MEASURES OF CENTRAL LOCATION AND DISPERSION

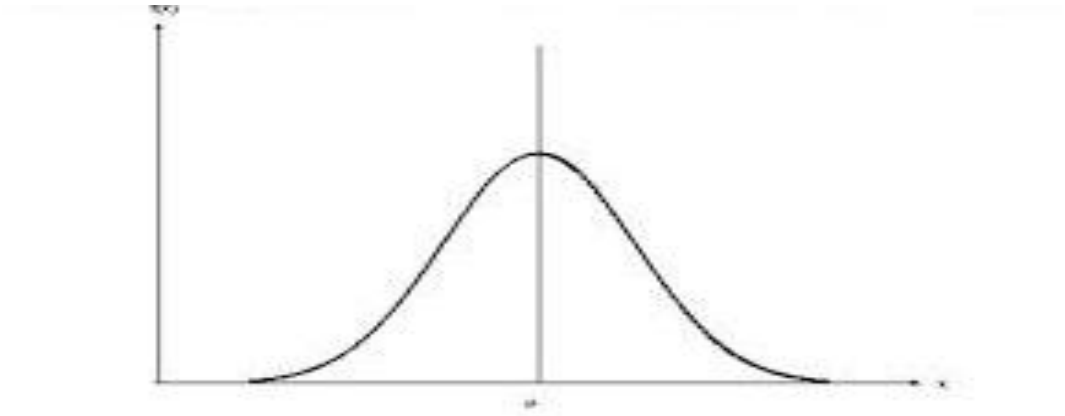
Improvement Cymru Academy

When collecting data, it can be useful to use summary statistics to describe the characteristics of that data set. Two key features that are often described are the measure of central location (also known as central tendency) and a measure of dispersion. An overview of these measures are provided below:

What is Central Location?

A measure of Central Location is a single value that represents the middle or centre of a dataset. The three main measures of central location are the mean, the median and the mode.

When data is normally distributed, the mean, median and mode should be identical; this can be seen in the Bell curve graph below.



It is particularly useful to describe a dataset using a measure of central location in conjunction with the measure of dispersion of a data set, which indicates the spread of the data.

How to use?

Mean

A very familiar measure of central location is the mean of a data set often known as the average value. It is calculated by dividing the sum of all values in a data set by the number of values.

So in a data set of 1, 3, 5, 6, 20, we would calculate the mean by adding the values (1+3+5+6 +20) and dividing by the total number of values (5). Our mean then is 35/5, which equals 7.

The mean is very useful as it is easy to calculate, easy to describe and uses every value in the data set.

The downside is that it is sensitive to outliers (observations which are markedly distant from the bulk of observations in a data set). Therefore, it is not appropriate to use when the data set is small or has a skewed distribution, rather than being of a normal distribution.

When to use it?

Statistical Process Control charts.

Median

The median of a data set is the middle value when arranged from smallest to largest. It means that half of the data set is above the median and half is below.

In the data set 1, 3, 5, 6, 20, the median is 5.

In a data set with *an even number of observations*, the median is calculated by dividing the sum of the two middle values by two. So in 1, 2, 3, 4, 5, 6, the median is (3+4)/2, which equals 3.5.

The median is useful where the data set has a skewed distribution or where there are outliers that would distort the mean.

When to use it?

Run charts.

Mode

The mode is the value in the data set that occurs most frequently.

It is useful to identify the popular or most frequently experienced option.

One drawback of the mode is that it is possible for two modes to appear in the one data set (e.g. in 1, 2, 2, 3, 4, 5, 5, both 2 and 5 are the modes).

When to use it?

The mode is an appropriate measure to use with categorical data and is visible on histograms.

Measures of Dispersion

Measures of dispersion describe the spread of data. Examples include the range, inter-quartile range and the standard deviation.

Range

The range describes the overall spread of data and is simply the difference between the maximum and minimum values in a data set.

Range = max - min

So in a data set of 2, 2, 3, 4, 5, 5, 6, 7, 8, 9, 11, 13, 15, 15, 17, 19, 20, the range is the difference between 2 and 20.

18 = 20 - 2

While it is useful in seeing how large the difference of values, the range can be distorted by extreme outliers and does not describe the spread of data values within that range e.g. if the data set is skewed.

References

Anon (2019) Designing and Conducting Health Systems Research Projects: Module 22 [online] Available at: https://www.betterevaluation.org/resources/guides/determine_reporting_requirements/design_conduct_heatlh_sys_res_projects [Accessed 12 Mar 2019]

Laerd Statistics (2013). 'Standard deviation' [online]. Available at: https://statistics.laerd.com/statistical-guides/measures-of-spread-standard-deviation.ph [Accessed 12 Mar 2019]