The Significance Test Controversy and the Bayesian Alternative

Bruno Lecoutre[1]
C.N.R.S. and Université de Rouen, France
Jacques Poitevineau[2]
C.N.R.S. and Université Pierre et Marie Curie, France

Keywords: Experimental data analysis, Statistical inference, Significance tests, Confidence intervals, Frequentist methods, Bayesian methods, Fisher, Predictive probabilities

The Significance Test Controversy

> "*It is very bad practice to summarise an important investigation solely by a value of* P" (Cox, 1982, page 327)

In spite of some recent changes, null hypothesis significance tests are again conventionally used in most scientific experimental publications. According to this publication practice, each experimental result is dichotomized: significant vs. nonsignificant. But scientists cannot in this way find appropriate answers to their precise questions, especially in terms of effect size evaluation. It is not surprising that, from the outset (e.g. Boring, 1919), significance tests have been subject to intense criticism. Their use has been explicitly denounced by the most eminent and most experienced scientists, both on theoretical and methodological grounds, not to mention the sharp controversies on the very foundations of statistical inference that opposed Fisher to Neyman and Pearson, and continue to oppose frequentists to Bayesians. In the sixties there was more and more criticism, especially in the behavioral and social sciences, denouncing the shortcomings of significance tests: *the significance test controversy* (Morrison & Henkel, 1970).

Significance Tests Are Not a Good Scientific Practice

> "*All we know about the world teaches us that the effects of A and B are always different - in some decimal place - for any A and B. Thus asking 'Are the effects different?' is foolish.*" (Tukey, 1991, page 100)

In most applications, no one can seriously believe that the different treatments have produced no effect: the point null hypothesis is only a *straw man* and a significant result is an evidence against an hypothesis known to be false before the data are collected, but not an evidence in favor of the alternative hypothesis. It is certainly not a good scientific practice, where one is expected to present arguments that support the hypothesis in which one is really interested. The real problem is to obtain estimates of the sizes of the differences.

The Innumerable Misuses of Significance Tests

> "*The psychological literature is filled with misinterpretations of the nature of the tests of significance.*" (Bakan, in Morrison & Henkel, 1970, page 239)

Due to their inadequacy in experimental data analysis, the practice of significance tests entails considerable distortions in the designing and monitoring of experiments. It leads to innumerable misuses in the selection and interpretation of results. The consequence is the existence of publication biases denounced by many authors: while non significant results are – theoretically – only statements of ignorance, only the significant results would really deserve publication.

The evidence of distortions is the use of the symbols *NS*, *, **, and *** in scientific journals, as if the degree of significance was correlated with the meaningfulness of research results. Many researchers and journal editors appear to be "star worshippers": see Guttman (1983), who openly attacked the fact that some scientific journals, and *Science* in particular, consider the significance test as a criterion of scientificness. A consequence of this overeliance on significant effects is that most users of statistics overestimate the probability of replicating a significant result.

The Considerable Difficulties Due to the Frequentist Approach

> "*What the use of* P *implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure.*" (Jeffreys, 1961/1939, Section 7.2)

Since the p-value is the proportion of samples that are "at least as extreme" as the observed data (under the null hypothesis), the rejection of the null hypothesis is based on the probability of the samples that *have not been observed*, what Jeffreys ironically expressed in the above terms. This mysterious and unrealistic use of the sampling distribution for justifying null hypothesis significance tests is for the least highly counterintuitive. This is revealed by questions frequently asked by students and statistical users: "why one considers the probability of samples outcomes that are more extreme than the one observed?"

Actually, due to their frequentist conception, significance tests involve considerable difficulties in practice. In particular, many statistical users misinterpret the p-values as inverse (Bayesian) probabilities: 1-p is "the probability that the alternative hypothesis is true." All the attempts to rectify this misinterpretation have been a loosing battle.

Significance Tests Users' Dissatisfaction

> "*Neither Fisher's null hypothesis testing nor Neyman-Pearson decision theory can answer most scientific problems.*" (Gigerenzer, 2004, page 599)

Several empirical studies emphasized the widespread existence of common misinterpretations of significance tests among students and scientists (for a review, see Lecoutre, Lecoutre & Poitevineau, 2001). Many methodology instructors who teach statistics, including professors who work in the area of statistics, appear to share their students' misinterpretations. Moreover, even professional applied statisticians are not immune to misinterpretations of significance tests, especially if the test is nonsignificant. It is hard to interpret these finding as an individual's lack of mastery: they reveal that significance test do not address the questions that are of primary interest for the scientific research.

