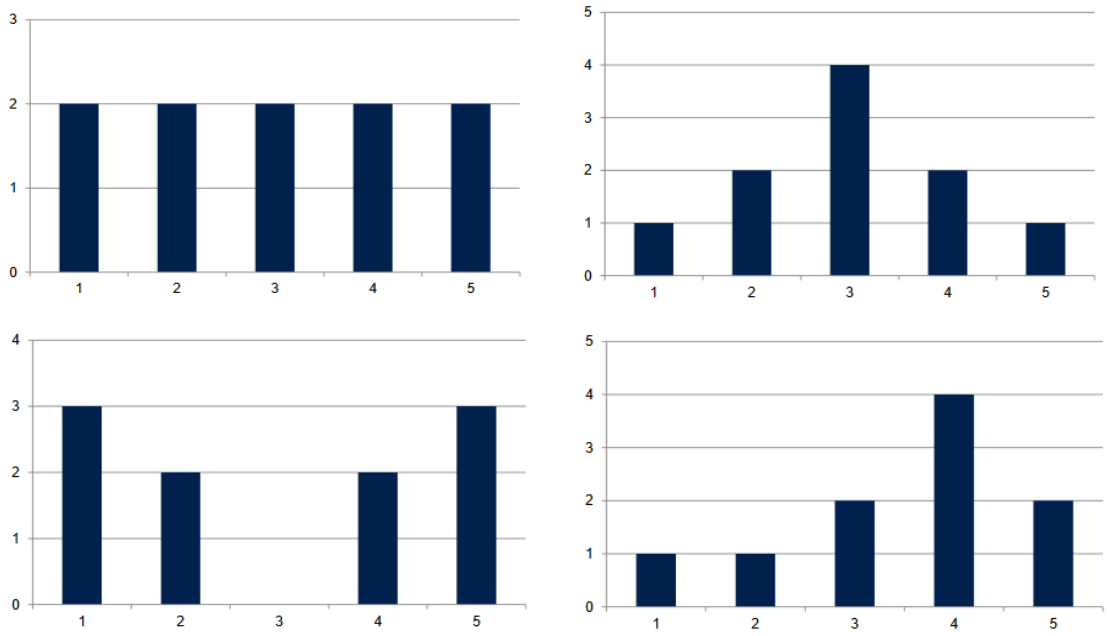
Data Analysis 1 *week 10*

System hypothesis experiment: collect data, visualise data, describe data, analyse data.

Visualising data includes looking at the distributions of answers. Distributions include uniform (top left), normal/Gaussian (top right), bimodal (bottom left), and skewed (bottom right).



Data distributions can give insights into which mathematical operations and statistical tests you can apply. Many tests require normal distribution of data, even things like the arithmetic mean, e.g. the normal distribution tells us that some people rated high, some rated low, but most rated average; the bimodal distribution shows us that half of the people rated high, while half rated low; however, the mean answer for both distributions is 3.

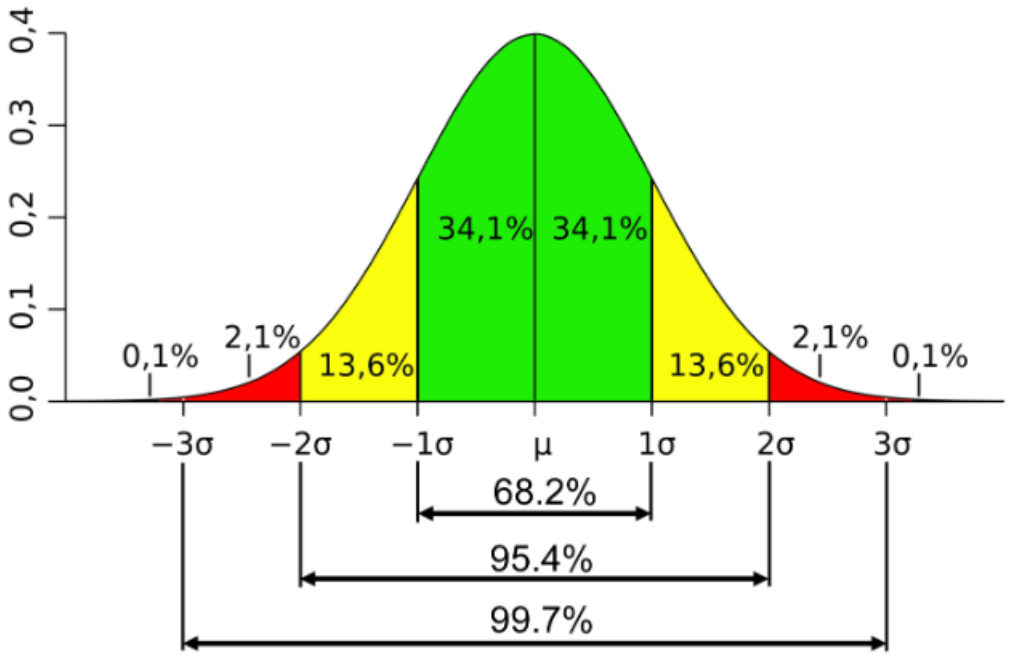
An average is a value that describes an entire distribution. The mean is the sum of all values divided by the amount of values. The mode is the most frequent value. The median is a value that splits the dataset at 50%. Depending on distribution, the chosen average leads to appropriate results. Means are used with normal distributions, medians with skewed, and modes for non-ordinal data, e.g. months.

Spread includes the range and deviation of a dataset. The range is the highest value and the lowest value of a dataset, e.g. 10 to 30 years old. The deviation tells us if the average model is a good representation of the data. Variance takes the sum of the squared (to account for direction differences) differences between the data and the average and divides it by N (whole population) or N-1 (population sample).

$$\frac{\sum(x_i-\bar{x})^2}{N-1}$$

Standard deviation takes the square root of variance to give a more realistic, smaller answer.

$$\sqrt{\frac{\sum(x_i-\bar{x})^2}{N-1}}$$



The above graph shows the percentage of answers within different amounts of the standard deviation (σ) from the mean (μ) in a normal distribution, i.e. 95.4% of answers are within two standard deviations either side of the mean.

Remember that correlation does not always mean causation. The two datasets could be affected by the same thing, e.g. ice cream sales and sunglasses sales increase during Summer, not because one causes the other, but because they are both influenced by the effects of more daily sun (hotter weather and brighter days).

Using mixed methods is good. Quantitative analysis helps explore, find patterns in, and generate high-level descriptions of data. Qualitative analysis helps interpret results and explains why the data is the way it is.