Retrospective power revisited
a Bayesian false non-significance rate

ROBERT GRANT
BayesCamp
April 12, 2018

Abstract

Statisticians in many fields are often asked to answer the question of whether a study, having obtained a non-significant hypothesis test, might in fact have been in error. Assessing type II error or power retrospectively is meaningless, but it seems a reasonable question to ask, and might be answered by examining Bayesian probabilities of such results in a hypothetical further identical study. I suggest a starting point here and propose the simpler term "false non-significance rate".

Keywords: Bayesian statistics , study design , statistical power

Introduction and notation

Power, the probability of obtaining a significant hypothesis test result if the population test statistic is equal to a minimally important value, is a ubiquitous concern in many fields of applied statistics, including my own, biomedical research. It is usually operationalised as a frequentist concept, and so calculating it after the study has been conducted — so-called retrospective power — is meaningless.

Let θ be the true value of the test statistic, and σ its true standard error given the sample size of the completed study. δ is the minimally important value upon which the power calculation for the study was based, and $\hat{\theta}$ and $\hat{\sigma}$ are the completed study's estimates of theta and sigma respectively. Here, for simplicity, they are regarded as transformed so that the null hypothesis is $H0 : \theta = 0$. I will also assume normality of the sampling distribution for illustrative purposes, although the formulas do not require that.The target difference in the original sample size calculation is delta.

$1 - P(\hat{\theta} < 1.96\,\hat{\sigma} \mid \theta = \delta, \sigma)$ is either 0 or 1 once you know $\hat{\sigma}$ and $\hat{\theta}$. Nevertheless there seems to be a widespread urge to answer this question: "could my study's non-significant result have been a mistake?" This seems a reasonable question, but to answer it requires something other than power.

Previous work

Ioannides considered possible ways of assessing the probability of the truth or falsehood of study results in 2005 in his widely cited paper "Why most published research findings are false"[1]. His formulas take into account various other forms of bias such as unacknowledged occult multiplicity, publication and reporting bias. However, he considers only a dichotomised finding (significant / not) and true value (effect / no effect), which limits the applicability of the approach to individual studies. This was one aspect criticised afterwards by Goodman & Greenland[2].

In a 2008 paper, expanding on an article in American Scientist magazine, and published only on the first author's own website, Gelman and Weakliem considered underpowered studies and set out the probabilities of various types of error: the familiar I and II as well as errors of magnitude (type M) and of sign (type S)[3]. They conclude that the system of type I and II has not been helpful. In most cases they consider, the probabilities of type M or S errors turn out to be so high as to call any conclusion of the study into question. In the same year, Ralph O'Brien spoke at the JSM conference on "crucial type I and II error rates". This proposal reverses the familiar formulas through Bayes' Theorem. The contemporaneous discussion on the author's blog sets the scene[4].

Objective

To address the question of whether a study's non-significant result could be a type II error, we must deal with a theoretical identical future study, in the same prospective way that classical power is calculated. The completed study's estimates of the parameter and its standard error are fixed values, and the true population values are unknown, but we can establish a distribution for the estimates of an identical future study.

The question "what is the probability that my study's non-significant result is wrong?" then can be rephrased as "given what we can infer about the true parameter given our data, what is the probability that the true effect is as big or bigger than the target difference and yet an identical study would yield a non-significant result?" This is then a form of Bayesian posterior predictive model checking.

Some theory

Let θ^* and σ^* be the estimates arising from an identical study. We are interested in: $P(\theta^* < 1.96\,\sigma^* \mid \hat{\theta}, \hat{\sigma}, \theta \geq \delta)$, which we can calculate from the sampling distribution $P(\theta, \sigma \mid \hat{\theta}, \hat{\sigma})$ and the conditional $P(\theta^*, \sigma^* \mid \theta, \sigma)$ and by integrating out the unknown θ and σ.

To emphasise the distinction from the type II error rate, I propose the clear term "false non-significance rate" for this.

