Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 1, pp. 1-9

CRISIS IN SCIENCE? OR CRISIS IN STATISTICS! MIXED MESSAGES IN STATISTICS WITH IMPACT ON SCIENCE

D. A. S. FRASER AND N. REID
Department of Statistical Sciences, University of Toronto, Toronto, Canada M5S 3G3, Canada
Email: dfraser@utstat.toronto.edu, reid@utstat.toronto.edu

SUMMARY

Gelman and Loken (2014) draw attention to a "statistical crisis in science" and describe how risks with multiple p-values can be present even in the analysis of a single data set. There is indeed a crisis, as p-values are everywhere, in science, engineering, medicine, social science, health care, and the media; and conflicting results are misrepresenting the importance of p-values, and indeed of many disciplines themselves. We argue that risks of misinterpretation are widespread, but that the crisis is really in the discipline of statistics, and starts with mixed messages about the meaning and usage of p-values. These mixed messages then have downstream effects that seriously misinform scientific endeavours. What are these mixed messages concerning p-values? And should statistics continue with such messages that compromise the discipline? We discuss this and offer recommendations.

Keywords and phrases: Bayesian, frequentist, likelihood, multiple testing, p-value-function. significance function

AMS Classification: 62A01, 62F99

1 Introduction

This article is a response to Gelman and Loken (2014), who drew attention to a "statistical crisis in science" and showed how multiple p-values can arise, in good faith, in the analysis of a single data set. At about the same time, the *Journal of Basic and Applied Social Psychology* made headlines in *Nature* (Woolston, 2015) by deciding to no longer publish papers containing p-values. This debate continues, and there were several media reviews of news in December 2015 from CERN's Large Hadron Collider about a possible discovery of a new particle, and the associated "5-sigma" criterion commonly applied in high-energy physics (Castelvecchi, 2015; Spiegelhalter, 2015).

There is a crisis as p-values are everywhere, in science, engineering, medicine, social science, health care, and in the standard media phrase "19 times out of 20" commonly appearing in the reporting of polls. Our view is that while the risks of misinterpretation of p-values are widespread, the crisis is really in the discipline of statistics, in providing mixed messages about the meaning of a p-value. These mixed messages have downstream effects that can seriously affect all applications. We discuss this and offer recommendations.

2 Multiple meanings

2.1 The p-value function

Suppose we observe a variable, say y, that measures an unknown θ of interest; thus y is accessible through measurement, but θ is only indirectly accessible, through inference from y. If we had unlimited time and resources we could collect a great many values of the variable y and obtain the probability distribution of the variable y. This density indicates the stochastic behaviour of the variable, and if we assume that the form of the density is known, but its location (for example) is not, by identifying this via an unknown parameter θ we can view learning where the distribution is located as learning the value of θ. This could be, and often is, formalized by having a hypothesis, called a null hypothesis and designated H_0, that the unknown true value θ is θ_0; an example is indicated in Figure 1.

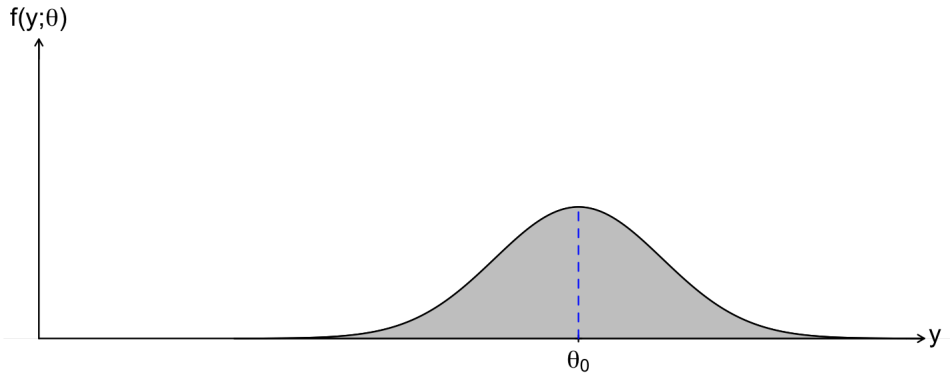

Figure 1: An accumulation of observations of y when the null hypothesis $H_0 : \theta = \theta_0$ holds.

