International Journal of Epidemiology 2003;**32**:699–702
DOI: 10.1093/ije/dyg294

Commentary: The *P*-value, devalued

Steven Goodman

Of the many honorifics bestowed on the articles in this historical series, it is doubtful that any have had applied the best—*funny*. The rhetorical zest and smiling outrage that Joseph Berkson brings to his puncturing of the quasi-religious precepts of traditional statistics in his classic article[1] recalls for me a public debate I witnessed in the 1980s between a highly respected statistician and a surgeon clinical-trialist. It was a debate on issues related to the adjustment of *P*-values in clinical trials, and what I remember best was the entrance of the physician in full surgical regalia; green operating scrubs, face mask, shoe covers, the whole bit. Playing effectively the role of the 'aw-shucks, I'm just a country doc who don't know nuthin' 'bout statistics' he parodied traditional statistical precepts so effectively, contrasting them unfavourably with common-sense judgements, that the statistician, however meritorious his rebuttal may have been, was left sputtering, helplessly pounding the lectern.

So it seems with this commentary, which asks in an innocent yet seemingly unanswerable way, 'If the population [of people] is not human, what *is* it?' This is the leading edge of an attack on Fisher's *P*-value which should still be required reading for all students of epidemiology and biostatistics today. The commentary shows us several things. First, it demonstrates just how old are some current criticisms, often presented as enlightened insights from a modern era. His first sentence has almost a nostalgic quality that looks surprising over 60 years later, 'There was a time when we did not talk about tests of significance; we simply did them.' These words described the future as much as the pre-1942 past.

Second, although it may not be immediately obvious, the argument presented here is closely related to ones that underlie modern recommendations to use CI and even Bayesian methods in lieu of *P*-values in biomedical research. Third, Berkson makes important distinctions between hypothesis testing and significance tests that continue to be ignored today. Fourth, and perhaps most subtly, he brings in a notion of 'evidence', a positive, relative concept that is critical to have on the table as separate and distinct from the *P*-value. And finally, he provides modern statisticians with a model for how to communicate technical concepts to applied users in an accessible and lively way.

All that said, it must be admitted that Berkson's critique is frustratingly incomplete. While he offers a scathing critique of the *P*-value, and shows us how standard interpretations contravene scientific intuition (grounded mainly in appeals to common sense) he does not offer a real alternative. He does call for more research, particularly into the meaning of what he calls 'middle *P*'s'. It is in this gap that I will spend most of my time in this commentary; linking his insights with the 'further research' that indeed occurred over the succeeding 60 years.

Department of Oncology, Division of Biostatistics, Johns Hopkins School of Medicine, 550 N. Broadway, Suite 1103, Baltimore, MD 21205, USA. E-mail: sgoodman@jhmi.edu

I must start by laying my own cards on the table. I look at most things statistical through a Bayesian/likelihood lens, a lens shaped by the writings of Jeffreys,[2] Good,[3] Savage,[4,5] Birnbaum,[6] Hacking,[7] Cornfield,[8,9] Edwards,[10] Berger,[11] and Royall.[12,13] But during the day I live and breathe *P*-values, for much the same reason that inhabitants of London continued breathing during the Great Fog of 1952; although it produced high morbidity and mortality, the alternative was even less attractive. Such is the predicament of those who try to banish *P*-values from epidemiological research, as the founding editor of the journal *Epidemiology* discovered; *P*-values are in epidemiologists' statistical air, and cannot be totally eliminated from the *corpus epidemiologicum* without unacceptable consequences.[14]

Let us start by putting Berkson's argument into context. The problem that he addresses, in multiple forms, is that the *P*-value is defined relative only to the null hypothesis and contains no information about an alternative. Berkson was certainly not the first one to try to remedy this; Jerzy Neyman and Egon Pearson introduced the notion of an alternative hypothesis and the associated 'power' concept with their hypothesis test procedure in 1933.[15] But they were concerned with creating procedures with certain long-run properties, not with measuring evidence.