There is a further complication to consider, which mirrors Ioannides's various biases. Retrospective power is usually only considered when $\hat{\theta} < 1.96\,\hat{\sigma}$ (that is, a non-significant result), and either $\hat{\sigma}$ is larger than expected (including problems of sample size) or $\hat{\theta}$ is close to, but smaller than, δ. This introduces a bias because we will only ask the retrospective power question of a subset of the possible values of $\{\hat{\theta}, \hat{\sigma}\}$. To counter this requires us to introduce a joint prior on $\{\hat{\theta}, \hat{\sigma}\}$, and so derive a posterior distribution for $\{\theta^*, \sigma^*\}$. This could be informed by previous research or opinion in the usual way, but it does not make sense for it to be diffuse. Sensitivity analysis with various priors is advisable.

It is almost always the case that there is some other information on anticipated findings, so that $\{\hat{\theta}, \hat{\sigma}\}$ is not the only information about $\{\theta, \sigma\}$ and hence $\{\theta^*, \sigma^*\}$. We should attempt to incorporate this as a prior distribution because it is hard to interpret retrospective power in the context of other studies, in the way that we expect people to interpret study findings (without informative priors).

Some discussion

As a former medical statistician in a university and hospital setting, I was regularly called on to advise on sample size and power calculations. I was, and still am, convinced that the great majority of these calculations were uninformative acts of sophistry, performed for the comfort of the tutor, ethics committee or funding body, and based on such an accumulation of assumptions as to be meaningless. My message to all colleagues and students in such a situation (because it is not their fault to expect a simple answer) is to think very carefully and in depth about what they are trying to investigate, and what they would do having found various potential results. This critical thought helps to inoculate them against the lure of simplicity that comes from one calculation on one hypothesis, under one set of assumptions. The calculations set out in this paper are no different, and require careful justification for all the assumptions behind them. Indeed, I approach publication of this proposition with some trepidation, lest what is intended as a stimulating exercise in defining slippery concepts is reduced instead to a catch-all formula that permits retrospective power calculations to proceed under a new name, and unhindered by cerebral activity. I hope that by encoding the most difficult part of this - the retrospective power bias - as an informative prior distribution, researchers will be forced to slow down and consider what has happened with their study, and what they are seeking to achieve by such calculation, very carefully, like the QWERTY keyboard, intended to slow typists sufficiently to avoid jamming of keys on a manual typewriter.

A final note of caution concerns the target difference δ, which appears in all the formulas here. A review of methods to establish this value, in the Health Technology Assessment series, is essential reading for everyone working with sample size and power calculations, because our recommendations for designing future studies and interpreting completed ones are undermined by irrelevant or unreliable target differences[5].

References

1. Ioannidis JPA (2005). Why Most Published Research Findings Are False. *PLoS Med* **2**(8): e124.
`journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124`

2. Goodman S, Greenland S (2007). *Assessing the unreliability of the medical literature: a response to "Why most published research findings are false"*. Johns Hopkins University, Department of Biostatistics Working Papers, number 135.
`biostats.bepress.com/jhubiostat/paper135/`

3. Gelman A, Weakliem D (2008). *Of beauty, sex, and power: Statistical challenges in estimating small effects*.
`stat.columbia.edu/~gelman/research/unpublished/power4r.pdf`

4. Gelman A, and various contributors. *What is the point, if any, of retrospective power calculations?*. Statistical Modeling, Causal Inference, and Social Science, 26 December 2008.
`andrewgelman.com/2008/12/26/what_is_the_poi/`

5. Cook JA, Hislop J, Adewuyi TE, Harrild K, Altman DG, Ramsay CR, Fraser C, Buckley B, Fayers B, Harvey I, Briggs AH, Norrie JD, Fergusson D, Ford I, Vale LD (2014). Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technology Assessment*; **18**(28).
`njl-admin.nihr.ac.uk/document/download/2002640`