location: FAQ / tdunpaired

Taken from Jake Westfall's comments on choice of Cohen's d here.

Five different “Cohen’s d” statistics for within-subject designs

Jeff Rouder poses an “effect size puzzler” where the puzzle is simply to compute a standardized effect size for a simulated dataset where subjects make 50 responses in each of 2 conditions. He offers a sweatshirt prize (???) to anyone who can do so “correctly.”

(Update: Jeff posted his answer to the effect size puzzler here.)

As it happens, I’ve been meaning for a while to write a blog post on issues with computing d-like effect sizes (or other standardized effect sizes) for within-subject designs, so now seems like a good time to finally hammer out the post.

Jeff didn’t actually say anything to restrict us to standardized mean difference type measures (as opposed to, say, variance explained type measures), and we can only guess whether the “correct” effect size he has in mind is a d-like measure or an R^2-like measure or what. But here I’ll focus on d-like measures, which are perhaps the most popular for studies with categorical predictors, and offer plenty of complications that I think are often under-appreciated (and it seems Jeff would agree).

What I’ll show here is that there are at least 5 different and non-equivalent ways that people might compute a d-like effect size (which they would invariably simply call “Cohen’s d”) for Jeff’s dataset, and the resulting effect sizes range from about 0.25 to 1.91. I’ll compare and contrast these procedures and ultimately choose one that I think is the least crazy, if you must compute a standardized effect size (more on that later). When I read a paper that reports “Cohen’s d” for a within-subject design, I usually have no idea which of these 5 procedures the authors actually applied unless I try to reproduce the effect size myself from some of the descriptives in the paper, which is often not possible.

Jeff’s dataset can be downloaded here. Here’s some code to load it into R, examine it, print the two conditions means, and save the mean difference as md:

```
1 dat <- read.table("http://pcl.missouri.edu/exp/effectSizePuzzler.txt",
header=TRUE)
2

3 str(dat)
4 # 'data.frame':       2500 obs. of  3 variables:
5 #  $ id  : int  1 1 1 1 1 1 1 1 1 1 ...
6 #  $ cond: int  1 1 1 1 1 1 1 1 1 1 ...
7 #  $ rt  : num  0.56 0.93 0.795 0.615 1.028 ...
8

9 (means <- with(dat, tapply(rt, cond, mean)))
10 #         1         2
11 # 0.8322560 0.8845544
12

13 md <- diff(means)
```

view raw esPuzzlerData.R hosted with ❤ by GitHub

d: Classical Cohen’s d

This is essentially a straightforward application of the formula provided by Cohen himself:

d=(M_1-M_2) /sigma

The numerator is the mean difference. Cohen never actually gives a clear, unambiguous definition of sigma in the denominator, but it is usually taken to be the pooled standard deviation, that is, the square root of the average variance in each condition (assuming equal sample sizes in both conditions). In R code:

```
(d <- md / with(dat, sqrt(mean(tapply(rt, cond, var)))))
# 0.2497971
```

The crucial thing to recognize about applying the classical Cohen's d is that it deliberately ignores information about the design of the study. That is, you compute d the same way whether you are dealing with a between-subjects, within-subjects, or mixed design. Basically, in computing d, you always treat the data as if it came from a simple two-independent-groups design. I don't want to get bogged down by a discussion of why that is a good thing at this point in the post—I'll get into that later on. For now I just note that, with this effect size, within-subject designs tend to be powerful not because they lead to larger effect sizes—if anything, the reverse is probably true, in that people elect to use within-subject designs when Cohen's d is particularly small, for example in many reaction time studies—but rather because they allow us to efficiently detect smaller effect sizes due to removing irrelevant sources of variation from the denominator of the test statistic.

