What's the True Effect Size?

Nathan (Nat) Goodman

September 15, 2019

What's the true effect size? That's my bottom line question when doing a study or reading a paper. I don't expect an exact answer, of course - this is statistics after all. What I want is a probability distribution telling where the true effect size probably lies. I used to think effitdence intervals answered this question, but they don't except under artificial conditions. A better answer comes from Bayes's formula.

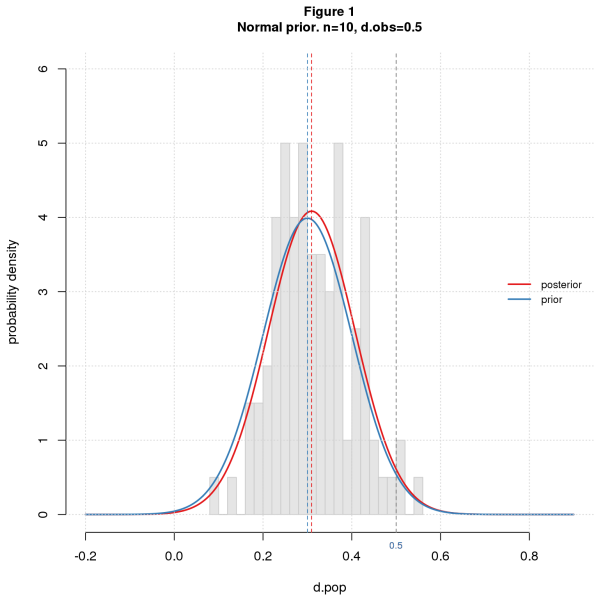
effitdence intervals, like other standard methods such as the t-test, imagine that we're repeating a study an infinite number of times, drawing a different sample each time from the same population. That's not how I do science, or rather, it's not how I do basic, exploratory research. What I do is run a study, compute an observed effect size from my data, and try to figure out what true effect sizes may have led to my observation. Mathematically, I consider all possible true effect sizes and ask, "What's the probability of getting my observed effect size for each true effect size?".

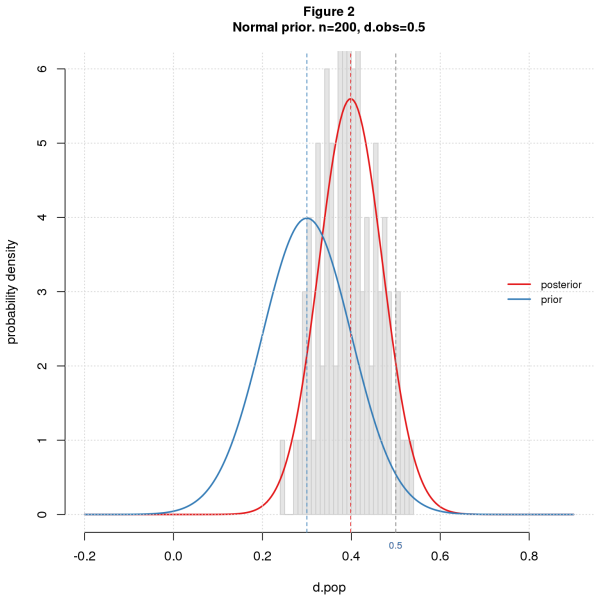
I model the scenario as a two stage random process. The first stage selects a population (aka "true") effect size, d_{pop}, from a distribution (the *prior* in Bayesian terminology); the second carries out a study with that population effect size resulting in an observed effect size, d_{obs}. Imagine I repeat the process infinitely, recording d_{pop} and d_{obs} each time. The result is a table with columns d_{pop} and d_{obs} showing which d_{pop}s give rise to which d_{obs}s. Now, pick a value for d_{obs}, say 0.5, and limit the table to rows where d_{obs} is near 0.5. The distribution of d_{pop} from this subset is the answer to my question (the *posterior* in Bayesian-speak).

For this post, the studies in the second stage are simple two group difference-of-mean studies with equal sample size and standard deviation, and the effect size statistic is standardized difference (aka *Cohen's d*). Concretely, each study selects two random samples of size n from standard normal distributions, one with $mean = 0$ and the other with $mean = d_{pop}$, and calculates d_{obs} as the difference of the group means divided by the pooled standard deviation.

Now for the cool bit. The Bayesian approach lets us pick a prior that represents our best guess of the distribution of true effect sizes in our research field. From what I read in the blogosphere, the typical population effect size in social science research is 0.3. I'll model this as a normal distribution with $mean = 0.3$ and small standard deviation, say 0.1.

Figures 1 and 2 show the results for small and large samples ($n = 10$ or 200) for $d_{obs} = 0.5$. Each figure shows a histogram of simulated data, the prior and posterior distributions (blue and red curves), the medians of the two distributions (blue and red dashed vertical lines), and $d_{obs} = 0.5$ (gray dashed vertical line). The posterior and data match pretty well, a good sign that my software works. For $n = 10$, the posterior mean is barely above the prior, while for $n = 200$, it's much closer to the observed (0.5).