> "*But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested.*" (Rozeboom, in Morrison & Henkel, 1970, page 221)

In particular, the dichotomous significant/non significant outcome of significance tests strongly suggests binary research decisions: "reject/accept the null hypothesis". The "reject/accept" attitude is obviously a poor and unfortunate decision practice.

• A statistically significant test provides no information about the departure from the null hypothesis. When the sample is large a descriptively small departure may be significant.

• A nonsignificant test is not evidence favouring the null hypothesis. In particular, a descriptively large departure from the null hypothesis may be nonsignificant if the experiment is insufficiently sensitive.

In fact, in order to interpret their data in a reasonable way, users must resort to a more or less naive mixture of significance tests outcomes and other information. But this is not an easy task! This leads users to make *adaptative distortions*, designed to make an ill-suited tool fit their true needs. Actually, many users explicitly appear to have a real consciousness of the stranglehold of significance tests: in many cases they use them only because they know no other alternative.

The Case for Confidence Intervals

On the one hand, it is not acceptable that the users of statistical inference methods will continue using non appropriate procedures because they know no other alternative. On the other hand, the times we're living in at the moment appear to be crucial. Changes in reporting experimental results are more and more enforced within editorial policies. Most of these changes are explicitly intended to deal with the essential question of *effect sizes*.

> "*Science is inevitably about magnitudes.*" (Cohen, 1990, page 1309)

Reporting an effect size estimate is one of the first necessary steps in overcoming the abuses of null hypothesis significance tests. It can effectively prevent researchers from unjustified conclusions in the conflicting cases where a non-significant result is associated with a large observed effect size. However, small observed effect sizes are often illusorily perceived by researchers as being *favorable* to a conclusion of no effect, when they can't in themselves be considered as sufficient proof.

Power studies can also be seen as a handrail to avoid hasty generalizations. However referring to statistical papers that discuss and compare procedures, a more and more widespread opinion is that the concept of power is inappropriate for interpreting results.

The majority trend, reinforced by editorial policies, is to advocate the use of confidence intervals, in addition to or instead of significance tests. Consequently, confidence intervals are more and more frequently reported, either about raw or standardized effect sizes. However, reporting confidence intervals appears to have very little impact on the way the authors interpret their data. Most of them continue to focus on the statistical significance of the results. They only wonder whether the interval includes the null hypothesis value, rather than on the full implications of confidence intervals: the steamroller of significance tests cannot be escaped.

> "*Inevitably, students (and everyone else except for statisticians) give an inverse or Bayesian twist to frequentist measures such as confidence intervals and* P *values.*" (Berry, 1997, page 242)

Furthermore, for many reasons due to their *frequentist* conception, confidence intervals can hardly be seen as the ultimate method. Indeed it can be anticipated that the conceptual difficulties encountered with the frequentist conception of confidence intervals will produce further dissatisfaction. In particular, users will realize that the appealing feature of confidence intervals is the result of a fundamental misunderstanding. As is the case with significance tests, the frequentist interpretation of a 95% confidence interval involves a long run repetition of the same experiment: in the long run 95% of computed confidence intervals will contain the "true value" of the parameter; each interval in isolation has either a 0 or 100% probability of containing it. Unfortunately treating the data as random even after observation is so strange this "correct" interpretation does not make sense for most users. Ironically it is the interpretation in (Bayesian) terms of "a *fixed* interval having a 95% chance of including the true value of interest" which is the appealing feature of confidence intervals. Moreover these incorrect natural interpretations of confidence intervals (and of significance tests) are encouraged by most statistical instructors who tolerate and even use them.

> "*It would not be scientifically sound to justify a procedure by frequentist arguments and to interpret it in Bayesian terms.*" (Rouanet, in Rouanet et al., 2000, page 54)

What a paradoxical situation! We then naturally have to ask ourselves whether the "Bayesian choice" (Robert, 2001) will not, sooner or later, be an unavoidable alternative.

