Crisis in Science? or Crisis in Statistics.  
Mixed messages and the user impact

D.A.S. Fraser and N. Reid *

December 16, 2014

Gelman and Loken (2014) draw attention to a statistical “crisis in science”, arguing that the risks with a multiplicity of \( p \)-values can arise even in a single data set. This is a crisis because \( p \)-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 reporting of polls. We argue that the risks of misinterpretation are widespread, but that the crisis is really in the discipline of statistics and starts with mixed messages about the meaning of a \( p \)-value. These mixed messages have downstream effects that seriously affect practitioners. What are these mixed messages around \( p \)-values? And how can statistics continue with mixed messages, without compromising the discipline?

**Responsibility for mixed messages from Statistics**

The discipline of statistics has and should acknowledge responsibility for the consequences of its methods and procedures in many areas of application. Fraser (2014) highlights three historically prominent cases where the responsibility seems overwhelming, even in the legal sense. 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 The pain relief drug Vioxx was approved by the US Food and Drug Administration in 1999, but withdrawn by the pharmaceutical company in 2004 after evidence for an elevated risk of heart attacks became overwhelming, although statistical assessments as early as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 (O’Neil, 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). Statistics can have serious consequences in lives lost and in billions of dollars in costs. These consequences can start with conflicting messages from statistics.

![Figure 1: The massive data picture when θ is the ‘true value’](image)

Meaning for *p*-value

Suppose we observe a variable say *y* that measures an unknown θ of interest; the Latin letter for something accessible and the Greek letter for something ‘inaccessible’.
If we had unlimited time and resources we could collect a great many values of the variable $y$ and stack these data into a mound along an axis for the $y$, obtaining what can be called the distribution of the variable $y$, as indicated in Figure 1. This mound of data would present the basic behaviour of the variable and the centre of the pile would be the unknown true $\theta$, assuming for this discussion that the basic shape is known beforehand. Then learning where the distribution is located, means learning the value of $\theta$. This is often formalized by having a hypothesis, called a null hypothesis and designated $H_0$, that the unknown true value $\theta$ is say $\theta_0$; see Figure 2.

Given a single observed measurement $y^0$, an investigator could then construct Figure 3, which shows a proportion 6.1\% of the distribution with $\theta = \theta_0$ 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 be $p^0 = 6.1\%$, and is simply recording the percentile or statistical position of the data point under $H_0$. 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 that there is indeed an approximating location model relevant to a single parameter of interest that can be calculated from more complex and realistic models quite routinely.

Common statistical custom and usage don’t usually stop with this percentile po-
Figure 3: An observed data point $y^0$ and proportions left and right of the data under the hypothesis $H_0$.

...ision, but use the position to draw conclusions with potentially huge impact. For example, in high-energy physics $\theta_0$ could represent the mean observation under background radiation, and larger values $\theta > \theta_0$ indicate the presence of a new particle, such as the Higgs boson. In what sense does the $p$-value provide support for this alternative claim? We can take a page from Fisher’s recommendation to have closer ties to the likelihood function.

**Likelihood view of $p$-value**

More informative than a single $p$-value is the $p$-value function $p(\theta)$, which records the statistical position of the observed data $y^0$ as a function of the unknown $\theta$: see Figure 4. This function simply records where the the “statistical position” of the observation, relative to various $\theta$ values (Fraser, 2014). This $p$-value function does not single out particular types of alternative to $\theta_0$, but leaves such follow-on choices to appropriate judgement in an application context. We could for example use the ‘19 out of 20’ convention and solve $0.95 = p(\theta)$, the solution of which, $\hat{\theta}_L$ say, is a lower confidence bound: under the model the interval $(\hat{\theta}_L, \infty)$ will include the true value 19 times out of 20, on average.
Decision theory view of \( p \)-value

Calculating observed proportions such as 0.061 and 0.939 as above was historically often challenging, and reference values corresponding to one or several simple 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 4. This practise evolved into a decision for or against the hypothesis \( H_0 \), at the chosen level of significance, and in due course acquired a formal status in a theory of hypothesis testing (Neyman and Pearson, 1967). The original concept of a \( p \)-value or observed level of significance changed its meaning and acquired status as a decision. Such decisions are often then taken at face value, as having explicit backing from the hypothesis testing theory of Neyman and Pearson, and even treated as part of the theory of inference.

