Abstract Title Page

Title: Statistical power calculations for evaluations that will interpret impact estimates using the BASIE (BAyeSian Interpretation of Estimates) framework

Authors and Affiliations:
John Deke, Mathematica Policy Research
Mariel Finucane, Mathematica Policy Research

Abstract Body

Research Context:

We know some evaluation findings are more reliable than others. But sorting out which findings deserve special attention is challenging. For nearly 100 years, the null hypothesis significance testing (NHST) framework has been used to determine which findings deserve attention (Fisher 1925; Neyman and Pearson 1933). Under NHST, findings worthy of attention are called "statistically significant" if the *p*-value is less than 0.05. But misinterpretation of *p*-values is so widespread that in 2016 the American Statistical Association (ASA) issued a statement on the subject (Wasserstein and Lazar 2016; Greenland et al. 2016).

The ASA statement referenced Bayesian methods as a potential solution to *p*-value misinterpretation, but did not provide *specific* guidance. This omission was perhaps unavoidable, given the diverse contexts in which *p*-values are used. However, the offhand reference to Bayesian methods raised concerns among some researchers. For example, Ionides et al. (2017) expressed concern that, although it is possible to use Bayesian methods in a way that is compatible with deductive science, "that is not currently the mainstream approach in the Bayesian community." Furthermore, the subjectivist Bayesian definition of probability as the intensity of one's personal belief regarding the truth of a proposition (de Finetti 1974) may seem inconsistent with objective science.

Fortunately, it is possible for evaluators to answer the question "what is the probability the intervention had a meaningful effect, given our impact estimate?" without misinterpreting p-values and without resorting to a subjectivist approach. In a forthcoming brief prepared for the Office of Planning Research and Evaluation (OPRE), an office of the Administration for Children and Families in the Department of Health and Human Services, we describe an alternative framework for interpreting impact estimates, which we call BASIE (BAyeSian Interpretation of Estimates).

Theoretical Background:

With BASIE, we seek to find common ground between the "Bayesian" and "frequentist" frameworks. BASIE is "Bayesian" in that it uses Bayes' Rule to leverage prior *evidence* (not *personal belief*) to assess the likelihood that an evaluated intervention has meaningful effects. But the framework might also appear "frequentist" because we define probability in terms of relative frequency (not belief), and emphasize the important of reporting impact estimates based only on study data.

BASIE includes 5 components providing specific guidance on 5 issues: (1) defining probability, (2) selecting a prior, (3) reporting impact estimates, (4) interpreting estimates, and (5) conducting sensitivity analyses (Table 1). While the components of BASIE are well-established in the literature (Gigerenzer and Hoffrage 1995; Gelman 2011, Gelman and Shalizi 2013; and Gelman 2016), we have brought them together in a way that provides an innovative solution to the problem of providing objective, scientific interpretation of findings from high stakes evaluations without some of the more controversial aspects of Bayesian methods.

Table 1. Components of the BASIE framework for impact evaluation

Component	BASIE involves…	BASIE does NOT involve…	Notes
Probability	A relative frequency (e.g., “21 out of 30 relevant studies in the WWC”)	Personal belief (e.g., “I am 70% sure that…”)	In this framework, a probability can generally be thought of as a number that is based on things that can be counted. When communicating probabilities, it’s important to make sure we are clear about what’s being counted.
Prior	Evidence	Personal belief	The prior evidence could be combined or refined using a model, but the fundamental basis of the prior is evidence, not belief.
Reported impact estimate	*Both* the impact estimated using only study data *and* the “shrunken” impact estimate that incorporates prior evidence	*Just* the impact estimated using only study data *or* the “shrunken” impact estimate	The relevance of the prior evidence base to the current study will dictate which estimate should be highlighted.
Interpretation	Bayesian posterior probabilities, Bayesian credible intervals	Statistical significance, *p*-values	Statistical significance and *p*-values are too easily misinterpreted and do not tell us what we really want to know, which is the probability that the intervention truly improved outcomes. We do not list confidence intervals in the “BASIE does NOT involve” column because they are useful for communicating the precision of an impact estimate, but we do not suggest using confidence intervals for interpretation/inference.
Sensitivity analysis	Reporting sensitivity of impact estimates and posterior probabilities to the selection and modeling of prior evidence	Reporting a single answer with no assessment of its robustness	Increasing the sample size of a study will reduce sensitivity to imperfect prior evidence.

