Understanding Cohen's d

last revised October 18, 2012

A measure of effect size that is often used in experiments where one is comparing the mean of one sample to another is that of Cohen's **d**, defined as

$$\mathbf{d} = \frac{\mu_1 - \mu_2}{\sigma}$$

where μ_1 is the mean of the first population, and μ_2 is the mean of the second population. Sigma, σ, is the population standard deviation that is assumed to be "common" across both populations. What does this mean, exactly? When an independent-samples t-test is performed, recall that one assumption of the t-test is that the variances in each population are the same. That is, to conduct the "ordinary" (as opposed to variance-corrected) t-test, we must assume the following null hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

If we are conducting the usual independent-samples t-test then, it is assumed that the population variances are equal. Given this, all we need to compute Cohen's **d** is the population standard deviation of either group.

Cohen's d with Pooled Standard Deviation

Given that the sample means $\overline{X}_1, \overline{X}_2$ corresponding to their respective population means μ_1, μ_2 are almost surely not equal to one another, it makes sense that instead of computing σ as the "common standard deviation," we instead compute what is known as the pooled standard deviation, s_p, which its square, the variance, is defined by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where s_1^2 is the variance of the first sample, and s_2^2 is the variance of the second sample. The sample size is defined by n_1 for the first sample, and n_2 for the second.

You can see that if $n_1 = n_2$, the pooled variance reduces to simply the *average* variance of both samples combined. That is, given that $n_1 = n_2$, we could write the pooled variance as simply

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
$$= \frac{s_1^2 + s_2^2}{2}$$

since if $n_1 = n_2$ are equal, it is understood that both variances s_1^2 and s_2^2 are *weighted* the same.

However, when $n_1 \neq n_2$, pooling variances gives more weight to the sample with the larger n.

Returning to Cohen's **d**, as suggested by Howell (2002), one can compute it using sample means and the pooled standard deviation as such:

$$\mathbf{d} = \frac{\overline{X}_1 - \overline{X}_2}{s_p}$$

The resulting statistic often goes by the name of ***Hedges' g*** who suggested the use of sample statistics and the pooling of the standard deviation as an estimate of the overall population standard deviation. So, often, the statistic is written as

$$\mathbf{g} = \frac{\overline{X}_1 - \overline{X}_2}{s_p}$$

Cohen's d vs. Hedge's g

There really isn't that much of a difference between these two statistics, and to say that Hedges "invented" a new statistic "**g**" would be a bit much. As well, it takes little to realize that one can usually never actually compute Cohen's **d**, since it involves population means in the numerator. How often do we have population means from which to compute a statistic? Hardly ever, which is why we're usually engaged in our experiment in the first place (i.e., we're trying to estimate what these population figures might be). As well, how often will the population standard deviation σ in Cohen's **d** be representative of both populations? Again, hardly ever, since even if the actual population variances are equal (which we know they probably aren't anyway), we usually only have the sample variances. Hence, using the sample variances as part of the pooled estimate, s_p, is usually the best approach anyway. So in reality, we're usually going to be computing Hedge's **g**, and not Cohen's **d**. However, reporting Cohen's **d** in a manuscript will be much more identifiable than reporting Hedge's **g**, and since Cohen defined the nature of the statistic (to be discussed in some detail below), it seems wiser to still credit it to him when reporting this kind of effect size measure, instead of crediting it to Hedges simply because of the use of sample statistics.

What Does Cohen's d Actually Tell Us?

Writing out a formula and plugging in numbers into it when conducting research unfortunately does not necessarily give us a feeling for what the formula actually means. This is the case with Cohen's **d**. We discuss the statistic in some detail, and point out why it is usually interpreted as the *standardized difference between means*.

