EFFECT SIZE AND STATISTICAL INFERENCE

Jeffrey A. Witmer
Mathematics Department
Oberlin College, Oberlin, OH 44074
jeff.witmer@oberlin.edu

There is wide agreement that P-values are over-used. Most introductory statistics courses spend a lot of time helping students understand how to find a P-value, and a good deal of time on thinking about what a P-value means (and doesn't mean). What matters in practice is the size of the effect. Effect sizes are just as important as P-values and should receive more attention.

P-VALUES VERSUS EFFECT SIZES

P-values are everywhere. The fact that people (students in particular) don't understand p-values is well documented. [See, for example, Wasserstein and Lazar (2016).] But even when p-values are understood, they are overrated. Quite often we know that the null hypothesis is not true. For example, a popular introductory textbook introduces hypothesis testing with an example in which we are to test H_0: μ = 80 versus H_a: μ > 80, where the variable is arsenic level, in parts per billion, in chickens sold in grocery stores. [See Lock (2017), page 265.] When reading this example, a student should not be asking "Is H_0 true?" but rather "Is it sensible to pretend that H_0 is true?" As George Box said, "All models are wrong; some models are useful." What we should care about is whether or not the effect being studied is large.

One sample mean

Returning to the chicken arsenic example, it is virtually impossible that the mean arsenic level in all chickens in the population is 80. The mean, μ, might well be between 79.9999 and 80.0001, but it is not *exactly* 80.000000... We know that. So why would anyone test H_0 μ=80? We would only care about such a test because the difference between μ and 80 might be small enough that it is OK to treat μ as being equal to 80.

The effect size here is the difference between μ and 80, scaled by the standard deviation. That is

$$\frac{|\mu - 80|}{\sigma}$$

If the effect size is small, say less than 0.20, then we would probably be willing to treat μ as being equal to 80. Or maybe some health official will only consider the difference between μ and 80 to be ignorable only if the effect size is less than 0.10. That's fine. If we report the observed (sample) effect size, then the reader can make his or her own conclusion.

On the other hand, if we only report the p-value then we've said essentially nothing. We know that even a small difference between μ and 80 will appear as "statistically significant" if the sample size is large enough. Consider the t-test statistic, which is directly related to the sample size:

$$\mathbf{t} = \frac{|\bar{\mathbf{y}} - 80|}{\mathbf{s}/\sqrt{\mathbf{n}}} = \sqrt{\mathbf{n}} * \frac{|\bar{\mathbf{y}} - 80|}{\mathbf{s}} = \sqrt{\mathbf{n}} * (\text{sample effect size})$$

Comparing two populations

Probably the most common statistical test involves comparing two sample means in order to infer whether or not two population means are equal. Again, we usually know that $\mu_1 - \mu_2 \neq 0$; what we care about is whether the difference is small enough that it can be ignored. In the two-sample setting there are many ways to define the effect size, with Cohen's d being widely used. Cohen's d is defined for populations as

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

$$\frac{|\mu_1 - \mu_2|}{\sigma}$$

and is defined for samples as

$$\frac{|\bar{y}_1 - \bar{y}_2|}{s}$$

where in each case we use the pooled standard deviation in the denominator.

Guidelines for Cohen's d are that the effect is

Small if d is around 0.2, which corresponds to $\Pr(Y_2 > \mu_1) = \Pr(Z > 0.2) \approx 40\%$

Medium if d is around 0.5, which corresponds to $\Pr(Y_2 > \mu_1) = \Pr(Z > 0.5) \approx 30\%$

Large if d is around 1, which corresponds to $\Pr(Y_2 > \mu_1) = \Pr(Z > 1) \approx 15\%$

Whether an effect is small, medium, or large depends on the context, of course. An average difference between groups of half of a standard deviation might be important in one case (e.g., a salary comparison of men and women) and not so important in another (e.g., a comparison of hours spent watching television on the weekend for men versus women).

Statisticians know that sample size matters along with the magnitude of any observed difference. I collected data by asking 26 statisticians about the following situation. Consider two drugs, A and B. Each is compared to a placebo in a clinical trial. For the trial involving drug A the two sample sizes were each 50 (i.e., 50 patients got the drug and 50 got the placebo). Drug A did better than the placebo, with the p-value from the clinical trial being 0.03. For drug B the two sample sizes were each 12, rather than 50. Drug B did better than the placebo; the p-value was 0.07. Each clinical trials yields evidence of a drug being better than the placebo, but we can't be sure that either drug actually is better than the placebo. If your company could invest in a large follow-up clinical trial, which drug, A or B, would you study? A related question is "Which drug, A or B, would you ask your doctor to prescribe for you?"