Fisher objected to the hypothesis test procedure on several grounds. First, there was its automatic, decision-making aspect, which has attracted endless consternation and commentary over the years, much of it fuelled by Fisher's rhetorical heat.[16] Fisher's claim was that the scientific process was about learning, not decision-making, and that *P*-values could assist the learning process by serving as a continuous measure of evidence. This is why Berkson's commentary attacking the latter notion is significant. Fisher's second objection was Neyman and Pearson's introduction of the idea of a hypothesis that was an 'alternative' to the null. Fisher derided this because there is an infinite number of such alternatives, and he felt that to specify one or a subset was subjective. He claimed that this property made it literally impossible to calculate power objectively, and his goal was to create only objective methods:

> The frequency of the first class [errors of Type I], relative to the frequency with which the [null] hypothesis is true, is calculable, and therefore controllable simply from the specification of the null hypothesis. The frequency of the second kind [of error] must depend not only on the frequency with which rival hypotheses are in fact true, but also greatly on how closely they resemble the null hypothesis. Such errors are therefore incalculable…. [17]

While Fisher's objections to choosing one of many alternatives as the one to test against are understandable, the problem is that without an alternative hypothesis, the logic of significance testing is incomplete. Berkson points out why. If one is in the 'rejection' business, one cannot reject a null hypothesis without

something else to accept. And if one is in the 'evidence' business, you cannot have evidence *against* a hypothesis without it being *for* another, unless the hypothesis renders the observation literally impossible.[13,18]

This issue can be framed another way. The claim that a measure of evidence requires an alternative hypothesis is equivalent to saying that the magnitude of an observed effect is relevant to the evidence against the null hypothesis. That is, for the same *P*-value, a larger effect should provide more evidence against the null than a smaller effect. This motivates the example Berkson presents in his Table 1, when he looks at a hypothetical experiment to test whether a physician could divine the sex of a child *in utero*. In a small experiment (n = 10), the physician guesses right 60% of the time. In a larger experiment (N = 1000), the physician guesses right 50.5% of the time. The *P*-value in both cases equals 0.38. Berkson claims the small experiment is essentially uninformative, and the larger experiment confirms that the physician cannot discriminate 'to any significant degree', and is 'convincing positive evidence of the truth of the null hypothesis within practical limits' (ref. 1, p. 332).

It is worth spending some time on this simple example. To modern eyes it may seem curious that Berkson does not present what a first-year statistics student would today; confidence limits. Neyman's paper on confidence limits (derived from the logic of hypothesis tests) was published 8 years before this talk, and Berkson was clearly aware of it, referring to them in his footnote 5. His failure to calculate them may be because he did not see how they addressed the issue of evidence measurement. But we will start there. The exact 95% binomial CI for 6/10 successes is 26% to 89%. For 505/1000 successes it is 47.4% to 53.6%. These provide quantitative correlates to Berkson's phrases 'significant degree' and 'practical limits'. He is essentially saying that a 3.6% absolute increment in discriminating ability is of minimal practical value. This is the same as an argument that we are rarely interested in a precise null hypothesis; that there is always some 'equivalence' region around a zero effect. Berkson does not specify what that equivalence region is, except to claim implicitly that it includes a 3.6% increment in predictive ability.

Fisher probably would have agreed that in a situation like this, estimation is the preferred approach. In the following passage, which followed the previous quote, he claims that when the alternative is along a quantifiable continuum, estimation is more appropriate than testing, and that since any hypothesis can be specified as a 'null' hypothesis, Type II errors can be viewed as a form of Type I error, albeit for a non-zero null hypothesis:

> It may be added that in the theory of estimation we consider a continuum of hypotheses each eligible as null hypothesis, and it is the aggregate of frequencies calculated from each possibility in turn as true—including frequencies of error, therefore only of the 'first kind' without any assumptions of knowledge *a priori*—which supply the likelihood function … and other indications of the amount of information available. The introduction … to errors of the second kind in such arguments is entirely formal and ineffectual.[17]

