6.6: Quantifying Effects

In the previous sections, we discussed how we can use data to test hypotheses. Those methods provided a binary answer: we either reject or fail to reject the null hypothesis. However, this kind of decision overlooks a couple of important questions. First, we would like to know how much uncertainty we have about the answer (regardless of which way it goes). In addition, sometimes we don't have a clear null hypothesis, so we would like to see what range of estimates are consistent with the data. Second, we would like to know how large the effect actually is, since as we saw in the weight loss example in the previous section, a statistically significant effect is not necessarily a practically important effect.

In this section, we will discuss methods to address these two questions: confidence intervals to provide a measure of our uncertainty about our estimates, and effect sizes to provide a standardised way to understand how large the effects are. We will also discuss the concept of *statistical power* which tells us how likely we are to find any true effects that actually exist.

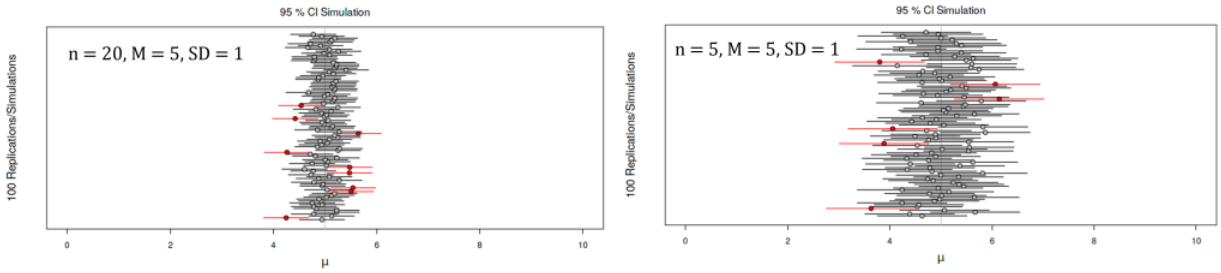

Figure 6.6.1. 95 % Confidence Intervals with different sample sizes (n) but with the same population parameters

Figure 6.6.2. Confidence intervals for the mean in jamovi

Relation of Confidence Intervals to Hypothesis Tests

There is a close relationship between confidence intervals and hypothesis tests. In particular, if the confidence interval does not include the null hypothesis, then the associated statistical test would be statistically significant. For example, if you are testing whether the mean of a sample is greater than zero with $_{\alpha}\alpha=0.05"\gt=0.05$, you could simply check to see whether zero is contained within the 95% confidence interval for the mean.

Things get trickier if we want to compare the means of two conditions or more (Schenker & Gentleman 2001).[3] In certain situations, statistical analysis is conducted by comparing the confidence intervals of the estimates to determine if there is any overlap. When the confidence intervals do not overlap, this is interpreted as indicating a statistically significant difference (as shown in Figure 6.6.3). It is generally accepted that non-overlapping confidence intervals signify statistical significance, but it's important to note that the reverse is not always true for overlapping confidence intervals (as depicted in Figure 6.6.3). For instance, what about the case where the confidence intervals overlap one another but don't contain the means for the other group? In this case, the answer depends on the relative variability of the two variables, and there is no general answer. To obtain a more precise assessment, an alternative method involves calculating the ratio or difference between the two estimates and constructing a test or confidence interval based on that particular statistic.

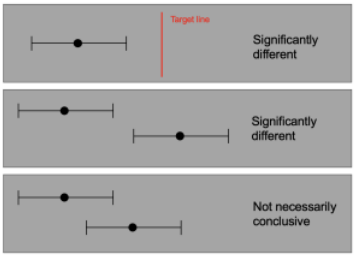

Figure 6.6.3. Using confidence intervals for making comparisons. The two top images show non-overlapping confidence intervals which can be statistically significant. The bottom image shows that overlapping confidence intervals do not always indicate a difference that is not statistically significant

While some academics suggest avoiding the "eyeball test" for overlapping confidence intervals (e.g., Poldrack, 2023), academics like Geoff Cummings are a strong advocate for using confidence intervals instead of NHST. [4]

$$d = \frac{M_1 - M_2}{S_{pooled}}$$

where M_1 and M_2 are the means of the two groups, and S_{pooled} is the pooled standard deviation (which is a combination of the standard deviations for the two samples, weighted by their sample sizes). Note that this is very similar in spirit to the t statistic – the main difference is that the denominator in the t statistic is based on the standard error of the mean, whereas the denominator in Cohen's d is based on the standard deviation of the data. This means that while the t statistic will grow as the sample size gets larger, the value of Cohen's d will remain the same.

