# Retrospective power analysis using external information[1]

Andrew Gelman and John Carlin[2]

11 May 2011

> "Power is important in choosing between alternative methods of analyzing data and in deciding on an appropriate size of experiment. It is quite irrelevant in the actual analysis of data." – D. R. Cox, 1958.
>
> "Power should play no role once the data have been collected." – S. N. Goodman and J. A. Berlin, 1994.
>
> "Dismayingly, there is a large, current literature that advocates the inappropriate use of post-experiment power calculations as a guide to interpreting tests with statistically nonsignificant results." – J. M. Hoenig and D. M. Heisey, 2001.
>
> "Power is indeed irrelevant in interpreting completed studies." – S. J. Senn, 2002.
>
> "Power is not useful in data analysis." – R. V. Lenth, 2007.

## Introduction

The present article proposes an ideal that every statistical analysis be followed up with a power calculation to better understand the inference from the data. As the quotations above illustrate, however, our suggestion contradicts the advice of many respected statisticians. Our resolution of this apparent disagreement is that we perform retrospective power analysis in a different way and for a different purpose than is typically recommended in the literature.

The starting point of any power analysis is the postulated effect size. As explained in the references above, post-hoc power calculations commonly make one of the following assumptions:

- Assuming an effect size equal to the estimate from the data;
- Determining power under an effect size deemed to be substantively important; or
- Computing the effect size required to reach a specified power.

We agree with the expert consensus that these approaches can be more confusing than helpful if used to supplement an existing data analysis.

Our suggestion is slightly different: we recommend performing a power analysis after the data, but *based on an effect size that is determined from literature review or other information external to the data at hand*.

[1] We thank John Carlin and Deborah Mayo for helpful comments and the Institute of Education Sciences, Department of Energy, National Science Foundation, and National Security Agency for partial support of this work.

[2] Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, http://www.stat.columbia.edu/~gelman/

The purpose of our retrospective power analysis is not to salvage non-significant results by reassuring ourselves that our sample size was too small to discover what we were looking for—we agree with the researchers quoted above that a simple confidence interval does a better job at this task—but rather to interpret the results of an experiment in light of prior information.

**Recommended procedure**

Suppose you perform a study that yields an estimate y with standard error s. The usual procedure is to compute the ratio z=y/s and report statistical significance if $|z|>2$, and if $|z|<2$, report the results as inconclusive.[3] If there is any retrospective power analysis (a procedure which, as noted above, is generally frowned upon by statistical experts), it would be performed in response to a finding of non-significance.

In contrast, we recommend performing a retrospective power analysis for all studies—significant and non-significant alike. In either case, we would use external information (other available data, literature review, and modeling as appropriate to extrapolate this additional information to apply to the problem at hand) to hypothesize an effect size, A. As with the usual prospective power analysis, it can make sense to consider several plausible effect sizes and to perform a power calculation for each. From a Bayesian perspective, it would be natural to express the estimated effect and its uncertainty as a prior distribution, but in the present article we will take a classical approach and consider a range of point assumptions.

Our retrospective power analysis is based on the distribution of the estimate y.rep from a hypothetical replicated study with effect size A and standard error s. There are three key summaries:

- The *power*: the probability that the replication y.rep is statistically significantly different from zero.
- The *Type S error rate*: the probability that the estimate has the correct sign, if it is statistically significantly different from zero.
- The *exaggeration factor*: the expected ratio of the estimate divided by the effect size, conditional on the estimate being statistically significantly different from zero.

We work out expressions for all three summaries under the assumption of normal distributions. It would not be difficult to compute these for other models using simulations.

The power is $\Pr(|y.rep/s|>1.96) = (1 - \Phi(1.96 - A/s)) + \Phi(-1.96 - A/s)$, where $\Phi$ is the normal cumulative distribution function.

The Type S error rate is the ratio of the first term in the above expression for power, divided by the sum of the two terms; that is, $(1 - \Phi(1.96 - A/s))/[(1 - \Phi(1.96 - A/s)) + \Phi(-1.96 - A/s)]$.