da: Cohen's d after averaging over replicates

For this next way of computing a d-like effect size, the section heading pretty much says it all. In Jeff's dataset, each participant makes a total of 100 responses, 50 in each condition. The computation of d_a proceeds by first averaging over the 50 responses in each subject-by-condition cell of the experiment, so that each subject's data is reduced to 2 means, and then applying the classical d formula to this aggregated data. In R code: sub_means <- with(dat, tapply(rt, list(id, cond), mean))

```
(d_a <- md / sqrt(mean(diag(var(sub_means)))))
# 0.8357347
```

Richard Morey blogged about some weird things that can happen when you compute a standardized effect size this way. The basic issue is that, unlike classical Cohen's d, d_a does not ignore all the design information: d_a will tend to be larger when there are more replicates, that is, when each subject responds more frequently in each condition.

dz: Standardized difference scores

A third way to compute a d-like effect size is to reduce each subject's data to a single difference score —the mean difference between their responses in each condition—and then use the standard deviation of these difference scores as the denominator of d. Cohen actually discusses this statistic in his power analysis textbook (Cohen, 1988, p. 48), where he carefully distinguishes it from the classical Cohen's d by calling it d_z. In R, we can compute this as:

```
(d_z <- md / sd(sub_means[,2] - sub_means[,1]))
# 1.353713
```

There is a straightforward relationship between d_z and the test statistic: t_w = d_z\sqrt{n}, where t_w is the paired-samples t-statistic from a within-subjects design and n is the number of subjects. One might regard this as a virtue of d_z. I will argue below that I don't think it's a good idea to use d_z.

dt: Naive conversion from t-statistic

In a simple two-independent groups design, one can compute the classical Cohen's d from the t-statistic using

```
d=t_b Sqrt[2/n]
```

where t_b is the independent-samples t-statistic for a between-subjects design and n is the number of subjects per group. Many, many authors over the years have incorrectly assumed that this same conversion formula will yield sensible results for other designs as well, such as in the present within-subjects case. Dunlap et al. (1996) wrote a whole paper about this issue. One might suppose that

applying this conversion formula to t_w will yield d_z, but we can see that this is not the case by solving the equation given in the previous section for d_z, which yields d_z=\frac{t_w}{\sqrt{n}}. In other words, naive application of the between-subjects conversion formula yields an effect size that is off by a factor of \sqrt{2}.

To compute d_t in R:

```
t_stat <- t.test(sub_means[,2] - sub_means[,1])$statistic
(d_t <- t_stat*sqrt(2/nrow(sub_means)))
# 1.914439
```

dr: Residual standard deviation in the denominator

Finally, one can compute a d-like effect size for this within-subject design by assuming that the \sigma in the classical Cohen's d formula refers to the standard deviation of the residuals. This is the approach taken in Rouder et al. (2012) on Bayes factors for ANOVA designs. In between-subjects designs where each subject contributes a single response, this is equivalent to classical Cohen's d. But it differs from classical Cohen's d in designs where subjects contribute multiple responses.

To compute d_r in R, we need an estimate of the residual standard deviation. Two ways to obtain this are from a full ANOVA decomposition of the data or by fitting a linear mixed model to the data. Here I do it the mixed model way:

```
options(contrasts=c("contr.helmert","contr.poly"))
library("lme4")
mod <- lmer(rt ~ cond + (cond||id), data=dat)
summary(mod)
# Random effects:
#  Groups   Name        Variance  Std.Dev.
#  id       (Intercept) 0.0026469 0.05145
#  id.1     cond        0.0001887 0.01374
#  Residual             0.0407839 0.20195
# Number of obs: 2500, groups:  id, 25
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept) 0.779958   0.016402   47.55
# cond        0.052298   0.008532    6.13
```

The residual standard deviation is estimated as 0.20195, which gives us d_r = 0.259. It turns out that, for this dataset, this is quite close to the classical Cohen's d, which was 0.25. Basically, classical Cohen's d is equivalent to using the square root of the sum of all the variance components in the denominator1,2, rather than just the square root of the residual variance as d_r uses. For this simulated dataset, the two additional variance components (intercepts and slopes varying randomly across subjects) are quite small compared to the residual variance, so adding them to the denominator of the effect size does not change it much. But the important thing to note is that for other datasets, it is possible that d and d_r could differ dramatically.