Given a single observed measurement y^0, an investigator could then construct Figure 2, which shows that a proportion 6.1% of the distribution $\theta = \theta_0$ falls to the left of the observed measurement y^0, and 93.9% falls to the right. The observed p-value associated with H_0 would then be $p^0 = 6.1\%$ and is thus presenting just the percentile or statistical position of the data y^0 under H_0, or recording just a pure statement of factual information. As a definition this aligns with Fisher's 1920 proposal, later clarified in Fisher (1956).

This example is simplified to an extreme, but asymptotic arguments developed in Fraser (1990), Fraser and Reid (1993), and Brazzale et al. (2007, Ch. 8) show in wide generality that there is in fact such an approximating location model relevant to a single parameter of interest and that it can can be calculated quite routinely with more complex and realistic models.

Common statistical custom and usage don't usually stop with this percentile position, but proceed from the statistical position to scientific statements with potentially huge impact. For example, in high-energy physics, θ_0 could represent the mean value under background radiation, and then

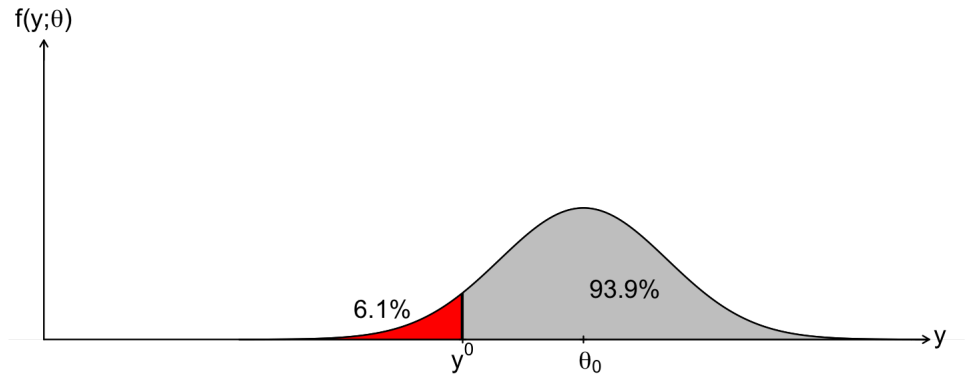

Figure 2: An observed data point y^0 and proportions left and right of the data under the hypothesis H_0.

larger values $\theta > \theta_0$ could indicate a new particle, such as the Higgs boson. In what sense does the p-value provide support for this new particle?

More informative than a single p-value, the p-value function $p(\theta)$, records the statistical position of the observed data y^0 for a range of values of θ: see Figure 3. This function presents the "statistical position" of the observation. It does not single out particular alternatives to θ_0, but leaves this choice to appropriate judgement in an application context Fraser (2014).

The p-value function presents in one plot all possible confidence bounds: we could for example solve $0.95 = p(\theta; y^0)$, the solution of which, $\hat{\theta}_L$ say, is a lower confidence bound at the conventional 95% limit. Under repeated observation of y from the model, the interval $(\hat{\theta}_L, \infty)$ will include the true value of θ 19 times out of 20, on average. The p-value function has also been called the confidence distribution function, e.g. in Cox (1958), Efron (1993), Xie and Singh (2013), Hjort and Schweder (2016). The p-value function or confidence distribution function has the added benefit that the direction of departure is recorded, as well as the magnitude.

2.2 Decision theory

Calculating observed proportions such as 0.061 and 0.939 as above was historically often challenging, and reference values corresponding to one or several standard values such as 5%, 10%, 90%, and 95% were derived and recorded in tables. Then in an investigation a statement such as "significant at the 10%" level, or "not significant at the 5% level", would be offered for the data point y^0 in Figure 2.

