Two cautions on the use of Cohen's *d*

E-mail distributed on 09-09-2024

Dear all,

When comparing two means with a *t*-test, we often report an effect size alongside the test statistics, as a way to judge if the difference in means is "small", "medium", or "large". Although one should prefer outcome measures where such qualitative judgments can be made directly on the raw outcome scale (e.g., number of units sold, number of patients recovered), many outcome measures in social sciences are (somewhat) abstract. In this case, a *standardized* effect size can facilitate qualitative interpretation. For a comparison of two means, the traditional effect size to report is **Cohen's *d***. Cohen's *d* simply takes the difference in means, and scales it by an appropriate standard deviation (SD), to obtain a value quite like a *z*-score.

$$d_z = \frac{M_1 - M_2}{SD}$$

For a one-sample *t*-test,[1] the appropriate SD would correspond merely to the sample SD. For a two-group, independent samples *t*-test, the appropriate SD is typically taken to be the *pooled* SD for both groups. This pooled SD is equivalent to the root mean square error (RMSE) of the regression model where the observed values of the two groups are the outcome, and a categorical factor identifying both groups is the predictor. By dividing our difference of means by this pooled SD, we obtain our Cohen's *d*. Qualitative interpretations of Cohen's *d* are popular, with cutoffs of 0.2, 0.5, and 0.8 for small, medium and large effects, respectively (Cohen, 1988).[2] Alternatively, the interpretation can be made directly in percentage of standard deviation, such that an effect of 0.2 means a difference in 20% standard deviation on the outcome's original scale. More formulas and clarifications can be found in the excellent paper by Lakens (2013), which will serve as a much better introduction to Cohen's *d* than this newsletter.

Instead, I would like to draw your attention today to two cautionary scenarios for Cohen's *d*, specifically **(a)** the paired *t*-test and **(b)** the independent *t*-test with unequal variances, for which reporting the *naïve* Cohen's *d* is not recommended!

[1] Where M_2 drops out (or can be said to equal 0).

[2] So my earlier newsletter on automated interpretation of different effect sizes

1. Paired t-test

A paired *t*-test is mathematically identical to a one-sample *t*-test, in that one calculates difference scores, *D,* between the paired samples, and then tests whether the mean of those difference scores differs significantly from 0. Naïvely, one could calculate Cohen's *d* likewise, by $d_z = D/\mathrm{SD}(D)$. However, this calculation will overestimate the effect size when the paired samples are moderately to strongly correlated ($r > 0.5$). Solutions have been proposed to this problem including **(1)** taking the SD from one sample only, **(2)** involving the correlation directly in the calculation of Cohen's *d*, and **(3)** taking the average of both SDs (Lakens, 2013).

The first follows the logic of Glass' Δ, by taking the SD of a "control" group as the reference SD. This is primarily suited to intervention designs where the pre-measurement can plausibly serve that function. For other within-subject designs, this choice may not be obvious. The second is known as the repeated measures Cohen's *d* (d_{rm}), but has undesirable properties and is not recommended over the third option, known as averaged Cohen's *d* (d_{av}). Although d_{av} does not involve the correlation of the paired samples, it will be smaller than ordinary d_z when $r > 0.5$. Importantly, it can be compared to the same d_z that would be obtained under a between-subject design (Lakens, 2013).[3] Note that, when $r < 0.5$ between the paired data, d_{av} will overestimate the effect size. However, unless your participants are responding randomly,[4] the correlation between repeated measures taken from the same participant is almost always substantial. Thus, reporting d_{av} is generally recommended.

In R, **caution is warranted!** Almost all packages that offer a function for Cohen's *d* report the naïve d_z for paired samples, including the functions `CohenD` (`DescTools`), `cohens_d` (`effectsize`), `cohen.d` (`psych`), and `cohens_d` (`rstatix`). Even when the function offers an argument for paired samples, this does not mean that d_{av} will be calculated! Calculating d_{av} can be achieved with the function `cohen.d` from package `effsize`, which will also print a confidence interval for the effect size. If you use online calculators to obtain Cohen's *d* for a paired *t*-test, trust only those that require you to enter both SDs and/or their correlation to be sure that the naïve d_z is not returned. Finally, it is relatively easy to calculate the appropriate SD by yourself, with the formula:

$$SD_{av} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}$$

2. Unequal variances

A pooled SD is appropriate for the two-group, independent samples *t*-test because, if the variance in both groups is equal, we gain more precision by using the entire sample to estimate the SD. This fails if the assumption of equal variance is violated, a scenario known more commonly as heteroscedasticity. I have previously discussed the problem of heteroscedasticity, and how to correct it in inferential testing. For a two-group, independent samples comparison of means, the Welch *t*-test is appropriate in this scenario, and will down-correct the degrees of freedom to a fractional value, depending on the severity of heteroscedasticity. It follows that one should also report a corrected value for Cohen's *d*, although this remains a rare practice. Delacre et al. (2021) recommend reporting

[3] This is ideal especially for the purpose of meta analysis.

[4] For example, on very difficult tasks where chance-level performance may be expected

Hedges' *g*, with a **non-pooled SD**, which they refer to as *g**. The basic Hedges' *g* corrects Cohen's *d* for its small-sample bias, by multiplying d_z with a factor that inversely depends on the total sample size *N*. Approximately, this factor is equal to:

$$H \approx 1 - \frac{3}{4N - 9}$$

Hedges' g is therefore generally recommended for small samples. Accounting for heteroscedasticity, in turn, can be achieved by replacing the pooled SD with the non-pooled SD, which is equal to:

$$SD_{av} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}$$

Remarkably, this is the same formula as the SD in d_{av}! Combining the Hedges small-sample correction with the non-pooled SD thus becomes straightforward:

$$g^* = \frac{M_1 - M_2}{SD_{av}} \times H$$

In R, there appears to be no active package to perform this calculation. However, the defunct package `deffectsize` (linked to the Lakens group) offered the function `datacohen_CI` to calculate all varieties of Cohen's *d* and Hedges *g* discussed in this newsletter. The source code was archived and can be extracted here, with a manual available here. Although the manual calculation of these effect sizes is easy, the R function also provides confidence intervals, yielding further information on the precision of the effect.

Lastly, I urge caution with back-transforming effect sizes from test statistics, since such conversions are typically blind to the underlying test or model. It would be wrong to transform a Welch *t*-value back to a naïve Cohen's *d*, or treat a paired *t*-value as if it was an independent samples *t*-value. Likewise, there are issues with transforming multilevel test statistics into effect sizes, as discussed in a previous communication.

References

Cohen, J.(1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

Delacre, M., Lakens, D., Ley, C., Liu, L., & Leys, C. (2021). Why Hedges' g*s based on the non-pooled standard deviation should be reported with Welch's t-test. *PsyArXiv*.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12.

--

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79