The Bayesian Choice

> "*At the very least, use of noninformative priors should be recognized as being at least as objective as any other statistical techniques.*" (Berger, 1985, page 110)

Time's up to come to a positive agreement for objective procedures of experimental data analysis that bypass the common misuses of significance tests. This agreement should fill up its role of "an aid to judgment," which should not be confused with automatic acceptance tests. Undoubtedly, there is an increasing acceptance that Bayesian inference can be ideally suited for this purpose. However, the contribution of Bayesian inference to experimental data analysis and scientific reporting has been obscured by the fact that many authors concentrate too much on the decision-theoretic elements of the Bayesian approach. This perpetuates the poor "reject/accept" attitude of significance tests. Without dismissing the merits of the decision-theoretic viewpoint, it must be recognized that there is another approach which is just as Bayesian which was developed by Jeffreys in the thirties (Jeffreys, 1961/1939). Following the lead of Laplace (Laplace, 1986/1825), this approach aimed at assigning "noninformative" prior probabilities when nothing was known about the value of the parameter. In practice, *noninformative* prior probabilities are vague distributions which, *a priori*, do not favor any particular value: they let the data "speak for themselves".

In this form the Bayesian paradigm provides *reference* methods appropriate for situations involving scientific reporting. Nowadays, thanks to the computer age, an *objective Bayes theory* is by no means a speculative viewpoint but on the contrary is perfectly feasible. Such a theory is better suited to the needs of users than the frequentist approach and provides scientists with relevant answers to essential questions raised by experimental data analysis. Bayesian methods allow users to overcome usual difficulties encountered with the frequentist approach. In particular, using the Bayesian interpretations of significance tests and confidence intervals in the language of probabilities about unknown parameters is quite natural for the users. In return, the common misuses and abuses of NHST are more clearly understood. In particular, users of Bayesian methods become quickly alerted that non-significant results cannot be interpreted as "proof of no effect."

Objective – or Fiducial – Bayesian Analysis: Reconciling Fisher and Bayes?

> "*A widely accepted objective Bayes theory, which fiducial inference was intended to be, would be of immense theoretical and practical importance.* [...] *A successful objective Bayes theory would have to provide good frequentist properties in familiar situations, for instance, reasonable coverage probabilities for whatever replaces confidence intervals.*" (Efron, 1998, pages 106, 112)

Routine Bayesian methods for the most familiar situations encountered in experimental data analysis are now available. Their aim is to let the statistical analysis express what the data have to say independently of any outside information (Lecoutre, 2008). They can be learned and used as easily, if not more, as the t, F or chi-square tests, and offer promising new ways in statistical methodology (Rouanet et al., 2000). Extensive applications to real data have been done and have been accepted well in experimental publications.

In order to promote these methods, it is important to give them an explicit name. Berger (2004) clearly denounced "the common misconception that Bayesian analysis is a subjective theory" and proposed the name *objective Bayesian analysis*. With the same incentive, an alternative name is *fiducial Bayesian methods*, which pays tribute to Fisher's work about "scientific inference for research workers" (Fisher, 1990/1925) and makes explicit that these methods are specially designed for use in experimental data analysis.

Other Bayesian Techniques are Promising

> "*An objective scientific report is a report of the whole prior-to-posterior mapping of a relevant range of prior probability distributions, keyed to meaningful uncertainty interpretations.*" (Dickey, 1986, page 135)

An analysis of experimental data should always include an objective Bayesian analysis in order to gain "public use" statements. However, informative Bayesian priors also have an important role to play in experimental investigations. They may help refining inference and investigating the sensitivity of conclusions to the choice of the prior. Informative Bayesian techniques are ideally suited for combining information from the data in hand and from other studies, and therefore planning a series of experiments. Ideally, when "good prior information is available", it could (should) be used to reach the same conclusion that an "objective Bayesian analysis", but with a smaller sample size.

> "*The essence of science is replication: a scientist should always be concerned about what would happen if he or another scientist were to repeat his experiment.*" (Guttman, 1983)

The predictive idea is central in experimental investigations. A major strength of the Bayesian paradigm is the ease with which one can make predictions about future observations. Bayesian predictive probabilities, because they relate observables between each other, are very intuitive and even more natural than posterior probabilities about parameters. They may wonderfully complement the Bayesian inference about parameters (Lecoutre, Lecoutre & Poitevineau, 2010).

In this way, interesting enough is the fact that Peter Killeen (for an up-to-date discussion, see Killeen, 2008) recently suggested a new statistic *p*rep (for "probability of replication") to present experimental results. This statistic is the fiducial Bayesian predictive probability of finding a same-sign effect in an exact replication of an experiment. It has been extensively used in *Psychological Science*, and more occasionally in other journals. It follows that for the first time a Bayesian probability has been routinely reported in scientific publications: a promising first step towards a broader recognition of the usefulness of the Bayesian approach?

