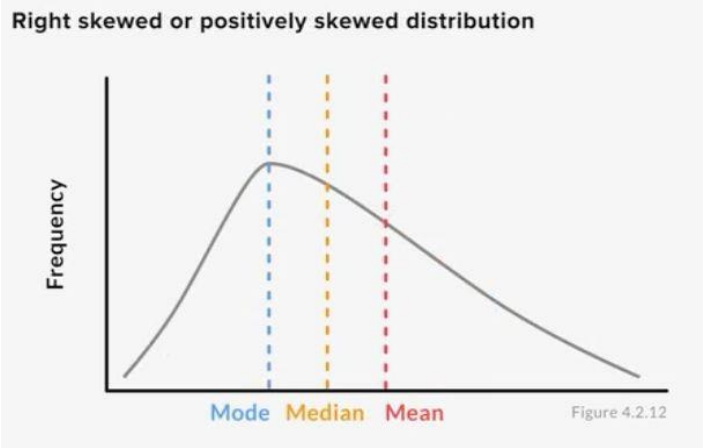
Left skewed because both the median and mean exist to the left of where most values occur the data set.

- **So what**. This distribution tells us that outliers exist to the left allowing us to factor this into any analysis we conduct. It also leads us toward using the median rather than the mean to measure the central tendency of our data and also to use quartiles rather than standard deviation to measure what constitutes relative high and low values in our data set.



Right skewed or positively skewed distribution

Shows the opposite shape to the left skewed distribution with the median and mean existing to the right of the highest frequency.



Bimodal distribution

In this distribution, we have two peaks with values distributed either side of these peaks. In this case the median and mean sit between the peaks with the modes existing at the peaks of each of the two mounds.

- **So what.** This distribution is telling us that neither the mean or median are an accurate measure of central tendency and that it will be difficult for us to use descriptive statistics to analyse our data without first segmenting it into separate two or more groups.