Effect Size Really Does Matter

Jeffrey A Witmer
Oberlin College, 10 N. Professor St. Oberlin, OH 44074

Abstract

The world is awash in p-values. 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). But we almost always know that the null hypothesis is false; what matters in practice is the size of the effect. Effect sizes are just as important as p-values but receive little attention. That should change.

Key Words: *Effect size, p-value, statistics education*

1. Effect size, not just p-values

Statisticians, and particularly statistics educators, make heavy use of hypothesis tests and p-values. But almost always, when a null hypothesis is tested we know in advance that it is false. Everything is statistically significant if the sample size is large enough. For example, when conducting a chi-square test, the test statistic goes up by a factor of 2 when the sample size is doubled. So what?

What matters is whether or not the effect that is being studied is large. Don't tell me that H_0 was rejected; tell me the effect size.

1.1 Comparing two populations

To keep things simple, let us concentrate on the setting of comparing two populations.[1] A commonly used measure of the size of the effect is Cohen's d, defined for the populations as

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

and defined for the samples as

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

We could use a pooled SD, but for simplicity we can just use the larger of the two sample SDs.

Guidelines for Cohen's d are that the effect is

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

[1] In a regression setting we might report r or the percentage of variation explained. In an ANOVA setting we might report eta-squared = SS(Trt)/SS(Total), i.e., the percentage of variation explained.

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 $\Pr(Y_2 > \mu_1) = \Pr(Z > 1) \approx 15\%$

Interpretation of effect size depends on context, of course. A shift of 1 standard deviation might be very important in a medical setting and might be unimportant in another setting. That’s OK. Indeed, that’s good. We should get our students to think about whether or not an effect is important, in context, and not just whether or not a difference is statistically significant.

2. Where we are

2.1 What do we teach?

Milo Schield put together a list of 12 big ideas and Rossi Hassad drafted a list of 11 key concepts for the session on Advancing Statistical Literacy at JSM in Chicago, but neither of them featured effect size.

I took a convenience sample of 19 introductory statistics textbooks and checked the index for “effect size.” I also paged through the inference chapters in the books. Only 3 of the 19 included a formal treatment of effect size as a topic – and I wrote one of those books, so 2/18 might be a better estimate of the fraction of textbook authors who promote student understanding of effect size.

2.2 What do we do in practice?

Mike Malek-Ahmadi, in a note on ASA Connect, responded to the question “How to upset the statistical referee?” by writing (with emphasis added):

> Having served as a reviewer for several different journals in the last few years, here are some things I have observed.
>
> 1. ***Interpreting the strength of an association based on how low a p-value is***. This is a very common mistake, especially among clinical papers.
>
> 2. Related to my point in #1, ***many papers still lack the reporting of effect sizes*** when comparing groups on continuous variables. I will say that papers submitted to psychology journals tend to be a little bit better about reporting a Cohen's d value with a t-test or an eta-squared with an ANOVA, but this practice is still lacking in other medical and health disciplines.

3. How do we react to data?

Consider two drugs, A and B. Each is compared to a placebo in a clinical trial. For the drug A trial 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; the p-value was 0.03. For drug B the two sample sizes were each 12 (i.e., 12 patients got the drug and 12 got the placebo). Drug B did better than the placebo; the p-value was 0.07. Each drug looks better than the placebo, but we can’t be sure that either drug actually works.

Which drug is more promising? Which should be used in a large follow-up trial? Which would you ask your doctor to prescribe for you? Etc.

I collected data on this question from 26 statisticians over the past couple of years. The data here make the comparison of A versus B a tough call, so I was not surprised that 13 statisticians chose A and 13 chose B.

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. That's why I like B: To get a p-value of 0.07 with only n_1=n_2=12, the effect size is likely to be somewhat large, perhaps around 0.75, 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 don't think that many statistics students are taught to think along these lines.

4. What should we teach?

Effect size is related to power but is more important. We often teach, or at least try to teach, power to our students, but power is a very difficult concept for them. Effect size is easier to grasp and is more important.

