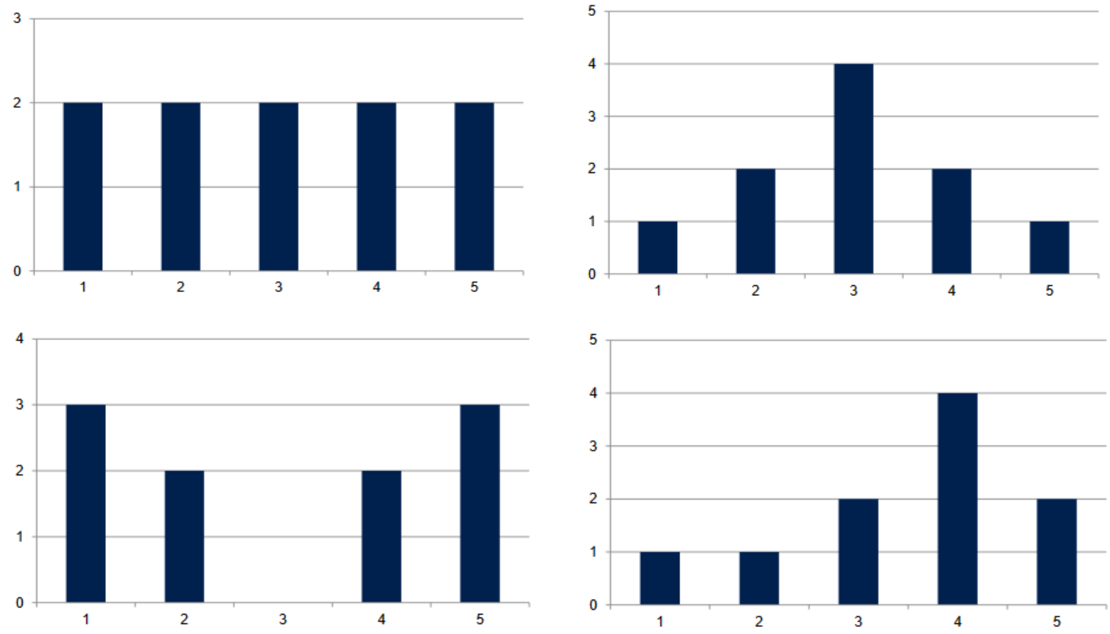
Data Analysis 1 *week 10*

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.

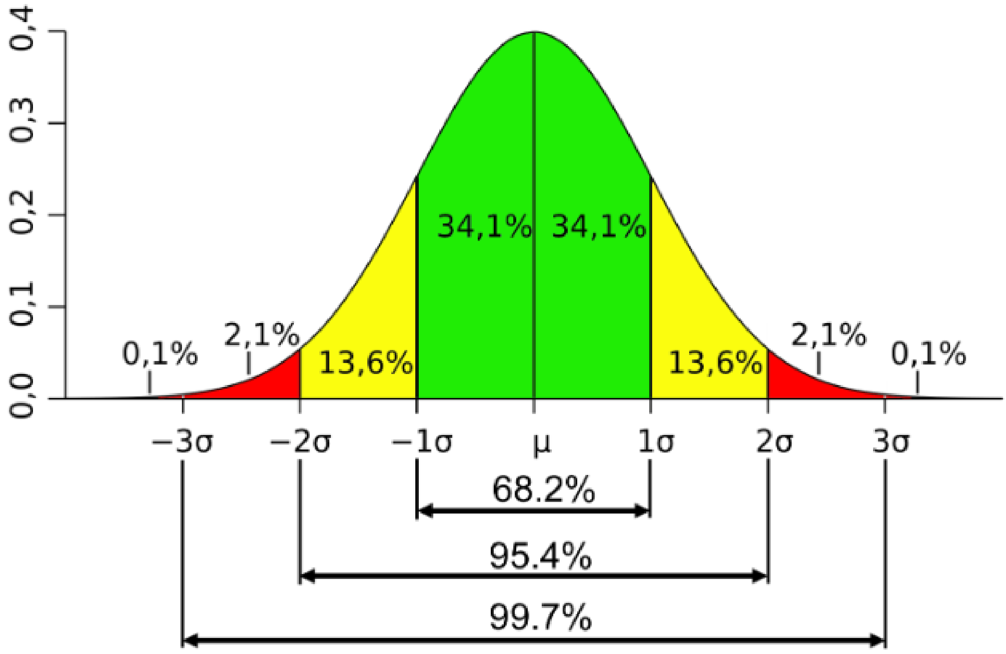
An average is a value that describes an entire distribution. The mean is the sum of all values divided by the amount of values. The mode is the most frequent value. The median is a value that splits the dataset at 50%. Depending on distribution, the chosen average leads to appropriate results. Means 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 differences (to account for direction 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.

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



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.

Data Analysis 2 *week 11*

Simple random sampling is selecting people at random from a known population. Convenience sampling is sampling people because they are convenient, e.g. physically close or part of an easily accessed group.

When comparing two sets of data, we want to know how large the effect is. What is the size of the difference and what impact is this likely to have?

Simple difference: $\boldsymbol{diff = |\bar{A} - \bar{B}|}$

Cohen's d: $\boldsymbol{d = \frac{\bar{A} - \bar{B}}{s_{pooled}}}$ $\boldsymbol{s_{pooled} = \sqrt{\frac{s_A^2 + s_B^2}{2}}}$

Confidence intervals describe the level of uncertainty in a sample parameter (estimate of the population parameter), e.g. the mean of a sample. For example, if the mean of a sample is 50, you could say that you are confident that the population's mean is between 40 and 60.

A null hypothesis is the hypothesis that there is no significant difference between conditions or populations. They are used when a hypothesis cannot be proved. For example, if you were trying to prove that A was better than B, your hypothesis (H_1) would be "we hypothesise that A is better than B". Your null hypothesis (H_0) would be "there is no real difference between A and B". If your study found that A was better than B, you could say "we reject the null hypothesis H_0 and accept H_1".

The p-value tells us if there is a significant difference between two conditions; it is the chance that null hypothesis is true. If $p < 0.05$, that means that the hypothesis H_1 is likely true and that the null hypothesis is likely false.

A study's design is between participants if one group completes one task while another group completes another task, and within participants if all participants complete both tasks.

Independent variables are controlled inputs. In the previous "A is better than B" study, we only had one independent variable (A or B), while if we also wanted to see how, for example, gender impacts the results of the study, we would have two independent variables ((A or B) and participant gender). As we are comparing two things, A and B, we say we have two levels of independent variable.

Parametric significance tests assume the data is normally distributed. Continuous data is typically parametric. Non-parametric significance tests do not rely on any distribution.

The t-test assesses whether the means of two datasets are significantly different. It uses the two datasets' means, standard deviations, and numbers of participants. This outputs the p-value.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Typically, you should aim to use a parametric test, unless the assumptions are specifically violated. First look at parametric tests, then non-parametric, as parametric tests have more statistical power.

The p-value says nothing about the magnitude of the effect. A smaller p-value does not imply a stronger effect than a larger p-value. P-values are not reliable indicators of replicability. Confidence intervals are a way of providing better information about replication.

Regarding correlations, an r value indicates the amount of variance in a set of results (a higher r value means a more significant correlation).