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 between he two datasets? What impact is this likely to have?

Simple difference is the absolute value of A's mean minus B's mean.

$$diff = |\overline{A} - \overline{B}|$$

Cohen's d:

$$d = \frac{\overline{A} - \overline{B}}{s_{pooled}} \qquad s_{pooled} = \sqrt{\frac{s_A^2 + s_B^2}{2}}$$

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) "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 one task while another group completed 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. The t-test 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).