Example

Imagine you have two independent samples of laboratory rats. To one sample, you give normal feeding and observe their weight over the next 30 days. To the other sample, you also feed normally, but also give them regular doses of a weight-loss drug. You are interested in knowing whether your weight-loss drug works or not. Suppose that after 30 days, on average, your rats in the non-treated group weigh 0.5 pounds. In the treated group, on average, the rats weigh 0.3 pounds. Is a difference of 0.5 – 0.3 = 0.2 pounds meaningful? How would you know? It's difficult to know how meaningful this difference is unless you have something you can compare it to. How big is a difference of 0.2 pounds for these groups? If the average difference in weight among rats in the population were very large, say, 0.8 pounds, then a mean difference of 0.2 pounds isn't that impressive. After all, if rats weigh very differently from one rat to the next, then really, finding a mean difference of 0.2 between groups isn't that exciting.

However, if the average weight difference between rats were equal to 0.1 pounds, then all of a sudden, a mean difference of 0.2 pounds seems rather impressive, because that size of difference is *atypical* in the population.

Consider the two computations below, one with a population standard deviation of 0.8 pounds vs. one with a population standard deviation of 0.1 pounds. In each case, we pretend as if we know the parameters (i.e., the population means), and compute Cohen's **d**:

$$\mathbf{d}_{Case1} = \frac{\mu_1 - \mu_2}{\sigma} = \frac{0.5 - 0.3}{0.8} = 0.25$$

$$\mathbf{d}_{Case2} = \frac{\mu_1 - \mu_2}{\sigma} = \frac{0.5 - 0.3}{0.1} = 2$$

The first thing to notice about the above two computations is that in both, the mean difference in the numerator is the *same* (0.5 – 0.3 = 0.2). The only difference is the population standard deviation, σ. It isn't surprising then that a difference in means of 0.2 is much more impressive if the population has *smaller variability* than if it has *larger variability*. Notice that in the first computation, where $\sigma = 0.8$, Cohen's **d** is equal to only 0.25. In the second computation, where we have much less population variability, $\sigma = 0.1$, Cohen's **d** is equal to 2. This is because in the population for which there is very little weight variability to begin with (0.1), it means a lot more (in a substantive sense) to observe a mean difference of 0.2 than in a population for which weight variability is very large (0.8).

Cohen's d as a Comparison of Two Deviations

A very powerful way to interpret and understand Cohen's **d** is by paying close attention to exactly what we are doing when we compute it. Look at the formula once more. What is the numerator? It is a difference, or *deviation* of means. Hence, we obtained a deviation of means of 0.2 in our example. How big is this deviation in relative terms? For that, why not compare it to another deviation, one that is more or less usual or *standard* in the population from which these data were drawn? That is exactly what the *standard deviation* tells us, which is in the denominator, σ . Hence, when we are computing Cohen's **d** (or Hedge's **g**), we are in actuality producing a *ratio of one deviation relative to another*, similar to how when we compute a z-score, we are comparing the deviation of $X_i - \mu$ to the *standard* deviation σ .

Cohen's **d** then, is a measure of the *standardized difference between means*. Literally, it is the difference between means divided by the standard deviation, and just like the z-score, when we divide a difference by the standard deviation, we are *standardizing* that difference.

Interpreting the Size of Effect

Cohen originally gave guidelines as to what constitutes a small, medium, and large effect. He gave the guidelines of 0.2 as small, 0.5 as medium, and 0.8 and higher as large. However, these guidelines are only potentially useful if you are doing research in an area that has no history of effect sizes, or that you are unable to evaluate for yourself whether the obtained effect is practically meaningful.

For example, if previous research in your area of investigation has found effect sizes of 0.8 and higher for the phenomenon you are studying, then reporting an effect of 0.5 would hardly be considered that meaningful, even if Cohen called it a "medium" effect. Obviously, your research is not finding the "norm" effect of 0.8 that other researchers are finding, and *that is the important matter, not some arbitrary designation as to what constitutes small, medium or large effects.*

So when interpreting effect sizes (whether Cohen's **d** or other), always *contextualize* your effects in relation to similar research in other areas to give the reader a sense of where your result stands relative to the field of study. Otherwise, your readers or listeners will not know how to evaluate your findings. There are usually no absolutely "big" or absolutely "small" effects (except in the limiting sense where each reaches its maximum and minimum). There are only meaningful effects relative to the research area under investigation. END.