So what does an effect size of 1.2, for example, look like? Figures 1 - 4 are some graphs that I want more students (and professors) to see. Going from Figure 1 to Figure 2, the p-value goes up by almost an order of magnitude as the effect size goes down somewhat. Comparing Figure 3 to Figure 2, a much smaller effect size coupled with much larger sample sizes yields (essentially) the same p-value. Finally, Figure 3 corresponds to the Drug A trial while Figure 4 corresponds to the Drug B trial. I find the visual evidence in Figure 4 to be more compelling.

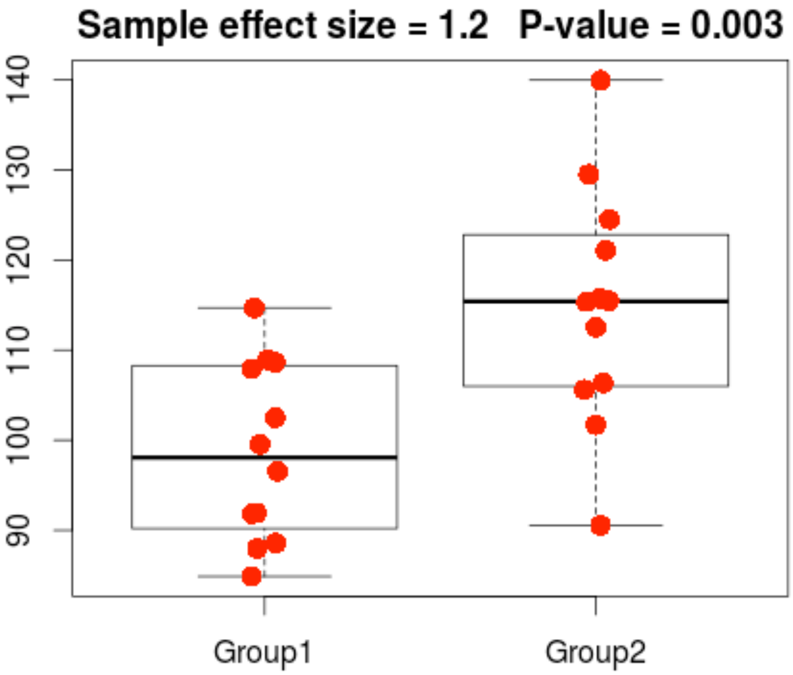

Figure 1: Parallel boxplots when the sample effect size is 1.2 and the p-value is 0.003.

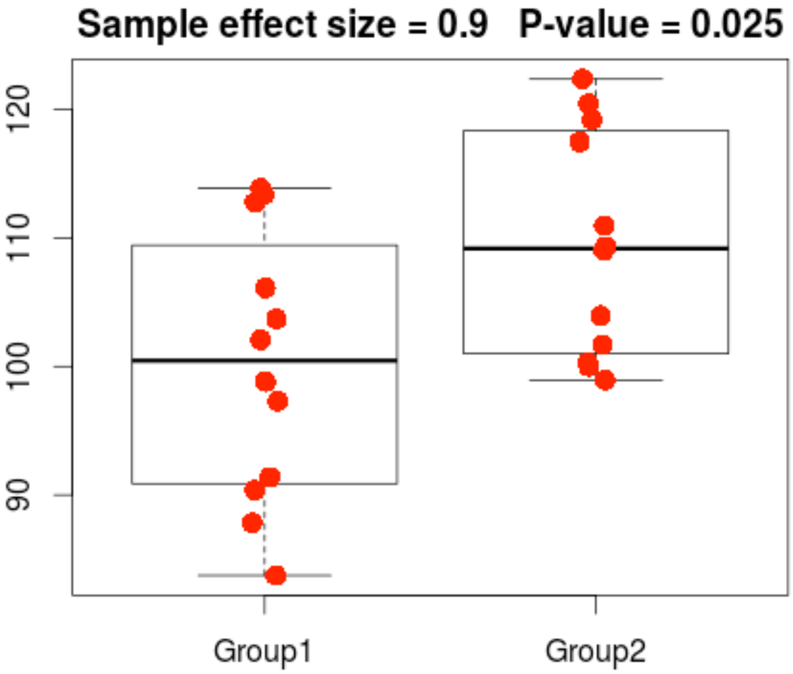

Figure 2: Parallel boxplots when the sample effect size is 0.9 and the p-value is 0.025.