The exaggeration factor is …

[3] The threshold for the z-ratio is not necessarily 2. It could be higher because of a t-distribution correction for degrees of freedom, a more stringent significance level, or a multiple comparisons correction. For our present purposes, however, the exact level of the statistical-significance cutoff is not important; all that matters is that some threshold is being used.

We have implemented these features in an function, `retropower()`, which is included in the `arm` (applied regression modeling) package in R. The arguments to the function are y (the estimate from the data), s (the standard error of the estimate), z (the statistical significance threshold, assumed above to be 1.96), and A (the hypothesized effect size). The function returns a list with three items: the power, the type S error rate, and the exaggeration factor.

**Simple numerical examples**

We first explore the mechanics of retrospective power analysis with some simple hypothetical examples of experimental data:

-

**Applied example: Beauty and sex ratios**

We demonstrate our recommended approach with an example from Gelman and Weakliem (2009). The story begins with a finding by Kanazawa (2007) from a sample of 2972 respondents from the National Longitudinal Study of Adolescent Health. Each of these people had been assigned an "attractiveness" rating on a 1-5 scale and then, years later, had at least one child. 56% of the first-born children of the parents in the highest attractiveness category were girls, compared to 48% in the other groups. This observed difference of 8 percentage points has a standard error of 3.5 percentage points (based on the binomial model, which we agree is appropriate for these data) and is statistically significant at the conventional 5% level, with a p-value of 0.015.

This p-value has been questioned for reasons of multiple comparisons (Gelman, 2007). Instead of comparing attractiveness category 5 to categories 1,2,3,4, the researcher also had the equally reasonable options of comparing 4,5 to 1,2,3, or comparing 3,4,5 to 1,2, or comparing 2,3,4,5 to 1, or comparing 5 to 1, or comparing 4 and 5 to 1 and 2, or simply fitting a linear regression. It turns out that *none* of these other potential analyses is statistically significant: thus, the p-value of 0.015 represents the winner among at least seven possible comparisons. We would judge the most reasonable summary of the attractiveness/sex-ratio pattern in these data to be the linear regression, which (when the predictor is standardized) yields an estimate of 0.047 (that is, 4.7 percentage points) with standard error 0.043, as illustrated in Figure 1.

For the purpose of illustration, however, we will stick with the original estimate of 8 percentage points with p-value 0.015. Kanazawa (2007) followed the usual practice and just stopped right here. After all, there's no need for a power analysis if you already have statistical significance, right? The sources quoted at the start of this article all assume that retrospective power analysis, if it is done at all, would be performed in response to a non-significant result.

But let's break the rules and do a power analysis anyway.

We need to postulate an effect size, which will *not* be 8 percentage points, or even the 4.7 percentage points that came from the regression. Instead, we form our assumptions the way that

would be done in a prospective power calculation, using the scientific literature. From Gelman and Weakliem (2009):

> There is a large literature on variation in the sex ratio of human births, and the effects that have been found have been on the order of 1 percentage point (for example, the probability of a girl birth shifting from 48.5 percent to 49.5 percent). Variation attributable to factors such as race, parental age, birth order, maternal weight, partnership status and season of birth is estimated at from less than 0.3 percentage points to about 2 percentage points, with larger changes (as high as 3 percentage points) arising under economic conditions of poverty and famine. That extreme deprivation increases the proportion of girl births is no surprise, given reliable findings that male fetuses (and also male babies and adults) are more likely than females to die under adverse conditions.

Given the generally small observed differences in sex ratios as well as the noisiness of the subjective attractiveness rating used in this particular study, we expect any population differences in the probability of girl birth to be well under 1 percentage point.

It is standard for prospective power analyses to be performed under a range of assumptions, and we shall do the same here, hypothesizing effect sizes of 0.1, 0.3, and 1.0 percentage points. Under each hypothesis, we consider what might happen in a study with sample size equal to that of Kanazawa (2007).

