LING82100: effect size and power analysis

Kyle Gorman

1 Effect size

An *effect size* can be defined in several closely-related ways. First, in simple one- and (unpaired) two-samples tests, it can be thought of as a measure of the magnitude—and sometimes, though not always, of the sign—of the differences between the sample statistic and the parameter (i.e., the population mean or median) of the null model. In a paired test, the sample parameter is the usually the mean or median of the differences between the paired samples, and the population parameter defining the null model is a mean or median of zero. When we consider more complex models such as ANOVAs or linear models, it can be thought of as a measure of the impact of the independent variable (or predictor) on the dependent variable (or outcome).

1.1 Absolute effect size

The simplest kind of effect size is the *absolute* effect size, which is usually expressed as a raw difference. For instance, imagine a study in which some students receive an intervention and others do not; this naturally corresponds to a two-sample, unpaired test. We could then compare the average (mean or median) grades between the two groups; e.g., "the intervention resulted in an average improvement of one letter grade".

1.2 Cohen's d

So, is that difference a lot or a little? *Standardized* (or *relativized*, or *relative*) effect size measures attempt to answer that by scaling the effect size by the standard deviation. One simple form of this used for t-tests is known as *Cohen's d.* It is defined as

$$d = \frac{|\mu_1 - \mu_2|}{\sigma} \tag{1}$$

$$\hat{d} = \frac{|\bar{X} - \mu|}{s} \tag{2}$$

$$\hat{d} = \frac{|\bar{X_1} - \bar{X_2}|}{s_{pooled}} \tag{3}$$

where the "pooled" standard deviation is given by

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}. \tag{4}$$

This statistic can be interpreted in terms of how much separation there are between the two Gaussians; see the lecture for some graphical examples. While this statistic is somewhat difficult to interpret, one commonly-used qualitative guideline (Cohen 1988:532) says that a $d < .2$ is “small”, $.2 < d < .8$ is “medium”, and $d > .8$ is “large”.

Effect sizes are strictly independent of sample size, and their relationship to p-values and statistical significance is complex. Thus both small and large effect sizes can give significant or non-significant results, depending on other factors to be defined below.

- As effect size increases, the expected magnitude of t increases, and
- as n increases, the expected magnitude of t increases as a function of $\sqrt{n/2}$.

1.3 Wilcoxon *r*

For median-based statistics, a statistic known as r is sometimes used. It is defined as

$$r = \frac{|Z|}{\sqrt{n}} \tag{5}$$

$$\hat{r} = \frac{\left|\hat{Z}\right|}{\sqrt{n_1 + n_2}}. \tag{6}$$

(and so on) where Z and $\hat{Z}$ are the Z-score for the p-value of the appropriate Wilcoxon test. There is no standard function for computing this statistic in R, but the following should suffice.

```
# NB: This long list of arguments just mirrors those of `wilcox.test`.
wilcox.r <- function(x, y = NULL,
                     alternative = c("two.sided", "less", "greater"),
                     mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
                     conf.int = FALSE, conf.level = 0.95) {
    test <- wilcox.test(x, y, alternative, mu, paired, exact, correct,
                        conf.int, conf.level)
    Z <- abs(qnorm(test$p.value / 2))
    n <- length(x)
    if (!is.null(y) & !paired) n <- n + length(y)
    Z / sqrt(n)
}
```

2 Power

The degree to which we are able to detect a significant difference is known as *power*. It can be thought of as the probability that the null hypothesis will be rejected given a particular effect size, and so is a function of α, d, and n. To increase power, one can

- increase α—at the cost of Type I error,

- increase *d*,
- increase precision of measurement,
- decrease variability (e.g., by using a within-subjects design), or
- increase *n*.

In fact, if one knows any three of $\{d, n, \alpha, \text{power}\}$, then you can solve for the fourth. A standard qualitative guideline, once again due to Cohen (1992:100), suggests fixing power at least .8. The following examples use the `pwr` library, available from CRAN.

```
> library(pwr)
> pwr.t.test(n = 25, d = .5, sig.level = .05, type = "paired")

     Paired t test power calculation

              n = 25
              d = 0.5
      sig.level = 0.05
          power = 0.6697077
    alternative = two.sided

NOTE: n is number of *pairs*

> pwr.t.test(d = .5, sig.level = .05, power = .8, type = "paired")

     Paired t test power calculation

              n = 33.36713
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

We see that at $d = .5$ and $\alpha = .05$, a 25-sample paired *t*-test is somewhat underpowered according to Cohen's rule of thumb, but would achieve adequate power at roughly $n = 33$.

References

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Associates, 2nd edition.

Cohen, Jacob. 1992. Quantitative methods in psychology: a power primer. *Psychological Bulletin* 112:155–159.