The difference between “significant” and “not significant” is not itself statistically significant*

Andrew Gelman†

November 29, 2005

Abstract

The second problem is that changes in statistical significance are not themselves significant. By this, we are not merely making the commonplace observation that any particular threshold is arbitrary—for example, only a small change is required to move an estimate from a 5.1% significance level to 4.9%, thus moving it into statistical significance. Rather, we are pointing out that even large changes in significance levels can correspond to small, non-significant changes in the underlying variables. We illustrate with a theoretical and an applied example.

Keywords: multilevel modeling, multiple comparisons, statistical significance

1 Introduction

A common statistical error is to summarize comparisons by statistical significance and to draw a sharp distinction between significant and non-significant results. The approach of summarizing by statistical significance has two pitfalls, one that is obvious and one that is less well known.

First, statistical significance does not equal practical significance. For example, if the estimated effect of a drug was to decrease blood pressure by 0.10 with a standard error of 0.03, this would be statistically significant but probably not practically significant. Conversely, an estimatd effect of 10 with a standard error of 10 would not be statistically significant, but it has the possibility of being practically significant.

The second problem, which is the subject of this article, is that changes in statistical significance are not themselves significant. By this, we are not merely making the commonplace observation that any particular threshold is arbitrary—for example, only a small change is required to move an estimate from a 5.1% significance level to 4.9%, thus moving

*We thank the National Science Foundation for financial support.

†Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, www.stat.columbia.edu/∼gelman

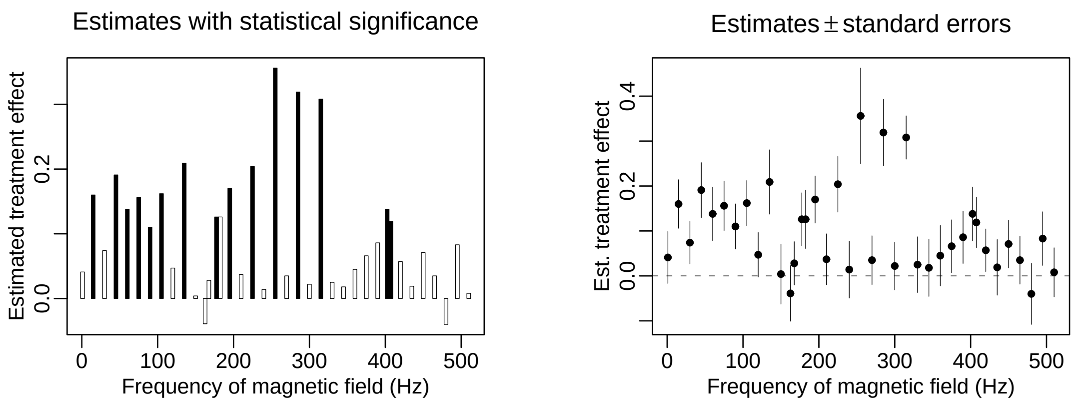

Figure 1: (a) Estimated effects of electromagnetic fields on calcium efflux from chick brains, shaded to indicate different levels of statistical significance, adapted from Blackman et al. (1988). A separate experiment was performed at each frequency. (b) Same results presented as estimates ± standard errors. As discussed in the text, the first plot, with its emphasis on statistical significance, is misleading.

it into statistical significance. Rather, we are pointing out that even large changes in significance levels can correspond to small, non-significant changes in the underlying variables. We illustrate with two examples.

2 Theoretical example

Consider two independent studies with effect estimates and standard errors of 25 ± 10 and 10 ± 10. The first study is statistically significant at the 1% level, and the second is not at all significant at one standard error away from 0. Thus it would be tempting to conclude that there is a large difference between the two studies. In fact, however, the difference is not even close to being statistically significant: the estimated difference is 15, with a standard error of $\sqrt{10^2 + 10^2} = 14$.