With the development of the theory of hypothesis testing by Neyman and Pearson (1933), this practice acquired a formal theoretical status. In due course the original concept of a p-value or observed level of significance as the position of the data with respect to the model changed its presentation into a decision for or against the hypothesis H_0, at some chosen fixed level of significance. The observed value y^0 then became a decision for, or against, some null value. Taking such de-

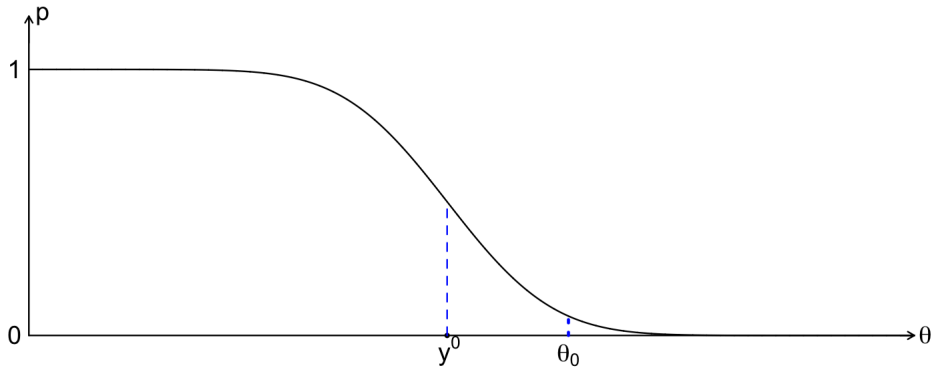

Figure 3: The observed p-value function from data y^0 as in Figure 3; height at θ_0 is 6.1%.

cisions at face value is a substantial change from the notion of statistical position, and has had the profound and unfortunate consequence of setting an arbitrary standard for determining the adequacy, or even publishibility, of the results from an experiment.

When p-values are used only to make a decision, and a larger sample size is viewed as a route to getting to the decision point faster, the results can be even more misleading.

Gelman and Loken express this concern for treating p-values from a decision theoretic viewpoint: "By convention, a p-value below 0.05 is considered a meaningful refutation of the null hypothesis: however, such conclusions are less solid than they appear". They do not, however, dwell further on this point. Many contemporary presentations of introductory statistics also overlook such concerns.

The point was emphasized famously in Ioannidis (2005), but there is a much earlier literature warning about this. Sterling (1959) wrote of "publication decisions and their possible effects on inferences drawn from tests of significance"; in particular "... (where) a borderline between acceptance and rejection is taken (at a) fixed point (say) 0.05 ... is interesting by itself ... (and when) used as a critical criterion for selecting reports for (publication) in professional journals (might result in) unanticipated results." Rozeboom (1960) wrote of "The fallacy of the null-hypothesis significance test" and quoted a famous philosophical epigram that the "accept-reject" paradigm is the "glory of science and the scandal of philosophy", meaning the glory of statistics and the scandal of logic and application.

2.3 Bayesian view of p-value

To this point we have assumed that the model for Figure 1 provides the full background information for θ. Another approach is available if we have a function $\pi(\theta)$ allegedly describing a probability density for potential values of θ. If the joint model is then accepted as valid, the application of the basic rules of conditional probability enable calculation of a probability distribution for θ, given the

observed measurement y^0, as $f(\theta \mid y^0) = c\pi(\theta)f(y;\theta)$. We can then compute, for example, the probability that θ is larger than θ_0, having observed y^0.

But where does such a probability density function $\pi(\theta)$ come from? Efron (2013) cites two possibilities: there may indeed be a case in some applications where randomness for the source of the true θ can be identified with a distribution $\pi(\theta)$: he calls this a genuine prior. If θ represents the rate of defectives in a manufacturing process, there may be enough data from previous manufacturing runs to identify such a distribution.