There is an interesting split here; Fisher clearly has the technical skills to have proposed a measure of comparative evidence if he wanted to, but he claims to be able to meet all his scientific needs with an informal combination of the *P*-value and estimation. Berkson, on the other hand, was perhaps in better touch with how *P*-values were viewed and (mis)used in the real world. Berkson saw the need for a measure of evidence that combined effect magnitude and precision, but it appears that he could not quite figure out exactly how to do it. He says repeatedly, in different ways, that the probability of the observed data under the null hypothesis needs to be compared to the probability under an alternative hypothesis. Yet he does not propose formally the statistic that embodies this, one that was already established as the mathematical foundation of the hypothesis-testing calculus, and for which Fisher could have easily adapted his own likelihood function; the likelihood ratio (LR).

It is instructive to look at this example through the Likelihood/Bayesian prism. The LR is defined as the ratio of the data's probability under two hypotheses. To emphasize its characteristic of measuring evidence *for* a hypothesis (relative to another), and minimize confusion with the *P*-value, I will calculate it with the alternative hypothesis in the numerator:

$$\frac{\text{Probability of the data under the alternative hypothesis}}{\text{Probability of the data under the null hypothesis}}$$

$$= \frac{\Pr(D|\text{Ha})}{\Pr(D|\text{Ho})}$$

The first question the LR forces us to answer is what alternative hypothesis to use. As Fisher pointed out, this is not a simple question, and what we call 'null' and 'alternative' can be arbitrary. However, recognizing that a question about evidence requires it to be answered, and that the evidence reported will differ according to that answer, is a critical step forward.

The null hypothesis in this example is p = 50%, i.e. that the physician's guesses are no better than chance. For didactic and computational purposes, I will change the data in the N = 1000 scenario slightly from 505 successes to 510, which will not affect the point. The LR for various non-null hypotheses versus the null in the two scenarios is reported in the Table.

The Table tells us quantitatively what Berkson tells us qualitatively. We see that for the experiment with N = 10 and six successes, the data is virtually uninformative about any hypothesis in the range 50% to 60%, with all the LR very close to one. The best-supported hypothesis, with a very weak LR of 1.22, is p = 60%, the hypothesis that the true success probability equals the observed proportion of 60%. (This hypothesis must have the biggest likelihood, since the observed proportion is the maximum likelihood estimate.) This same LR (i.e. 1.22) is

Table Likelihood ratios for the listed alternative hypotheses versus the null hypothesis, Ho: p = 50%. These probabilities are the chance that a physician can guess the sex of a child *in utero*. See Appendix for calculations. (Note that the 'p' refers to the probability of a correct prediction, not the *P*-value)

		Alternative hypothesis		
Observed data	**95% CI**	p = 51%	p = 55%	p = 60%
$\hat{p}$ = 60%, N = 10	26% to 89%	LR = 1.04	1.16	1.22
$\hat{p}$ = 51%, N = 1000	48% to 54%	1.22	0.05	8×10^{-8}

obtained in the N = 1000 experiment for a success proportion of 51%, the maximally supported alternative when there are 510/1000 successes. The evidence against the null hypothesis in the two experiments looks the same, but only when we measure it for different alternative hypotheses; Ha: p = 60% in the N = 10 case, and Ha: p = 51% in the N = 1000 case.

If we measure the evidence for the same alternative hypothesis in the two experiments, the evidence is totally different; the N = 10 experiment favours Ha: p = 60% 1.22 times more strongly than Ho, but the N = 1000 experiment speaks overwhelmingly *against* Ha: p = 60% and *for* Ho by a factor exceeding 100 million! This vindicates Berkson's intuition that when considering the evidence *against* the null hypothesis, it is critical to understand what the evidence is *for*.