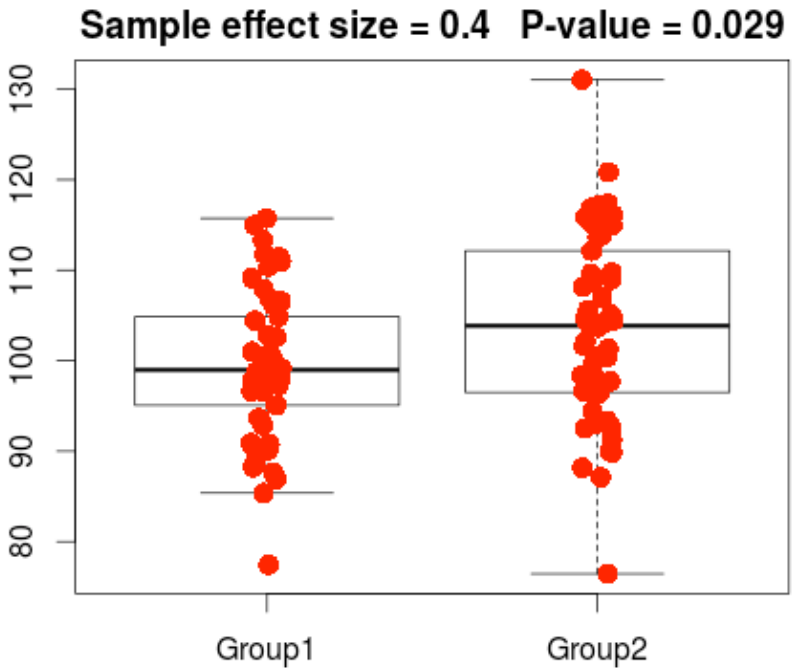

Figure 3: Parallel boxplots when the sample effect size is 0.4 and the p-value is 0.029.

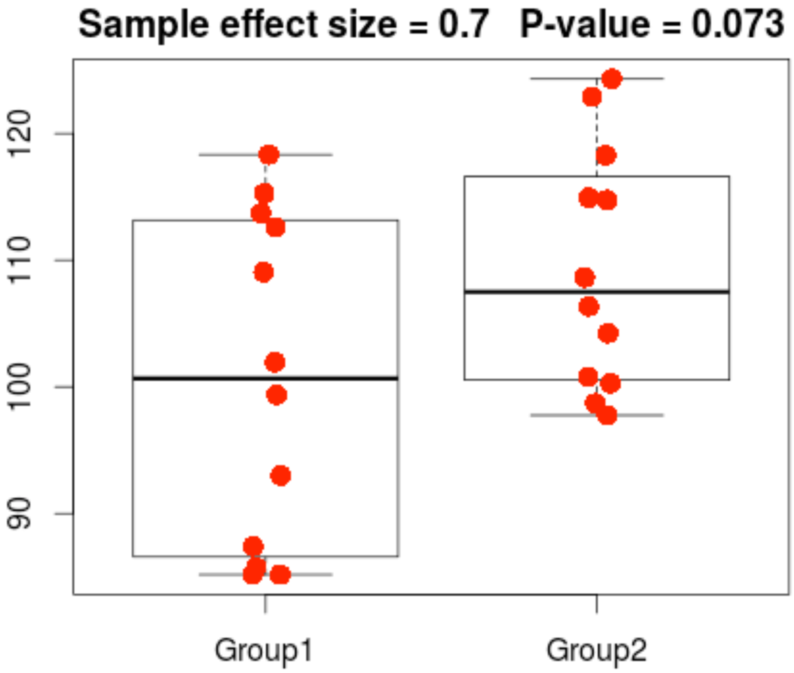

Figure 4: Parallel boxplots when the sample effect size is 0.7 and the p-value is 0.073.

References

Schield, Milo (2016), “Twelve Big Ideas for Introductory Statistics.” JSM Session 129, Chicago.

Hassad, Rossi A. (2016), “Making Connections and Understanding Statistics: Students' Ratings of the Utility of Key Concepts in the Introductory Statistics Course.” JSM Session 129, Chicago.

Malek-Ahmadi, Mike, 29 January 2016 comment posted on ASA Connect.