The 26 statisticians I surveyed were evenly divided between A and B, with each being chosen by 13 of them. Note that if $n_1=n_2=50$ and the (population) effect size is 0.44, then 0.03 is the median p-value that a study would produce. If $n_1=n_2=12$ and the (population) effect size is 0.77, then the median p-value is 0.07. I suspect that this is why many statisticians choose B, despite its larger p-value. To get a p-value of 0.07 with sample sizes of only 12, the effect size is likely to be somewhat large, with the two populations being shifted by perhaps around three-fourths of a standard devation, but with $n_1=n_2=50$ a small p-value can easily arise when the effect size is modest. My concern is that I we don't teach statistics students to think along these lines.

STATISTICS EDUCATION – THE STATUS QUO

A couple of years ago I took a convenience sample of introductory statistics textbooks. All of them discussed hypothesis tests, p-values, and confidence intervals at length. Some included cautions about interpretation of a p-value. Some stressed the importance of looking at a confidence interval to estimate how large the difference between that $\mu_1 - \mu_2$ might be. But only 2 out of 18 books included a formal treatment of effect size.

We statistics teachers often talk about how much one mean might differ from another, but we rarely talk about that difference in terms of the underlying variability in the data. For people who think that inherent variability is important to understand, this is a surprising state of affairs.

STATISTICS EDUCATION – THE FUTURE?

In years gone by, I spent time in my introductory courses talking about power, which is a difficult concept (as if p-values aren't hard enough!). But I've come to understand that effect size is both more important than power and is easier to grasp.

When I teach correlation, I always spend time with my students looking at scatterplots and developing a sense of what a correlation of r = 0.80, say, looks like. [See, for example, http://www.istics.net/Correlations/.] Likewise, I want my students to get a feel for what different effect sizes look like. One way to do this is to present overlapping normal curves, as in Figure 1.

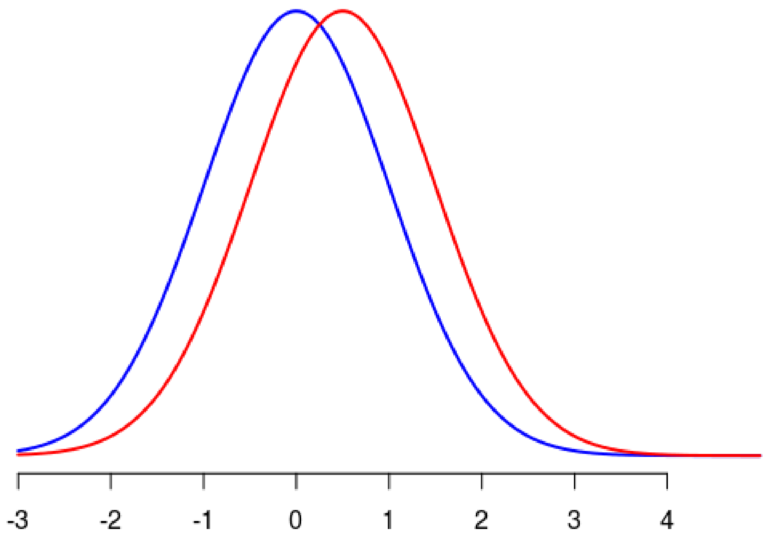

Figure 1: Overlapping normal curves when the effect size is 0.5.

Another way to develop a sense of effect size is to look at parallel dotplots and boxplots. Figures 2 and 3 show the kinds of graphs that I have in mind. Figure 2 shows a moderate effect size (and small p-value), corresponding to the Drug A trial, while Figure 3 shows a fairly large effect size (and somewhat small p-value), corresponding to the Drug B trial.