Again, we will ignore multiple comparisons issues and take his claim of statistical significance at face value: from the reported estimate of 8% and t-statistic of 2.44, we can deduce that the standard error of the difference is 3.28%. Such a result will be statistically significant only if the estimate is at least 1.96 standard errors from zero; that is, the estimated difference in proportion girls, comparing beautiful parents to others, would have to be more than 6.43 percentage points or less than -6.43.

*True difference of 0.1%.* Suppose the probability of girl births in the population is 0.1 percentage points higher among attractive than among unattractive parents. Then an unbiased estimate with standard error 3.28 percentage points will have a 2.68% chance of being statistically significantly positive—and a 2.32% chance of being statistically significantly negative. In either case, the estimated effect, of at least 6.43 percentage points, will be over 60 times higher than the true effect, and with a 46% chance of going in the wrong direction. If the result is not statistically significant, the chance of the estimate being the wrong sign (sometimes called a Type S error; see Gelman and Tuerlinckx, 2000) is 49% percent, so that the direction of the estimate would provide almost no information on the sign of the true effect.

*True difference of 0.3%.* Now suppose the probability of girl births in the population is 0.3 percentage points higher among attractive than among unattractive parents. Based on the same analysis based on the normal distribution, there is a 3.1% chance the result will be statistically significantly positive and a 2.0% chance of being statistically significantly negative. Thus, even a statistically significant result has roughly a 40% chance of being in the wrong direction. In addition, any statistically significant finding necessarily overestimates the magnitude of the true

effect (in this scenario) by more than a factor of 20. If the result is not statistically significant, there would be a 46% chance that the point estimate is in the opposite direction of the true effect.

*True difference of 1.0%.* A similar calculation shows that, under a true difference of 1% (larger than could be realistically expected given the literature on sex ratios), there would be a 4.9% chance of the result being statistically significantly positive and a 1.1% chance of a statistically significantly negative result. A statistically significant finding in this case has a 19% chance of appearing with the wrong sign (and, in addition, would overestimate the magnitude of the true effect by more than a factor of 6).

Our retrospective power analysis has shown that, *even if* the true difference were as large as 1 percentage point, and *even if* there were no multiple comparisons problems, that the sample size of this study is such that a statistically significant result has a one-in-five chance of having the wrong sign and, in any case, would overestimate the magnitude of the effect by more than a factor of six.

We have learned something from the retrospective power analysis, beyond what was revealed by the estimate, confidence interval, and p-value that came from the original data summary.

**Comparing our procedure to usual practice**

The method illustrated above differs from the usual retrospective power analysis in two ways:

1. We hypothesize an effect size using prior information on possible effects. Compare this to the sometimes-recommended strategy of considering a minimal effect size deemed to be substantively important. Both these approaches use substantive knowledge but in different ways. For example, in the beauty-and-sex-ratio example, our best estimate from the literature is that any true differences are less than 0.3 percentage points in absolute value. Whether this is a substantively important difference is another question entirely. Conversely, suppose that a difference in this context were judged to be substantively important if it were at least 5 percentage points. We have no interest in computing power under this assumption, since our literature review suggests it is extremely implausible.

2. We recommend a retrospective power analysis whether or not the data result is statistically significant. This is different than the usual procedure of accepting $p<0.05$ results at face value and only performing power analysis to understand the limitations of non-significant outcomes. When studies are noisy, as in the beauty and sex ratio example, a power analysis can reveal the fragility even of statistically-significant findings.

**The statistical significance filter: overestimating the magnitude of small effects**

As the above example illustrates, underpowered studies suffer from three deficiencies.

First, and most obviously, a study with low power is unlikely to "succeed" in the sense of yielding a statistically significant result.

Second, any statistically significant findings are likely to be in the wrong direction. It is quite possible for a result to be significant at the 5% level—with a 95% confidence interval that is entirely excludes zero—and for there to be a chance of 40% or more that this interval is on *the wrong side* of zero. Even sophisticated users of statistics can be unaware of this point, that the probability of a type S error is not the same as the p-value or significance level.[4]

The third problem with underpowered studies is that when they do happen to result in statistical significance, they can drastically overestimate the magnitudes of any effects. This filtering effect of statistical significance may very well contribute to the decreasing trends that have been observed in the magnitudes of reported medical research findings (Lehrer, 2010). We hope that retrospective power analysis can give a clue as to how much this might be a problem in any particular case.