An alternative construction of a distribution $\pi(\theta)$ is by describing symmetries among various θ values: Efron (2013) calls these Laplace priors, as they received special support from Laplace (1812). In that case the construction of $f(\theta \mid y^0)$ can be regarded as a completely formal exercise, not embodying any probability interpretation. In this setting the best we could argue is that these probabilities have a meaning in as much as they lead to identical conclusions as the p-value function. Then the probability interpretation of the result is vacuous, but not misleading.

This is the case, for example, in a simple location model with a uniform prior for θ. The frequency calculation and the Bayes posterior probability calculation are computational reflections of each other; thus $s^0(\theta_0) = \int_{\theta_0} f(\theta \mid y^0)d\theta$ attaches the same value, 6.1%, to the statement that θ is larger than θ_0 as the argument above attaches to the probability under the model $f(y;\theta_0)$ that y is less than y^0: the Bayes posterior bound is in fact exactly a confidence bound: see Figure 4.

In our view the two ingredients $\pi(\theta)$ and $f(y;\theta)$, even if $\pi(\theta)$ is a genuine prior, should be left separate, rather than being combined into a joint model $\pi(\theta)f(y;\theta)$ describing the pair (y,θ). This makes available the full background information, and leaves to the concerned user the option to combine them if desired. This point is discussed from a slightly different point of view in Cox (2006, Ch. 5) and Cox and Reid (2015),where it is argued that "personalistic" priors have a different logical status from probability density functions.

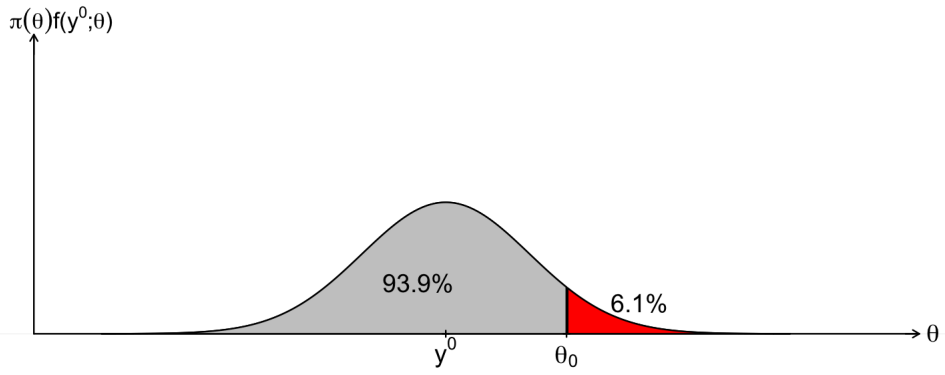

Figure 4: The Bayes calculation with the Laplace noninformative prior can with location symmetry duplicate the frequency calculation, thus giving a confidence result.

3 Responsibility and Risks

We have discussed three different interpretations for "p-value" or "level of significance": (i) The frequency view: The statistical position of the observed data with respect to a θ_0 value being tested; (ii) The decision theory view: The conventional level at which the data is just significant with respect to a θ_0 value being tested; and (iii) The Bayes view: The Bayes survivor calculation at a θ_0 value using some prior distribution for θ.

It is our view that the discipline of statistics should acknowledge responsibility for the consequences of the confusion, in many areas of application, caused by these multiple meanings. Fraser (2014) highlights three historically prominent cases where responsibility for statistical steps seems overwhelming, even in legal senses. The launch of the space-shuttle Challenger failed on January 28, 1986, causing seven deaths: statistical data available before the flight indicated a concern with the effect of low temperatures on critical O-rings, but the statistical warnings were by-passed (Dalal et al., 1989). The pain relief drug Vioxx was approved by the US Food and Drug Administration in 1999, but withdrawn by the pharmaceutical company in 2004 after evidence for an elevated risk of heart attacks became overwhelming, although statistical assessments as early as 2000 had indicated heightened risk of such serious events (Abraham, 2005). An estimated 40,000 people died and a five billion dollar settlement with the pharmaceutical company was obtained for those injured or the survivors (ONeil, 2012). Before the L'Aquila earthquake on April 5, 2009 an official committee with statistical expertise underemphasized in public statements the risk of an imminent major earthquake; some 300 died in that earthquake, and seven committee members were convicted of manslaughter (Marshall, 2012; Prats, 2012), a conviction that was overturned on appeal for six of the members (Abbott and Nosengo, 2014).

