The difference between "significant" and "not significant" is not itself statistically significant*

Andrew Gelman† Hal Stern‡

August 22, 2006

Abstract

It is common to summarize statistical comparisons by declarations of statistical significance or non-significance. Here we discuss one problem with such declarations, namely that changes in statistical significance are often not themselves statistically 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.

The error we describe is conceptually different from other oft-cited problems—that statistical significance is not the same as practical importance, that dichotomization into significant and non-significant results encourages the dismissal of observed differences in favor of the usually less interesting null hypothesis of no difference, and that any particular threshold for declaring significance is arbitrary. We are troubled by all of these concerns and do not intend to minimize their importance. Rather, our goal is to bring attention to what we have found is an important but much less discussed point. We illustrate with a theoretical example and two applied examples.

Keywords: multilevel modeling, multiple comparisons, replication, statistical significance

1 Introduction

A common statistical error is to summarize comparisons by statistical significance and then draw a sharp distinction between significant and non-significant results. The approach of summarizing by statistical significance has a number of pitfalls, most of which are covered in standard statistics courses but one that we believe is less well known. We refer to the fact that changes in statistical significance are not themselves significant. By this, we are not

*We thank Howard Wainer, Peter Westfall, and an anonymous reviewer for helpful comments and 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

‡Department of Statistics, University of California, Irvine, sternh@uci.edu, www.ics.uci.edu/∼sternh

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 shall illustrate with three examples.

This article does not attempt to provide a comprehensive discussion of significance testing. There are several such discussions; see, for example, Krantz (1999). Indeed many of the pitfalls of relying on declarations of statistical significance appear to be well known. For example, by now practically all introductory texts point out that statistical significance does not equal practical importance. If the estimated effect of a drug is to decrease blood pressure by 0.10 with a standard error of 0.03, this would be statistically significant but probably not important in practice (or so we suppose, given our general knowledge that blood pressure values are typically around 100). Conversely, an estimated effect of 10 with a standard error of 10 would not be statistically significant, but it has the possibility of being important in practice. As well, introductory courses regularly warn students about the perils of strict adherence to a particular threshold (the point mentioned above regarding 5.1% and 4.9% significance levels). Similarly most statisticians and many practitioners are familiar with the notion that automatic use of a binary significant/non-significant decision rule encourages practitioners to ignore potentially important observed differences in favor of the usually less interesting null hypothesis. Thus, from this point forward we focus only on the less widely known but equally important error of comparing two or more results by comparing their degree of statistical significance.

2 Theoretical example: comparing the results of two experiments

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 statistically significant, being only 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 $\sqrt{10^2 + 10^2} = 14$.

Additional problems arise when comparing estimates with different levels of information. Suppose in our example that there is a third independent study with much larger sample size that yields an effect estimate of 2.5 with standard error of 1.0. This third study attains

the same significance level as the first study, yet the difference between the two is itself also significant. Both find a positive effect but with much different magnitudes. Does the third study replicate the first study? If we restrict attention only to judgments of significance we might say yes, but if we think about the effect being estimated we would say no, as noted by Utts (1991). In fact, the third study finds an effect size much closer to that of the second study, but now because of the sample size it attains significance.

Declarations of statistical significance are often associated with decision making. For example, if the two estimates in the first paragraph concerned efficacy of blood pressure drugs, then one might conclude that the first drug works and the second does not, making the choice between them obvious. But is this obvious conclusion reasonable? The two drugs do not appear to be significantly different from each other. One way of interpreting lack of statistical significance is that further information might change one's decision recommendations. Our key point is not that we object to looking at statistical significance but that comparing statistical significance levels is a bad idea. In making a comparison between two treatments, one should look at the statistical significance of the difference rather than the difference between their significance levels.