Purpose and Research Questions:

The power analyses we conduct to inform evaluation design should be aligned with the framework that will be used to interpret evaluation findings. Historically, findings have been interpreted using the NHST framework and power analyses have been aligned with that framework. For evaluations that will use BASIE to interpret findings a new approach to power analysis is needed.

Our three main research questions are:

1. How should prior evidence be selected and analyzed to form a prior distribution?
2. What is the power to detect a given effect size under the BASIE framework, using prior evidence drawn from the What Works Clearinghouse (WWC) database?
3. How sensitive is power to which prior evidence is used to calculate posterior probabilities?

Research Design and Methods:

Before we can conduct power analyses, we must first define what it means to "detect" an effect using a posterior probability. An effect will be detected if there is a sufficiently large probability that the true effect is greater than the smallest magnitude deemed meaningful. For example, an evaluator might specify that an effect is detected if there is a 97.5 percent probability that the true effect is greater than zero (this is what many people misinterpret $p < 0.05$ from a two-tailed test to mean).

To calculate the power to detect a specified effect size (d) given a specified sample size, we use this Monte Carlo simulation algorithm:

1. Generate an impact estimate ($\hat{d}$) drawn from the normal distribution with mean d and variance based on a specified sample size, ICC, and regression R^2.
2. Calculate a posterior probability using $\hat{d}$ and a prior distribution based on prior evidence from the WWC. To select prior evidence, we must: (1) decide which evidence to include and (2) make statistical adjustments for variations in the accuracy and reliability of past findings. We will suggest considerations and strategies for each of these.
3. Given the posterior probability and criterion for detecting an effect, assess whether an impact has been detected.
4. Repeat steps 1–3 10,000 times and calculate the proportion of times that an effect is detected. That proportion is power.

We examine the second research question using Monte Carlo simulations to assess how power changes if we modify which prior evidence is specified in step 2.

Findings and Conclusions:

We find that:

1. When selecting prior evidence, it is generally best to cast a wide net and include a broader range of past studies than what might first seem intuitive.
2. Prior evidence needs to be adjusted for variation in accuracy and reliability. Once this is done, we find that the distribution of findings in the WWC is centered near zero
3. If we specify that an effect is detected if there is a 97.5 percent probability that the true effect is greater than zero, then power under BASIE with an evidence-based prior tends to be a little lower than under the NHST

4. The power to detect effects under the BASIE framework becomes less sensitive to prior selection as sample size increases

References

de Finetti, B. *Theory of Probability: A Critical Introductory Treatment.* New York: Wiley, 1974.

Fisher, R. A. *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd, 1925.

Gelman, A. "Induction and Deduction in Bayesian Data Analysis." Special topic issue, Statistical Science and Philosophy of Science: Where Do (Should) They Meet in 2011 and Beyond? *Rationality, Markets and Morals,* vol. 2, 2011, pp. 67–78.

Gelman, A., and C. Shalizi. "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology,* vol. 66, 2013, pp. 8–38.

Gelman, A. "What is the 'True Prior Distribution'? A Hard-Nosed Answer." *Statistical Modeling, Causal Inference, and Social Science* blog, 2016. Available at http://andrewgelman.com/2016/04/23/what-is-the-true-prior-distribution-a-hard-nosed-answer/. Accessed June 7, 2018.

Gigerenzer, G., and U. Hoffrage. "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats." *Psychological Review,* vol. 102, 1995, pp. 684–704.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. "Statistical Tests, *p*-Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology,* vol. 31, no. 4, 2016, pp. 337–350.

Neyman, J., and E. S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society,* Series A, Containing Papers of a Mathematical or Physical Character, vol. 231, 1933, pp. 289–337.

Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA's Statement on *p*-Values: Context, Process, and Purpose." The *American Statistician,* vol. 70, no. 2, 2016, pp. 129–133.