These examples emphasize that a misleading use of statistics can have serious consequences in lives lost and in billions of dollars in costs. These consequences can start with conflicting messages from statistics, and in particular the message that "statistical significance" is treated as an absolute, a decision, and that the goal of the statistical analysis of an observed set of data is to reach that elusive bar: a theme very common to applied work, especially among those new to the research process.

Gelman and Loken (2014) focus their discussion on the decision theoretic interpretation and address the consequences from this approach, emphasizing in particular the problem that for a given scientific or social scientific problem, the translation of "interesting science" to "statistical hypothesis" can, and often does, involve several hypotheses, and hence the calculation of several p-values, with a particular data set. They write "It would take a highly unscrupulous researcher to perform test after test in a search for statistical significance ... at the 0.05 level ... The difficult challenge lies elsewhere". They further note "it is reasonable for scientists to refine their hypotheses in light of the data". Their assessment of the risks emphasizes that the formulation of an hypothesis in science or social science is not as straightforward as identifying a single θ_0, and as a result multiple testing is implicit in a great many analyses, and more subtle than carrying out several tests in search of "$p < 0.05$".

We agree with them that the risks of using arbitrary p-values to define 'significance', and using these as decisions is very serious when multiple formulations of hypotheses lead explicitly or implicitly to large numbers of p-values. Among their recommended strategies of pre-registration,

authentic replication, and analysis of "all data", they include a claim "that p-values should not necessarily be taken at face value". This last we disagree with! It is the conventional but unwarranted attribution of decision, and the use of p-values for journal management, that are at the heart of the problem.

The p-value and p-value function is simply recording the statistical position of data relative to an hypothesis; it is elemental and provides an appropriate starting point for inference conclusions. It can guide the judgments about scientific conclusions, but cannot replace them. The consensus judgment in high-energy physics is that a 'discovery' is claimed when the p-value is less than 1 in 3.5 million: it is called "5-sigma" as this is the probability that a normal variable is greater than five standard deviations from the mean, the normal here being an approximation to the Poisson count of number of observed particles. Another physics example that received wide publicity in the popular media of the time was Eddington's verification of Einstein's theory of general relativity. The orbit of Mercury had been known in the 18-hundreds to precess at a rate different from that predicted by Newtonian mechanics, and Einstein's general relativity provided an adequate explanation. But further corroboration seemed appropriate to the physics community. General relativity also predicts the bending of light rays as they pass near a large mass; this provided, then, an appropriate variable to measure, and in May 1919 Eddington was able to carefully measure the apparent position of stars in the sky as indicted by light from the stars after it had passed adjacent to the sun during a solar eclipse.

Suppose, as viewed, the star light was passing on the right side of the hidden sun where general relativity would indicate that its apparent position in the sky was displaced to the right. Then if a 5-sigma event had been observed, the statistical position of the observed data would have been $p = 0.999,999,7$, indicating the large departure to the right; this value is the complement of $0.000,000,3$, in turn the reciprocal of 1 in 3.5 million. This p value records that data value was large, near 1; it is in the right tail of the null distribution under the standard theory of the time. The statistical position version of the p-value is appropriate and indicates the magnitude of the departure as well as the type of departure.

We believe the discussion is more urgent now, in the era of Big Data. As a reviewer has emphasized, the use of false discovery rates has been developed as a method of protecting against multiple hypothesis tests. In applications of many similar tests to a single set of data, for example in genome-wide association studies, this has provided some protection against claims of discoveries that could not subsequently be validated. Indeed the conventional, if somewhat arbitrary, 5-sigma rule of high energy physics is an *ad hoc* correction for multiple testing to protect exactly against false discoveries. This seems not to solve the issue, but rather to move the decision boundary.