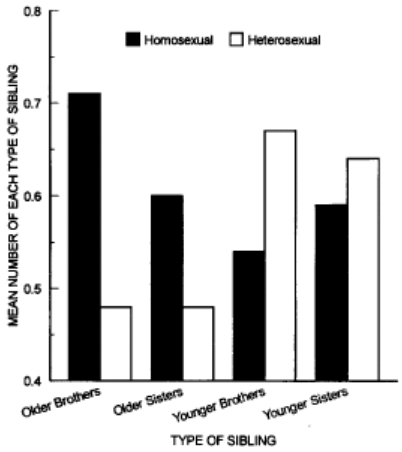
3 Applied example: homosexuality and the number of older brothers and sisters

The paper, "Biological versus nonbiological older brothers and men's sexual orientation," (Bogaert, 2006), appeared recently in the Proceedings of the National Academy of Sciences and was picked up by several leading science news organizations (Bower, 2006, Motluk, 2006, Staedter, 2006). As the article in *Science News* put it:

> The number of biological older brothers correlated with the likelihood of a man being homosexual, regardless of the amount of time spent with those siblings during childhood, Bogaert says. No other sibling characteristic, such as number of older sisters, displayed a link to male sexual orientation.

We were curious about this—why older brothers and not older sisters? The article referred back to Blanchard and Bogaert (1996), which had the graph and table shown in Figure 1, along with the following summary:

> Significant beta coefficients differ statistically from zero and, when positive, indicate a greater probability of homosexuality. Only the number of biological



Predictor	β	SE	Wald statistic	p	e^{β}
Initial equation					
Number of older brothers	0.29	0.11	7.26	0.007	1.33
Number of older sisters	0.08	0.10	0.63	0.43	1.08
Number of younger brothers	−0.14	0.10	2.14	0.14	0.87
Number of younger sisters	−0.02	0.10	0.05	0.82	0.98
Father's age at time of proband's birth	0.02	0.02	1.06	0.30	1.02
Mother's age at time of proband's birth	-0.03	0.02	1.83	0.18	0.97
Final equation—number of older brothers	0.28	0.10	8.77	0.003	1.33

Figure 1: From Blanchard and Bogaert (1996): (a) mean numbers of older and younger brothers and sisters for 302 homosexual men and 302 matched heterosexual men, (b) logistic regression of sexual orientation on family variables from these data. The graph and table illustrate that, in these data, homosexuality is more strongly associated with number of older brothers than with number of older sisters. However, no evidence is presented that would indicate that this difference is statistically significant.

> older brothers reared with the participant, and not any other sibling characteristic including the number of nonbiological brothers reared with the participant, was significantly related to sexual orientation.

The conclusions appear to be based on a comparison of significance (for the coefficient of the number of older brothers) with nonsignificance (for the other coefficients), even though the differences between the coefficients do not appear to be statistically significant. One cannot quite be sure—it is a regression analysis and the different coefficient estimates are not independent—but based on the picture we strongly doubt that the difference between the coefficient of the number of older brothers and the coefficient of the number of older sisters is significant.

Is it appropriate to criticize an analysis of this type? After all, the data are consistent with the hypothesis that only the number of older brothers matters. But the data are also consistent with the hypothesis that only the birth order (the total number of older siblings) matters. (Again we cannot be certain but we strongly suspect so from the graph and the table.) Given that the 95% confidence level is standard (and we are pretty sure the paper would not have been published had the results not been statistically significant at that level), it is appropriate that the rule should be applied consistently to hypotheses consistent with the data. We are speaking here not as experts in biology but rather as statisticians: the published article and its media reception suggest unquestioning acceptance of a result (only the number of older brothers matters) which, if properly expressed as a comparison, would be better described as "suggestive." For example, the authors could have written that the sexual preference of the men in the sample is statistically significantly related to birth order and, in addition, more strongly related to number of older brothers than number of older sisters, but with the latter difference not being statistically significant.

4 Applied example: health effects of low-frequency electromagnetic fields

The issue of comparisons between significance and non-significance is of even more concern in the increasingly common setting where there are a large number of comparisons. We illustrate with an example of a laboratory study with public health applications.