Figure 6.6.4 shows that the two distributions are quite well separated, though still overlapping, highlighting the fact that even when there is a very large effect size for the difference between two groups, there will be individuals from each group that are more like the other group.

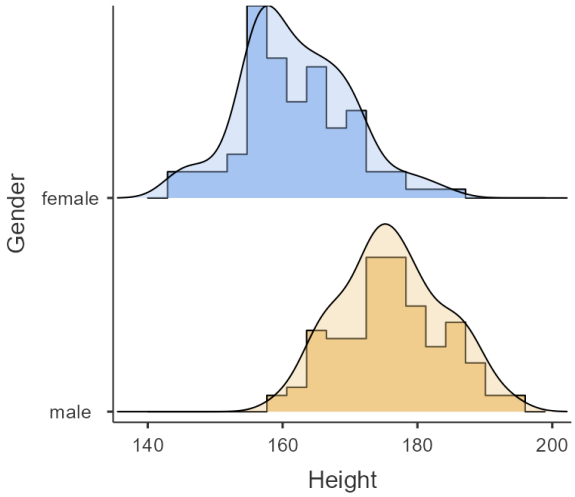

Figure 6.6.4 Histogram with density plots for male and female heights in the NHANES dataset, showing distinct but also clearly overlapping distributions. Screenshot from the jamovi program