An approach more directly aligned with the presentation of the p-value function is a method to correctly combine many such functions into a single summary p-value function. Methods of combination motivated by developments in the theory of composite likelihood are in development (Fraser and Reid, 2016).

For a great many settings where Big Data is available for analyses, the calculation of the dimensionality for possible hypotheses may be difficult or impossible, and the potential for making incorrect decisions is enormous. Attributing significance or decision to a comparison selected from

among millions of potential hypotheses suggests serious rethinking of the exploration process, the evaluation process, and the decision process. The risks for misleading decisions seem large; we could have mega p-values, mega decisions and mega wrong 'answers'. Scientists and social scientists are making serious efforts to address these issues; see for example the *Science* editorial McNutt (2014), and Gelman and Loken (2014)'s suggestions around pre-registration. Perhaps Statistics should stand up for its responsibilities before a Big Data Disaster.

4 Acknowledgement

We gratefully acknowledge discussions with Ian Spence in the Department of Psychology at the University of Toronto. This research has received support from the National Science and Engineering Research Council of Canada and the Senior Scholars Funding of York University. We thank reviewers of an earlier version for helpful suggestions for improvement.

References

[1] Abbott, A. and Nosengo, N. (2014). Re: Acquittal of 6 of the members of the committee. *Nature*, 515:7526.

[2] Abraham, C. (2005). Study finds Vioxx took deadly toll. *The Globe and Mail* 25 January 2005.

[3] Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics*. Cambridge University Press, Cambridge.

[4] Castelvecchi, D. (2015). LHC sees hint of boson heavier than Higgs. *Nature News* 15 December 2015.

[5] Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29:357–372.

[6] Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press: Cambridge.

[7] Cox, D. R. and Reid, N. (2015). On some principles of statistical inference. *Intern. Statist. Rev.*, 83:293–308.

[8] Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. *J. Am. Statist. Assoc.*, 84:945–957.

[9] Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80:3–26.

[10] Efron, B. (2013). Bayes' theorem in the 21st century. *Science*, 340:1177–1178.

[11] Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

[12] Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, 77:65–76.

[13] Fraser, D. A. S. (2014). Why does statistics have two theories? In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present and Future of Statistical Science*, pages 237–252. CRC Press., Florida.

[14] Fraser, D. A. S. and Reid, N. (1993). Third order asymptotic models: likelihood functions leading to accurate approximation to distribution functions. *Statistica Sinica*, 3:67–82.

[15] Fraser, D. A. S. and Reid, N. (2016). On combining likelihoods and p-values. unpublished;

[16] Gelman, A. and Loken, E. (2014). The statistical crisis in science. *Amer. Scientist*, 102:460–465.

[17] Hjort, N. L. and Schweder, T. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press: Cambridge.

[18] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2. e124. doi:10.1371/journal.pmed.0020124.

[19] Laplace, P. S. d. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier.

[20] Marshall, M. (2012). Seismologists found guilty of manslaughter. *New Scientist*, 22 October 2012.

[21] McNutt, M. (2014). Journals unite for reproducibility. *Science*, 346:679.

[22] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of a statistical hypothesis. *Phil. Trans. Roy. Soc. A*, 231:289–337. Reprinted in *Joint Statistical Papers of J. Neyman and E.S. Pearson*, Cambridge University Press, Cambridge, 1967.

[23] O'Neil, C. (2012). How Big Pharma cooks data – the case of Vioxx and heart disease. `http://www.nakedcapitalism.com/2012/02/25244.html`.

[24] Prats, J. (2012). The L'Aquila earthquake: Science or risk on trial? *Significance*, 9:13–16.

[25] Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57:416–428.

[26] Spiegelhalter, D. (2015). What are sigma levels? *Plus Magazine*, 18 December 2015.

[27] Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *J. Amer. Statist Assoc.*, 54:30–34.

[28] Woolston, C. (2015). Psychology journal bans P values. *Nature News*.

[29] Xie, M. G. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review (with discussion). *Intern. Statist. Rev.*, 81:3–39.