3 Applied example

In the wake of concerns about the health effects of low-frequency electric and magnetic fields, Blackman et al. (1988) performed a series of experiments to measure the effect of electromagnetic fields at various frequencies on the functioning of chick brains. At each of several frequencies of electromagnetic fields (1 Hz, 15 Hz, 30 Hz, ..., 510 Hz), a randomized experiment was performed to estimate the effect of exposure, compared to a control condition of no electromagnetic field. The estimated treatment effect (the average difference

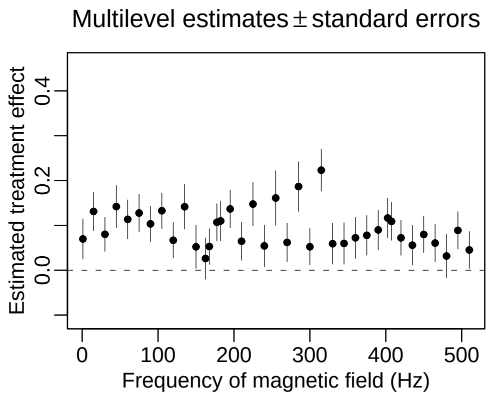

Figure 2: Multilevel estimates and standard errors for the effects of magnetic fields, partially pooled from the separate estimates displayed in Figure 1. The standard errors of the original estimates were large, and so the multilevel estimates are pooled strongly toward the common average which is near of 0.1.

between treatment and control measurements) and the standard error at each frequency were reported.

Blackman et al. (1988) summarized the estimates at the different frequencies by their statistical significance, using a graph similar to Figure 1a with different shading indicating results that are more than 2.3 standard errors from zero (that is, statistically significant at the 99% level), between 2.0 and 2.3 standard errors from zero (statistically significant at the 95% level), and so forth. The researchers used this sort of display to hypothesize that one process was occurring at 255, 285, and 315 Hz (where effects were highly significant), another at 135 and 225 Hz (where effects were only moderately significant), and so forth. The estimates are all of relative calcium efflux, so that an effect of 0.1, for example, corresponds to a 10% increase compared to the control condition.

The researchers in the chick-brain experiment made the common mistake of using statistical significance as a criterion for separating the estimates of different effects, an approach does not make sense. At the very least, it is more informative to show the estimated treatment effect and standard error at each frequency, as in Figure 1b.

One way to handle the multiple-comparisons aspect of this problem is to fit a multilevel model of the sort used in meta-analysis. If at each frequency j, we label the estimated effect and standard error as y_j and σ_j, then the simplest multilevel model is $y_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$, $\theta_j \sim \mathrm{N}(\mu, \tau^2)$, and the resulting Bayesian estimates for the effects θ_j are partially pooled toward the average of all the data (see, for example, Gelman et al., 2003, chapter 5). The posterior estimates and standard errors are shown in Figure 2.

The multilevel model can be seen as a way to estimate the effects at each frequency j, without setting "non-significant" results to zero. Some of the apparently dramatic features of the original data as plotted in Figure 1a—for example, the negative estimate at 480 Hz and the pair of statistically-significant estimates at 405 Hz—do not stand out so much in the multilevel estimates, indicating that these features could be easily explained by sampling variability and do not necessarily represent real features of the underlying parameters.

4 Discussion

It is standard in applied statistics to evaluate inferences based on their statistical significance at the 5% level. There has been a move in recent years toward reporting confidence intervals rather than p-values, and the centrality of hypothesis testing has been challenged (see Krantz, 1999, for a review of these issues) but even when using confidence intervals it is natural to check whether they include zero. Statistical significance, in some form, is a way for us to assess the reliability of statistical findings. However, as we have seen, comparisons of the sort, "X is statistically significant but Y is not" can be misleading.

References

Blackman, C. F., Benane, S. G., Elliott, D. J., House, D. E., and Pollock, M. M. (1988). Influence of electromagnetic fields on the efflux of calcium ions from brain tissue in vitro: a three-model analysis consistent with the frequency response up to 510 Hz. *Bioelectromagnetics* **9**, 215–227.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association* **94**, 1372–1381.