A reason why not to ban Null Hypothesis Significance Tests

Bruno Lecoutre[1], Jacques Poitevineau[2], Marie-Paule Lecoutre[3]

[1] ERIS, Laboratoire de Mathématiques Raphaël Salem
UMR 6085 C.N.R.S. et Université de Rouen
Avenue de l'Université, BP 12, 76801 Saint-Etienne-du-Rouvray
bruno.lecoutre@univ-rouen.fr
Internet : http ://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris
[2] ERIS, LAM/LCPE
UMR 7604, C.N.R.S., Université de Paris 6 et Ministère de la Culture
11 rue de Lourmel, 75015 Paris
poitevin@ccr.jussieu.fr
[3] ERIS, Laboratoire Psy.Co, E.A. 1780,
Université de Rouen
UFR Psychologie, Sociologie, Sciences de l'Education
76821 Mont-Saint-Aignan Cedex
marie-paule.lecoutre@univ-rouen.fr

Abstract

It is shown that an interval estimate for a contrast between means can be straightforwardly computed, given only the observed contrast and the associated t or F test statistic (or equivalently the corresponding p-value). This interval can be seen as a *frequentist* confidence interval, as a standard *Bayesian* credibility interval, or as a *fiducial interval*.

This gives Null Hypothesis Significance Tests (NHST) users the possibility of an easy transition towards more appropriate statistical practices. Conceptual links between NHST and interval estimates are outlined.

Introduction

Many recent papers have stressed on the necessity of changes in reporting experimental results. A more and more widespread opinion is that inferential procedures that provide genuine information about the size of effects must be used in addition or in place of Null Hypothesis Significance Tests (NHST). So, in psychology, this has been recently made official by the American Psychological Association Task Force on Statistical Inference. The Task Force has proposed guidelines for revising the statistical section of the American Psychological Association Manual. Following these guidelines, "interval estimates should be given for any effect sizes involving principal outcomes" (Wilkinson *et al.*, 1999).

Therefore a salutary project should be to equip NSHT users with tools that should facilitate a smooth transition towards interval estimates. In this perspective a surprisingly simple and virtually ignored result is the easiness to get an interval estimate for a difference between two means (and more generally for a contrast between means) from the associated t or F test.

Such an interval estimate can receive different justifications and interpretations. It can be seen as a *frequentist* confidence interval as well as a standard *Bayesian* credibility interval, or again as a Fisher's *fiducial* interval (Fisher, 1990). Theoretical discussions about these frameworks are outside the scope of this paper. Let us mention however that the authors' opinion is that a Bayesian approach with a fiducial flavor is ideally suited for experimental data analysis and scientific reporting. The interested reader can be refered to Lecoutre *et al.* (2001) and Rouanet *et al.* (2000). Here we shall use the expression "interval estimate", leaving the reader free to choose between the justification and interpretation frameworks.

From *F* Ratios to interval estimates for contrasts between means

As an illustration consider an experiment involving two crossed factors *Age* and *Treatment*, each with two modalities. The means of the four experimental conditions (with 10 subjects in each) are respectively 5.77 (a1,t1), 5.25 (a2,t1), 4.83 (a1,t2) and 4.71 (a2,t2).

The following typical comments based on ANOVA F tests are found in an experimental review :

> "the only significant effect is a main effect of treatment (F[1,36]=6.39, p=0.016), reflecting a substantial improvement"

and again

> "clearly, there is no evidence ($F_{[1,36]} = 0.47$, $p = 0.50$) of an interaction".

Such comments are commonly found in experimental publications. It is strongly suggested to a reader, little informed of the rhetoric going with the use of NHST, that it has been demonstrated both a large main effect of treatment and a small interaction effect. But there is nothing of the kind !