Bayesian view of \( p \)-value

To this point we have assumed that the model for Figure 1 is the full background information for \( \theta \). Another approach is available if we have a distribution \( \pi(\theta) \) describing a source for the value of \( \theta \). With these two ingredients, \( \pi(\theta) \) and \( f(y; \theta) \), we are faced with a modelling situation: should these be combined into a joint model \( \pi(\theta)f(y; \theta) \)
describing the pair \((y, \theta)\); or should they be left separate? But where would the \(\pi(\theta)\) come from? Efron (2013) cites two possibilities: the distribution \(\pi(\theta)\) describes identified randomness for the source of the true \(\theta\), what he calls a genuine prior; or the distribution \(\pi(\theta)\) describes symmetries among \(\theta\) values, which Efron calls Laplace priors, as they received special support from Laplace (1812).

If the joint model is accepted as valid then conventional calculations with the conditional probability lemma give a probability distribution \(f(\theta \mid y^0) = c\pi(\theta)f(y; \theta)\) for \(\theta\) given the observed measurement \(y^0\). This is fine in the genuine prior case, but if the prior is just expressing the symmetries that Laplace found attractive, then the conditional probability lemma says absolutely nothing. In spite of this the formal calculations in the location model have the property that 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 \(\theta\) is larger than \(\theta_0\) as the argument above attaches to the probability under the model \(f(y; \theta_0)\) that \(y\) is less than \(y^0\). Thus the Bayes limits here are in fact just confidence limits. Laplace certainly found his calculations appealing but he did not at his time have available the relevant confidence theory, developed much later by Fisher.

![Figure 5: The Bayes calculation with the Laplace noninformative prior can with symmetry duplicate the frequency calculation, thus giving the confidence result.](image-url)
Multiple meanings for \( p \)-value

We now have three somewhat different interpretations for “\( p \)-value” or “level of significance”: (i) The likelihood view: The statistical position of the observed data with respect to a \( \theta_0 \) value being tested; (ii) The decision theory view: The conventional level at which the data is just significant with respect to a \( \theta_0 \) value being tested; and (iii) The Bayes view: the Bayes survivor calculation at \( \theta_0 \) value using some prior distribution for \( \theta \). Gelman and Loken focus their discussion on the second or decision theoretic interpretation and address serious consequences from this approach, emphasizing the case of multiple hypotheses with a particular data set.

Threats from multiple meanings

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 are misleading.

Gelman and Loken express this concern for treating \( p \)-values from the 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. Much of contemporary statistics also largely overlooks such concerns. Statistics texts quite widely record the “accept-reject” approach, perhaps with cautions, but then leave the overview in the language of decision.

There is a long literature warning about this, however, and much earlier than Ioannidis (2005). Sterling (1959) discusses “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.” Indeed Gelman and Loken have stronger words: “It would take a highly unscrupulous researcher to perform test after test in a search for statistical significance … at the (0.05 level say and achieve publication)” . Rozeboom (1960) warns of the “The fallacy of the null-hypothesis sig-
nificance test” and quotes a famous philosophical epigram that the “Accept-Reject” is the “glory of science and the scandal of philosophy” meaning here the glory of statistics and the scandal of logic and application.

And then with “big data”

Certainly the risks of using arbitrary $p$-values to define ‘significance’, and to use these as decisions is very serious when multiple hypotheses lead to large numbers of $p$-values, as indicated by Gelman and Loken (2014). But just contemplate the potential for misleading results if all this is scaled up to Big Data. With ‘Big Data’ an anomaly can be anywhere, of any form, at any time, and could be something of significance or could be a total illusion. The calculation of the dimensionality for possible hypotheses, possible anomalies, and possible misleading decisions is challenging. And attributing significance or decision to selected millions of these suggests serious rethinking of the exploration process, the evaluation process, and the decision process. The risks for misleading decisions seem endless; we could have mega $p$-values, mega decisions and mega wrong ‘answers’.

Gelman and Loken (2014) do recommend some strategies: pre-registration; authentic replication; analysis of ‘all data”; an awareness “that $p$-values should not necessarily be taken at face value”. This last we would disagree with! The $p$-value is simply recording a statistical position of data relative to hypothesis; it is elemental and provides an appropriate starting point. The conventional but unwarranted attribution of decision, and the use of $p$-values for journal management, are at the heart of the problem, and neither has received the appropriate attention of the statistical community.

McNutt (2014) reports on an effort to help journal editors address the acceptance procedure for research reports. Adjusting the cut off levels from 5% or 1% would just be altering the rules and not correcting the problem; the recommendations, fortunately, go far beyond this.

The downstream effects are serious. Science has addressed some of the risks (McNutt, 2014). Perhaps Statistics should stand up for its responsibilities before a Big Data Disaster.

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.

References