**Power analysis and costs**

Traditionally, one performs a power analysis *before* embarking on a study that will be expensive in money, effort, or the well-being of research participants. For example, the National Institutes of Health will typically only fund a study if its sample size is deemed large enough based on a reasonable power calculation.

But when analyzing existing data (as in the beauty and sex-ratio analysis) or when costs are low, researchers are not restricted by power considerations. And when sample sizes are large, statistically significant results can greatly overestimate the magnitude of effects.

Thus, retrospective power calculations may be particularly relevant for observational analyses of existing datasets, where there was no initial power calculation required to justify the study in the first place.

**Discussion**

Power calculations and null hypotheses are among the few places in which prior incorporation can be used in classical inference. From this perspective, it is unremarkable that we recommend retrospective power analysis as a way of supplementing classical confidence intervals and hypothesis tests when samples are small. Any statistical method is sensitive to its assumptions, and so one must carefully examine the prior information that goes into these analyses, just as one must scrutinize the assumptions that go into any estimator or likelihood function.

[4] For example, in a paper attempting to clarify p-values for a clinical audience, Froehlich (1999) describes a problem in which the data have a one-sided tail probability of 0.46 (compared to a specified threshold for a minimum worthwhile effect) and incorrectly writes: "In other words, there is a 46% chance that the true effect" exceeds the threshold. The mistake here is to treat a sampling distribution as a Bayesian posterior distribution—and is particularly likely to cause a problem in settings with small effects and small sample sizes, when statistically significant results can be dramatic overestimates.

The present article has focused on power analyses with assumptions determined by a literature review. In other settings, postulated effect sizes could be set using auxiliary data, meta-analysis, or a hierarchical model. It should also be possible to perform retrospective power analysis on secondary data analyses: for example, if subgroup analysis is performed in a randomized trial, one can apply a power analysis based on an external estimate of the magnitude of interactions.

Power analysis requires an assumed effect size and adds nothing to a data analysis if the postulated effect size is estimated from the very same data. This is the point made in the quotations that lead off this article. But when power analysis is recognized as a generally-accepted way of adding prior information, we see an opening for a big step forward in studies with small effects and small sample sizes. The relevant question is not, "What is the power of a test?" but rather, "What might be expected to happen in studies of this size?" Also, contrary to the common impression, retrospective power analysis is appropriate for statistically-significant as well as non-significant findings—as long as the power analysis uses real external information.

**References**


Cox, D. R. (1958). *Planning of Experiments.* New York: Wiley.

Froehlich, G. W. (1999). What is the chance that this study is clinically significant? A proposal for Q values. *Effective Clinical Practice* 2, 234-239.

Gelman, A. (2007). Letter to the editor regarding some papers of Dr. Satoshi Kanazawa. *Journal of Theoretical Biology* 245, 597-599.

Gelman, A., and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15, 373-390.

Gelman, A., and Weakliem, D. (2009). Of beauty, sex, and power: statistical challenges in the estimation of small effects. *American Scientist* 97, 310-316.

Goodman, S. N., and Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine* 121, 200-206.

Hoenig, J. M., and Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *American Statistician* 55, 1-6.

Ioannides, J.

Kanazawa, S. (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. *Journal of Theoretical Biology* 244, 133-140.

Lehrer, J. (2010). The truth wears off. *New Yorker* 13 Dec, 52-57.

Lenth, R. V. (2007). Statistical power calculations. *Journal of Animal Science* 85, E24-E29.

Leventhal, I. (2009). Statistical power calculations: comment. *Journal of Animal Science* 87, 1854-1855.

Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *British Medical Journal* 325, 1304.

Stearne and Davies-Smith (2001) BMJ.

Vul, E., Harris, C., Winkelman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition (with discussion). *Persepctives on Psychological Science* 4, 274-290.