The difference between the two observed treatment means is

$$d = \frac{1}{2}(5.77 + 5.25) - \frac{1}{2}(4.83 + 4.71) = +0.74$$

and the interaction effect can be characterized by the difference of differences

$$d = (5.77 - 4.83) - (5.25 - 4.71) = +0.40$$

A simple and general result is that the $100(1-\alpha)\%$ interval estimate for the true effect δ can be deduced from the F ratio (with one and q degrees of freedom). It is (assuming $d \neq 0$) :

$$\left[d - (|d|/\sqrt{F})t_{q;\frac{\alpha}{2}}\ ,\ d + (|d|/\sqrt{F})t_{q;\frac{\alpha}{2}}\right]$$

where $t_{q;\frac{\alpha}{2}}$ is the $(\frac{\alpha}{2})\%$ upper point of the standard Student's distribution with q degrees of freedom (recall that the square of $t_{q;\frac{\alpha}{2}}$ is the $\alpha\%$ upper point of the F distribution with one and q degrees of freedom). Moreover a good approximation can be straightforwardly obtained (i.e. without referring to statistical tables) by replacing $t_{q;\frac{\alpha}{2}}$ with $1.96\sqrt{q/(q-2)}$, or again more simply with 2 when q is large.

This result brings to the fore the fundamental property of the F test statistic of being an estimate of the experimental accuracy, *conditionally on the observed value* d. More explicitly d^2/F estimates the sampling error variance of d. The same result applies to usual Student's t tests, replacing $|d|/\sqrt{F}$ with d/t.

From $t_{36;0.025} = 2.028$, we get here the 95% interval estimates [+0.15 , +1.33] for the difference between the two treatments and [-0.78 , +1.58] for the interaction effect. This clearly shows that it cannot be concluded both to a substantive difference between treatment means and to a small, or at least relatively negligible, interaction effect.

Significance levels and interval estimates

Since the t or F test statistics can be computed from the p-value (assumed known with a sufficient degree of accuracy), interval estimates can be deduced directly from the p-value. It follows that, *given the observed* d, the p-value is also an estimate of the experimental accuracy. Hence, intuitively, the more significant the result (the more p less than α), the more δ should be close to d. It is enlightening to remark that the $100(1-\alpha)\%$ interval estimate can again be written as $[d-d_\alpha\ ,\ d+d_\alpha]$, where $d_\alpha = (|d|/\sqrt{F})t_{q;\frac{\alpha}{2}}$ is the (positive) critical value of d such that the test is declared significant at two-sided level α if $|d|$ exceeds d_α.

As an other illustration consider a study designed to test the efficacy of a new drug by comparing two groups (new drug *vs.* placebo) of 20 subjects each. The new drug is considered as efficient (clinically interesting) by experts in the field if the raw difference is more than +2. Four possible cases of results are constructed by crossing the outcome of the t test (significant, $p = 0.001$ *vs.* nonsignificant, $p = 0.60$, two-sided) and the observed difference between the two means d (large, $d = +4.92$ *vs.* small, $d = +0.84$).

The corresponding 95% interval estimates for the true difference δ are given in Table 1. This table illustrates the shift from the knowledge of d and p (or t) to a conclusion about the magnitude of δ (the efficacy of the new drug). From this table, it becomes easy to avoid erroneous conclusions based on hasty interpretations of NHST. The following general rules can be deduced.

Table 1 - 95% interval estimates for δ in the four cases of results
($t_{38;0.025} = 2.024$, hence $d_{0.05} = 2.024d/t$)

case	t	p	d	$d_{0.05}$	95% *interval estimate*	*conclusion*
1	+3.566	0.001	+4.92	2.79	[+2.13 , +7.71]	efficient
2	+3.566	0.001	+0.84	0.48	[+0.36 , +1.32]	inefficient
3	+0.529	0.60	+4.92	18.83	[-13.91 , +23.75]	no firm conclusion
4	+0.529	0.60	+0.84	3.22	[-2.38 , +4.06]	no firm conclusion

Case 1 (significant test, large positive d**)**. Such results seem generally very favorable to NHST users. This is justified here, since $d - d_{0.05}$ is greater than +2. However it must be stressed that asserting a large difference needs some cautious. The test must be "sufficiently significant", i.e. p sufficiently smaller than α, to imply a large $d - d_\alpha$ value. Indeed, in the limiting case where d is positive and the test is just significant at two-sided level α, it can only be concluded that δ is positive.