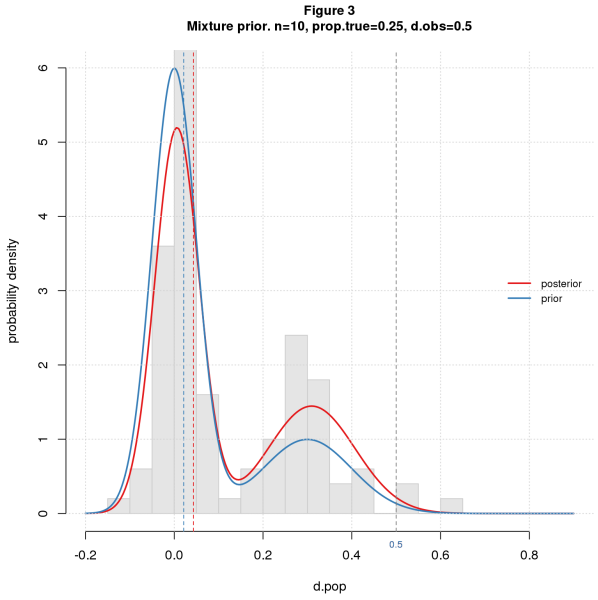


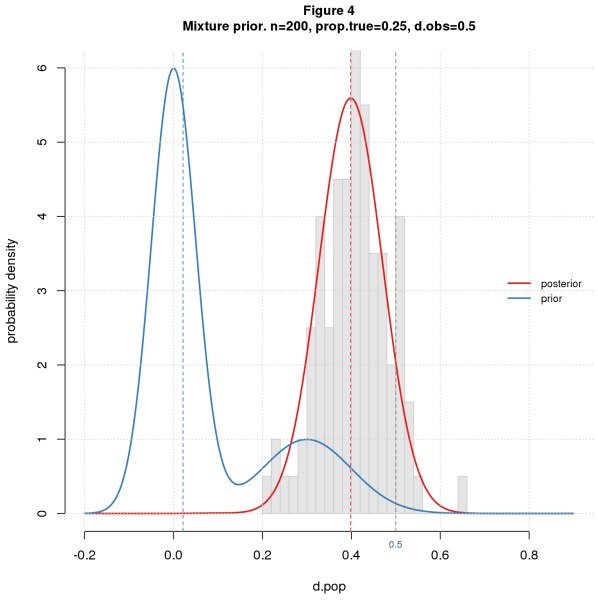


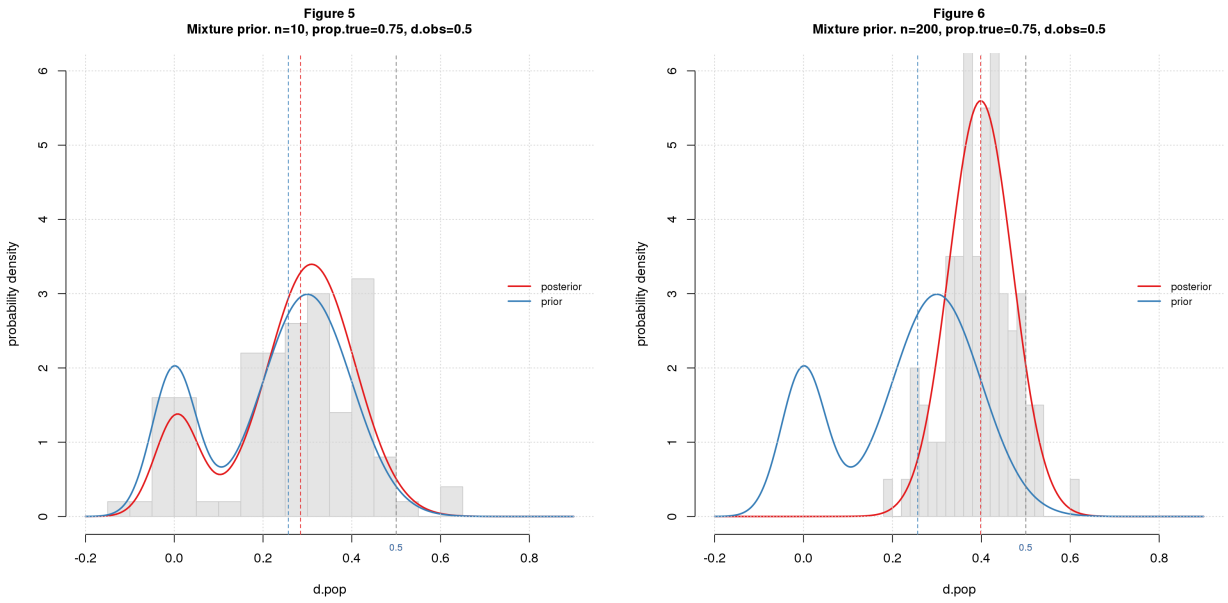
I think of it as a tug-of-war: for small samples, the prior wins and pulls the posterior down; for large samples, the data is stronger and keeps the posterior up. Completely intuitive.

But wait. I forgot an important fact: some of the problems we study are "false" ("satisfy the null") with effect sizes of 0 or close to 0, while others are "true" with effect sizes as above. No worries. I model the "false" effect sizes as normal with $mean = 0$ and very small standard deviation, say 0.05, and the totality of effect sizes as a *mixture* of this distribution and the previous one. To complete the model, I have to specify the proportion of "true" vs. "false" problems.

Figures 3-6 show results for small vs. large samples (10 or 200) and low vs. high proportions of true problems (25% or 75%) for $d_{obs} = 0.5$.







The distributions are more complicated than the previous model, but they show the same tug-of-war: small samples lose to the prior, while big samples hold their own.

The priors are bimodal, reflecting the two classes. For small samples, the posterior is also bimodal and barely different from the prior. For big samples, the posterior is unimodal, essentially normal; further, the posterior medians are almost identical for all three priors. Though not obvious from the figures, the entire posterior distributions are nearly identical. In other words, if the sample is big enough, the prior barely matters.

These results are simple, intuitive, and give reasonable answers. The analysis can compute the probability that the true effect size is less than 0 or any other smallest effect size of interest, which can be used for a yes/no classification - akin to a significance test - if you're so inclined. The code for the core Bayesian analysis (available here) is simple, too, just a few lines of R. What more do you want?

The real head scratcher is why this isn't the standard way to estimate true effect size.

Comments Please!

Please post comments on Twitter or Facebook, or contact me by email natg@shore.net.