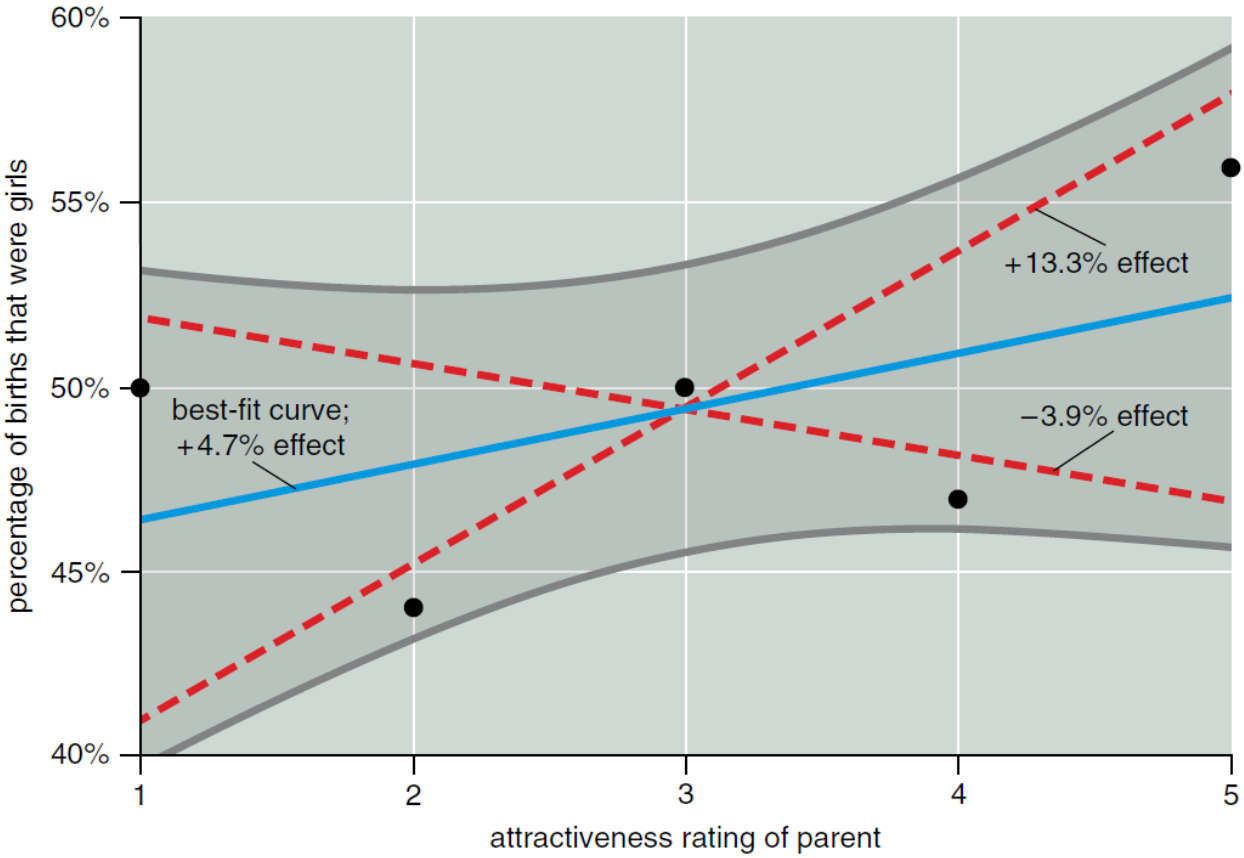


Figure 1. Using a sample of 2,972 respondents from the National Longitudinal Study of Adolescent Health, each of whom had been rated on a five-point scale of attractiveness, Kanazawa (2007) compared the most attractive group to the average of the other four groups and concluded that there was a higher probability that the first-born child of the most attractive respondents would be a girl. The difference is 8 percentage points with p-value 0.015. This is, however, only one of the many comparisons that could be performed on these data. A regression analysis gives an estimated slope (after standardizing the predictor) of 4.7 percentage points with a standard error of 4.3 percentage points. From Gelman and Weakliem (2009).