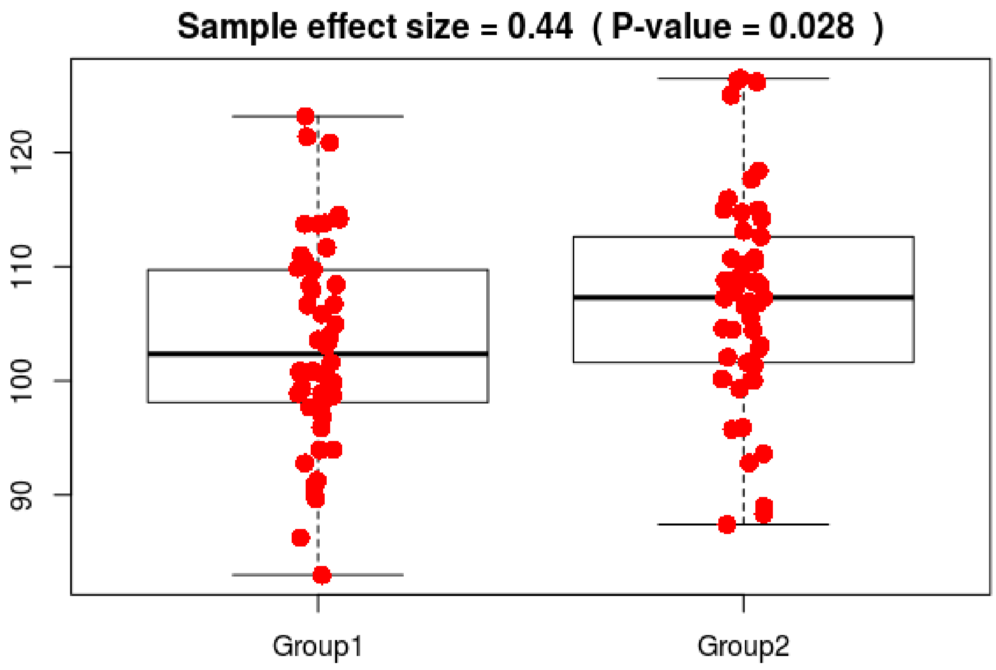

Figure 2: Parallel boxplots when the sample effect size is 1.2 and the p-value is 0.003.

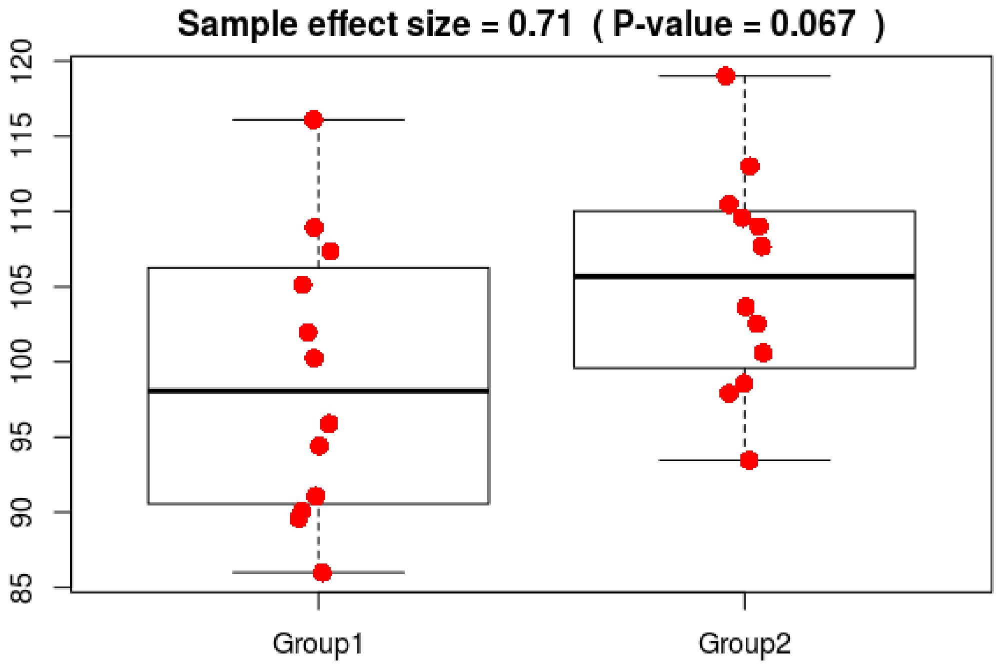

Figure 3: Parallel boxplots when the sample effect size is 0.71 and the p-value is 0.067.

Of course, effect size arises in situations beyond the study of means. When teaching categorical data analysis I present Cramer's V:

$$\mathbf{V} = \sqrt{\frac{\chi^2}{\mathbf{n} * (\mathbf{m} - 1)}}$$

where m = min(r , c), with guidelines for small, medium, and large effects sizes being V = 0.10, 0.25, or 0.50, respectively. In the setting of regression, I use the correlation coefficient as a measure of effect size. R^2 is useful in multiple regression. R^2 can also be used in analysis of variance, but there I like to extend Cohen's d to multiple groups, comparing the largest and smallest samples means to each other, scaled by the square root of mean squared error.

CONCLUSION

P-values are not going away, nor are confidence intervals. But they are not enough, and they can distort the picture. We need to be teaching students about effect size.

REFERENCES

Lock, R.H., et al. (2017). *Statistics: Unlocking the Power of Data*. New York: Wiley.

Wasserstein, R. L, and Lazar, N. A., (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129-133, DOI: 10.1080/00031305.2016.1154108