> "*An essential aspect of the process of evaluating design strategies is the ability to calculate predictive probabilities of potential results.*" (Berry, 1991, page 81)

Bayesian predictive procedures give users a very appealing method to answer essential questions such as: "how big should be the experiment to have a reasonable chance of demonstrating a given conclusion?", "at an interim stage, given the current data, what is the chance that the final result will be in some sense conclusive, or on the contrary inconclusive?" These questions are unconditional in that they require consideration of all possible values of parameters. Whereas traditional frequentist practice does not address these questions, predictive probabilities give them a direct and natural answer.

For instance, from a pilot study, the predictive probabilities on credible limits for the parameter of interest give a useful summary to help in the choice of the sample size of an experiment. The predictive approach is also a valuable tool for interim analyses, and more generally for missing data imputation. It is a very appealing method to aid the decision to stop an experiment at an interim stage. On the one hand, if the predictive probability that it will be successful appears poor, it can be used as a rule to abandon the experiment for futility. On the other hand, if the predictive probability is sufficiently high, this suggests to early stop the experiment and conclude success. This is of primary importance in medical studies where ethical questions are concerned.

Based on an article from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science +Business Media, LLC.

References

Berger, G.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New-York: Springer Verlag.

Berger, J. (2004). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 1–17.

Berry, D. A. (1991). Experimental design for drug development: a Bayesian approach. *Journal of Biopharmaceutical Statististics* **1**, 81–101.

Berry, D.A. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician*, **51**, 241–246.

Boring, E.G. (1919). Mathematical versus scientific significance. *Psychological Bulletin*, **16**, 335–338.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304–1312.

Cox, D.R. (1982). Statistical significance tests. *British Journal of Clinical Pharmacology*, **14**, 325–331.

Dickey J.M. (1986) - Discussion of Racine, A., Grieve, A.P., Flühler, H., & Smith, A.F.M., Bayesian methods in practice: Experiences in the pharmaceutical industry. *Applied Statistics*, **35**, 93–150.

Efron, B. (1998). R.A. Fisher in the 21st century [with discussion]. *Statistical Science*, **13**, 95–122.

Fisher, R. A. (1990/1925). *Statistical Methods for Research Workers* (Reprint, 14th edition, 1925, edited by J.H. Bennett). Oxford: Oxford University Press.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, **33**, 587–606.

Guttman, L. (1983). What is not what in statistics? *The Statistician*, **26**, 81–107.

Jeffreys, H. (1961/1939). *Theory of Probability* (3rd edition; 1st edition: 1939). Oxford: Clarendon.

Killeen, P.R. (2008). Replication statistics as a replacement for significance testing: Best practices in scientific decision-making. In J.W. Osborne (Ed.), Best practices in quantitative methods. Thousand Oaks, CA: Sage Publishing, 103–124.

Laplace, P.-S. (1986/1825). *Essai Philosophique sur les Probabilités* (Reprint, 5th edn, 1825). Christian Bourgois, Paris (English translation: *A Philosophical Essay on Probability*, 1952, Dover, New York).

Lecoutre, B. (2008). Bayesian methods for experimental data analysis. In C.R. Rao, J. Miller & D.C. Rao (eds), *Handbook of statistics: Epidemiology and Medical Statistics Vol. 27.* Elsevier: Amsterdam, 775–812.

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, B., Lecoutre, M.-P. & Poitevineau, J. (2010). Killeen's probability of replication and predictive probabilities: How to compute, use and interpret them. *Psychological Methods*, **15**, 158–171.

Morrison, D.E., & Henkel, R.E. (Eds.) (1970). *The Significance Test Controversy - A Reader*. London: Butterworths.

Robert, C.P. (2001). The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (2nd edition). New York: Springer.

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

Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science*, **6**, 100-116.

A detailed bibliography can be found at address:

`http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.htm`

---

[1] Bruno Lecoutre, ERIS and UPRESA 6085, Laboratoire de Mathématiques Raphaël Salem, C.N.R.S. et Université de Rouen, Mathématiques, Site Colbert, 76821 Mont-Saint-Aignan Cedex, France.

[2] Jacques Poitevineau, ERIS and UMR 7190, IJLRA/LAM/LCPE, C.N.R.S., Université Pierre et Marie Curie, 11 rue de Lourmel, 75015 Paris, France.