Case 2 (significant test, small positive d). Since $0 < d_\alpha < d$, these conditions imply that d_α and $d - d_\alpha$ are small. Moreover, in the present results, $d + d_{0.05}$ is also small (less than +2), so that a small difference can be asserted. Since there is apparently conflict between the small observed difference and the statistically significant outcome, this case generally appears to NHST users to be cumbersome. There is no paradox however, since this can only occur when the experimental accuracy is "very good" (i.e. when the sampling error variance is a small). Therefore it is in fact a privileged case. But, as a consequence the test is very powerful (d_α small), so that even a small observed difference can be statistically significant.

Case 3 (nonsignificant test, large positive d). As a general rule no firm inductive conclusion can be reached : it is obviously out of the question that a small difference can be asserted. Actually, these results indicate an insufficient experimental accuracy and therefore are not really contradictory (only a very large observed difference should be statistically significant). However, many NHST users feel this case cumbersome because they cannot generalize the descriptive conclusion of a large difference.

Case 4 (nonsignificant test, small positive d). These conditions only imply that d is less than d_α. But they can correspond to small as well as large $d - d_\alpha$ and $d + d_\alpha$ values. in the present results, $d + d_{0.05}$ is distinctly larger than +2, so that no firm conclusion can be reached. Nevertheless, like in the first case, the apparent convergence between the observed difference and the test outcome seems often favorable to NHST users who tend to erroneously conclude that the drug is inefficient.

Conclusion

In a sense, a p-value cannot be regarded as a rational measure of weight of evidence (see *e.g.*, Hacking, 1965 ; Spielman, 1974). It must also be stressed that a p-value *in itself* says nothing about the magnitude of the effect. However it must be acknowledged that in many usual cases the test statistic, or equivalently the p-value, can be straightforwardly combined with a descriptive statistic to obtain an interval estimate. Unlike power this interval estimate is directly and easily interpretable with respect to effect sizes. In actual fact Seldmeier & Gigerenzer (1989) deplored the neglect of power in experimental publications. In front of the misuses of NHST, they stated that "given such misconceptions, the calculation of power may appear obsolete because intuitively the level of significance already seems to determine all we want to know" (page 314). A more relevant assertion appears to be "given such misconceptions, the calculation of power may appear obsolete because *formally* the level of significance *may determine what we want to know*". This confirms Goodman and Berlin's assertion (1994) that "for interpretation of observed results, the concept of power has no place" (this does not mean that power cannot be useful for sample size calculations).

In particular, regarding the common misuses of NHST (see e.g., Lecoutre *et al.*, 2003), it follows that a "very significant" result generally allows the descriptive result to be extended. However, depending on the observed effect size, this can lead as well to assert a large, a medium, or a small effect. On the contrary a "very nonsignificant" result will lead to assert a small effect only if the observed effect is very small. In practice it will correspond more often than not to a statement of ignorance.

In conclusion, even if banning NHST in expérimental publications would be without doubt a shock therapy (see Shrout, 1997), t statistics, F ratios and p-values would remain useful, at least for computations of interval estimates and reanalyzes of previously published results. Ironically reporting them with sufficient accuracy appears then to be a valuable practice for subsequent analyses about effect sizes.

References

Fisher R. A. (1990) – *Statistical Methods, Experimental Design, and Scientific Inference* (Re-issue). Oxford : Oxford University Press.

Goodman, S.N. & 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.

Hacking, I. (1965) – *The Logic of Statistical Inference*. Cambridge, England : Cambridge University Press.

Lecoutre, B., Lecoutre, M.-P., & Poitevineau, J. (2001) – Uses, abuses and misuses of significance tests in the scientific community : won't the Bayesian choice be unavoidable ? *International Statistical Review*, **69**, 399-418.

Lecoutre M.-P., Poitevineau J., & Lecoutre B. (2003) – Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology* **38**, 37-45.

Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., & Le Roux, B. (2000) – *New Ways in Statistical Methodology : From Significance Tests to Bayesian Inference*, $2^n d$ edition. Bern, CH : Peter Lang.

Spielman, S. (1974). The Logic of Tests of Significance – *Philosophy of Science*, **41**, 211–226.

Seldmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies - *Psychological Bulletin*, **105**, 309–316.

Shrout, P.E. (1997). Should significance tests be banned ? – Introduction to a special section exploring the pros and cons, *Psychological Science*, **8**, 1–2.

Wilkinson, L. and Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical Methods in Psychology Journals : Guidelines and Explanations. *American Psychologist*, **54**, 594–604.