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

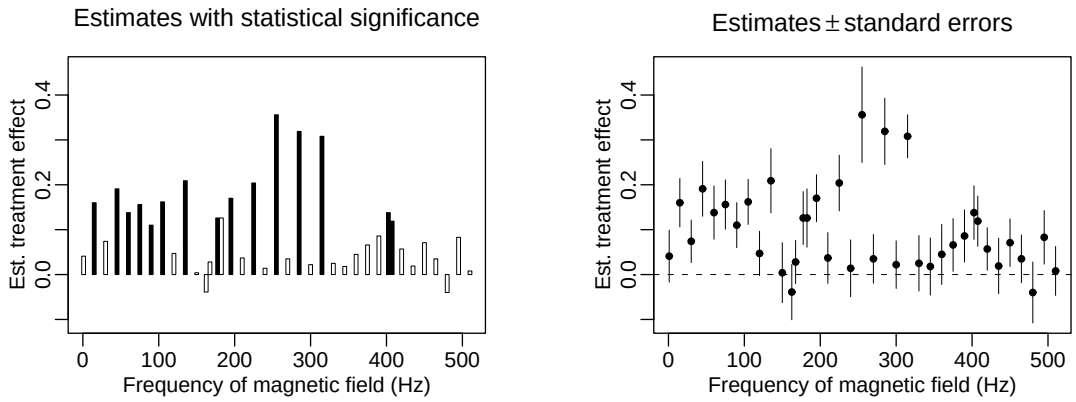


Figure 2: (a) Estimated effects of electromagnetic fields on calcium efflux from chick brains, shaded to indicate different levels of statistical significance, adapted from Blackman et al. (1988). A separate experiment was performed at each frequency. (b) Same results presented as estimates ± standard errors.
As discussed in the text, the first plot, with its emphasis on statistical significance, is misleading.

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 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 2a 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 that 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 2b. This display makes the key features of the data clear. Though the size of the effect varies, it is just about

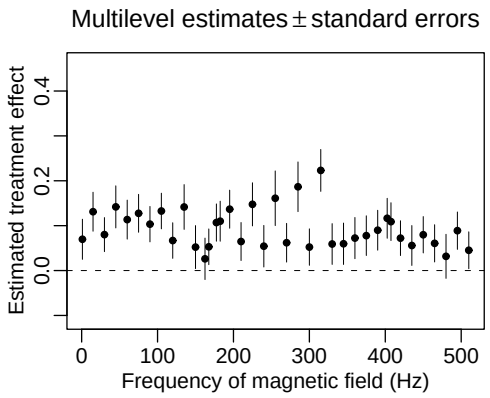


Figure 3: Multilevel estimates and standard errors for the effects of magnetic fields, partially pooled from the separate estimates displayed in Figure 2. The standard errors of the original estimates were large, and so the multilevel estimates are pooled strongly toward the common average which is near of 0.1.

always positive and typically not far from 0.1.

What should one do instead? One natural idea is to fit a model in which the effect is a smooth function of frequency. The data, however, appear to jump around quite a bit and we do not have the scientific expertise to justify the appropriateness of such a model. Another way to handle the large number of related experiments in a single data analysis 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 (using a flat prior distribution on μ and τ) are partially pooled toward the average of all the data (see, for example, Gelman et al., 2003, chapter 5, for more discussion of models of this type). The posterior estimates and standard errors are shown in Figure 3. Some might object to this exchangeable model for the experiments at different frequencies, but this is consistent with the original Blackman et al. analysis, which also makes this assumption of exchangeability. (That is to say, neither of the analyses distinguish between the experiments based on the underlying frequency.) Our simple hierarchical model is not intended to be definitive, merely a model that we believe improves upon the separate judgments of statistical significance for each experiment. A subject-matter expert can perhaps use Figure 3 (rather than Figure 2a) to formulate further hypotheses and models.

The multilevel analysis can be seen as a way to estimate the effects at each frequency j, without setting apparently "non-significant" results to zero. Some of the most dramatic

features of the original data as plotted in Figure 2a—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 explained by sampling variability and do not necessarily represent real features of the underlying parameters.

5 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, but even when using confidence intervals it is natural to check whether they include zero. Statistical significance, in some form, is a way 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.

Blanchard, R., and Bogaert, A. F. (1996). Homosexuality in men and number of older brothers. *American Journal of Psychaitry* **153**, 27–31.

Bogaert, A. F. (2006). Biological versus nonbiological older brothers and men's sexual orientation. *Proceedings of the National Academy of Sciences* **103**, 10771–10774.

Bower, B. (2006). Gay males' sibling link: men's homosexuality tied to having older brothers. *Science News* **170** (1), 3.

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.

Motluk, A. (2006). Male sexuality may be decided in the womb. *New Scientist*, online edition, 26 June.

Staedter, T. (2006). Having older brothers increases a man's odds of being gay. *Scientific American*, online edition, 27 June.

Utts, J. M. (1991). Replication and meta-analysis in parapsychology (with discussion). *Statistical Science* **6**, 363–403.