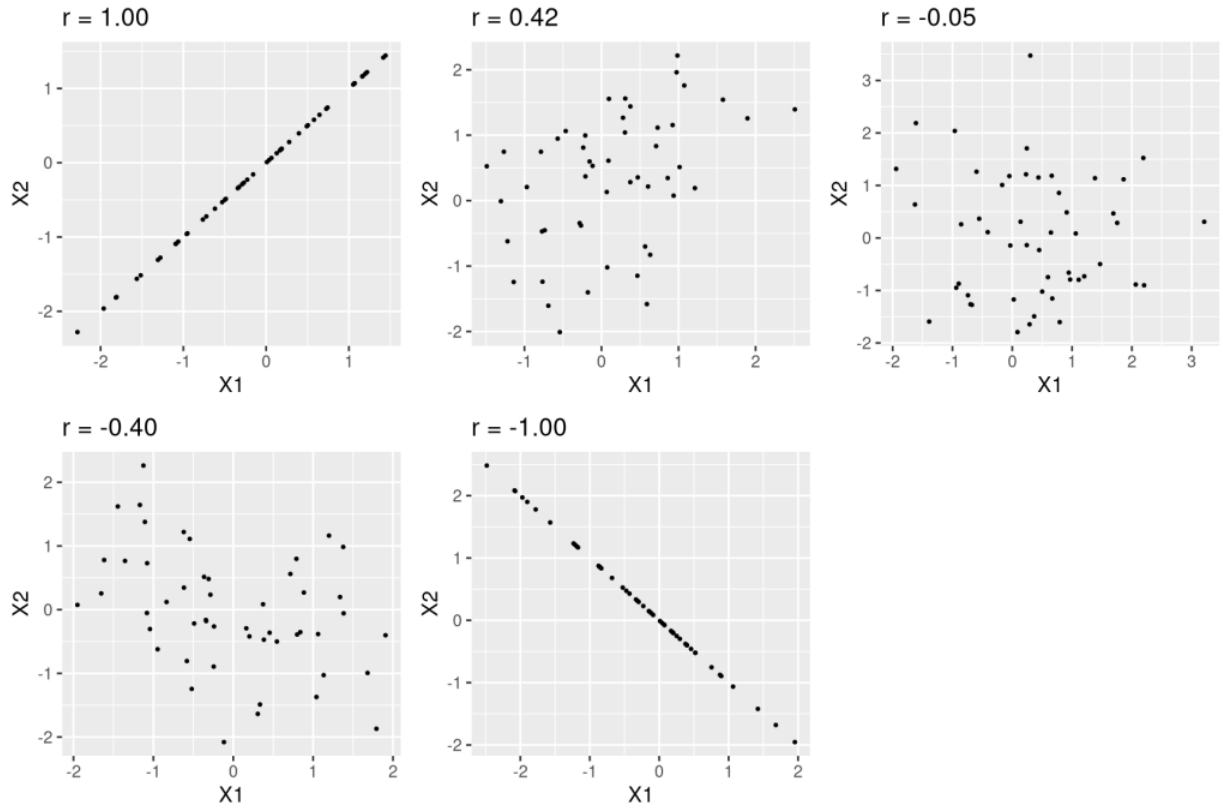

Figure 6.6.5. Examples of various levels of Pearson's r. Image by Poldrack, licenced under CC BY-NC 4.0

Figure 6.6.6 shows an example of how power changes as a function of these factors.

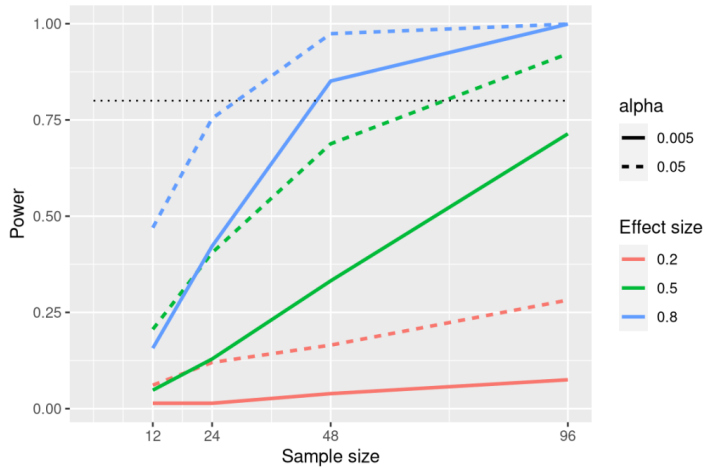

Figure 6.6.6. Results from power simulation, showing power as a function of sample size, with effect sizes shown as different colours, and alpha shown as line type. The standard criterion of 80 per cent power is shown by the dotted black line. Image by Poldrack, licensed under CC BY-NC 4.0

This simulation shows us that even with a sample size of 96, we will have relatively little power to find a small effect (d=0.2) with α=0.005. This means that a study designed to do this would be *futile* – that is, it is almost guaranteed to find nothing even if a true effect of that size exists.

There are at least two important reasons to care about statistical power. First, if you are a researcher, you probably don't want to spend your time doing futile experiments. Running an underpowered study is essentially futile because it means that there is a very low likelihood that one will find an effect, even if it exists. Second, it turns out that any positive findings that come from an underpowered study are more likely to be false compared to a well-powered study.

Power Analysis

Fortunately, there are tools available that allow us to determine the statistical power of an experiment. The most common use of these tools is in planning an experiment (i.e., *a priori* power analysis), when we would like to determine how large our sample needs to be to have sufficient power to find our effect of interest. We can also use power analysis to test for sensitivity. In order words, a priori power analysis answers the question, "How many participants do I need to detect a given effect size?" and sensitivity power analysis answers the question, "What effect sizes can I detect with a given sample size?"

In jamovi, a module called jpower allows users to conduct power analysis when conducting an independent samples t test, paired samples t test and one sample t test. This module is a good start – however, if you need another software that can accommodate other statistical tests, *G*Power* is one of the most commonly used tools for power analysis. You can find the latest version using this link.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

1. Neyman, J. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 236*(767), 333–80. doi.org/10.1098/rsta.1937.0005 ↵
2. shiny.rit.albany.edu/stat/confidence/ ↵
3. Schenker, N., & Gentleman J. F. 2001. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182–86. www.jstor.org/stable/2685796 ↵
4. Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29. doi.org/10.1177/0956797613504966 ↵
5. Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the p value Is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. doi.org/10.4300/JGME-D-12-00156.1 ↵
6. Cohen, J. (1994). The earth is round (p< 0.05). *American Psychologist*, *49*(12), 997. ↵
7. Wakefield, A. J. (1999). MMR vaccination and autism. *The Lancet*, *354*(9182), 949–950. https://doi.org/10.1016/S0140-6736(05)75696-8 ↵