So which one should I compute?

Contrary to Jeff, I don't really think there's a "correct" answer here. (Well, maybe we can say that it's hard to see any justification for computing d_t.) As I put it in the comments to an earlier blog post:

Basically all standardized effect sizes are just made-up quantities that we use because we think they have more sensible and desirable properties for certain purposes than the unstandardized effects. For a given unstandardized effect, there are any number of ways we could "standardize" that effect, and the only real basis we have for choosing among these different effect size definitions is in choosing the one that has the most sensible derivation and the most desirable properties relative to other candidates.

I believe that classical Cohen's d is the option that makes the most sense among these candidates. Indeed, in my dissertation I proposed a general definition of d that I claim is the most natural

generalization of Cohen's d to the class of general ANOVA designs, and I considered it very important that it reduce to the classical Cohen's d for datasets like Jeff's. My reasoning is this. One of the primary motivations for using standardized effect sizes at all is so that we can try to meaningfully compare effects from different studies, including studies that might use different designs. But all of the effect size candidates other than classical Cohen's d are affected by the experimental design; that is, the "same" effect will have a larger or smaller effect size based on whether we used a between- or within-subjects design, how many responses we required each subject to make, and so on. Precisely because of this, we cannot meaningfully compare these effect sizes across different experimental designs. Because classical Cohen's d deliberately ignores design information, it is at least in-principle possible to compare effect sizes across different designs. Morris and DeShon (2002) is a nice paper that talks about these issues. Bakeman (2005) also has some great discussion of essentially this same issue, focused instead on "variance explained"-type effect sizes.

Although I don't really want to say that there's a "correct" answer about which effect size to use, I will say that if you choose to compute d_z, d_r, or anything other than classical Cohen's d, just please do not call it Cohen's d. If you think these other effect sizes are useful, fine, but they are not the d statistic defined by Cohen! This kind of mislabeling is how we've ended up with 5 different ways of computing "Cohen's d" for within-subjects designs.

Finally, there is a serious discussion to be had about whether it is a good idea to routinely summarize results in terms of Cohen's d or other standardized effect sizes at all, even in less complicated cases such as simple between-subjects designs. Thom Baguley has a nice paper with some thoughtful criticism of standardized effect sizes, and Jan Vanhove has written a couple of nice blog posts about it. Even Tukey seemed dissatisfied with the enterprise. In my opinion, standardized effect sizes are generally a bad idea for data summary and meta-analytic purposes. It's hard to imagine a cumulative science built on standardized effect sizes, rather than on effects expressed in terms of psychologically meaningful units. With that said, I do think standardized effect sizes can be useful for doing power analysis or for defining reasonably informative priors when you don't have previous experimental data.

Footnotes

1 In general, this is really a weighted sum where the variance components due to random slopes must be multiplied by a term that depends on the contrast codes that were used. Because I used contrast codes of -1 and +1, it works out to simply be the sum of the variance components here, which is precisely why I changed the default contrasts before fitting the mixed model. But be aware that for other contrast codes, it won't simply be the sum of the variance components. For more info, see pp. 20-21 of my dissertation.

2 If you actually compute classical Cohen's d using the square root of the sum of the estimated variance components, you will find that there is a very slight numerical difference between this and the way we computed d in the earlier section (0.2498 vs. 0.2504). These two computations are technically slightly different, although they estimate the same quantity and should be asymptotically equivalent. In practice the numerical differences are negligible, and it is usually easier to compute d the previous way, that is, without having to fit a mixed model.

None: FAQ/tdunpaired (last edited 2017-06-05 14:41:57 by PeterWatson @ pc0021.mrc-cbu.cam.ac.uk[172.31.10.21]: PeterWatson)