Because the *P*-value is not comparative, it is not an evidential measure and does not behave like one. But it is often a monotonic function of the maximum LR, the LR of the hypothesis with maximum likelihood (i.e. the MLE) versus the null. For example, when the statistic of interest is Gaussian, the maximum LR equals $\exp(Z^2/2)$ and the *P*-value value equals $2\,(1 - \Phi(|Z|))$ (i.e. twice the tail area), where Z = standard Z-score and $\Phi(|Z|)$ = the cumulative normal distribution from $-\infty$ to $|Z|$. This monotonic relationship tells us that the fixed sample-size *P*-value is a transformation of a measure (the LR) that compares the likelihood of the null hypothesis to a *post-hoc,* data-suggested hypothesis, i.e. the MLE. Thus, a fixed sample-size *P*-value might be regarded as a surrogate for a kind of evidential measure, but it is a measure that violates a prime dictum of scientific research: to pre-specify hypotheses. A true measure of evidence uses a pre-specified alternative not dictated by the data.

Operationally, how do we pre-specify alternative hypotheses? The easiest way is to choose a single non-null parameter value, e.g. the 'minimum important difference'. But the alternative hypothesis typically encompasses more than one non-null parameter value, e.g. 'Ha = treatment difference greater than zero'. In those situations, measuring the likelihood ratio for Ha versus Ho requires that we average the likelihood function over all values of the treatment difference included in the alternative hypothesis. The resulting LR is called a weighted likelihood ratio. If a Bayesian prior probability distribution is used as the averaging function, the resulting LR is called the 'Bayes Factor'.[19,20,21] The reason for the 'Bayes factor' name is that this ratio appears in Bayes theorem as the factor that multiplies the prior odds of the truth of two hypotheses to generate their posterior odds of being true.

Interestingly, if we use a pre-specified likelihood averaging function, many issues that pose a problem for conventional statistics vanish. In particular, the problem of 'multiple looks' disappears, as the probability of a Type I error is bounded by $1/LR_t$, where LR_t = the threshold degree of evidence to stop a study.[13,22] The problem of measuring evidence for a composite alternative hypothesis (i.e. one that encompasses more than one parameter value) is therefore analogous to that associated with exploration of multiple subgroups. If we go on a data-dredging expedition and report only which subgroup effects are significant, we are certain to produce a welter of false findings. But if we pre-specify which subgroups will be analysed, the chance of a spurious claim is greatly reduced. Similarly, if we pre-specify a weight function over a range of simple alternatives, and then measure the evidence comparatively, we can look at the data as frequently as we want without fear of an unbounded Type I error. If we find such pre-specification difficult, then any resulting problems should not be viewed as caused by our exploration or our evidential measure, but rather by our weak background knowledge.[20]

Berkson also brings to our attention that more complex features of the data other than just a simple effect measure can affect the alternative we consider. In the *Drosophila* example, he points out that by itself, the low *P*-value for the test of linearity was meaningless in the absence of a competing explanation for the observed pattern. Interestingly, Fisher seized on this in his rebuttal to Berkson's article, reprinted here, to claim that Berkson had ignored the example's biological nuances.[23] But a close reading of Berkson's piece shows that Fisher did not fairly represent him; Berkson does not dismiss the non-linearity, but rather states that the meaning of the low *P*-value depended on exactly which explanation for the non-linearity was entertained, and it was only such explanations that would justify rejection of the null hypothesis, not the small *P*-value alone. He and Fisher are in agreement about the importance of searching for those explanations; they differ only on the interpretation of the *P*-value in the absence of one.

Using the relationship between the fixed sample size *P*-value and the maximum LR, an idiosyncratic view is that the *P*-value is a kind of evidential measure, but an extraordinarily fickle one; flitting from one alternative hypothesis to another wherever the data wind blows. Unfortunately, in their traditional guise, *P*-values obscure the idea that true evidence is relative, and that it is literally nonsensical to speak of evidence against the null hypothesis without considering what it is *for*. Even without a modern statistical perspective or tools, Berkson saw this clearly, and communicated it in an eminently readable and entertaining way. The *International Journal of Epidemiology* is to be applauded for bringing this article again to our collective attention, and I hope that those who see it 60 years hence will not still be lamenting its unlearned lessons.

Appendix

Calculations underlying the Table.

$$\mathrm{LR}(\mathrm{Ha}: p = p_a \text{ vs. } \mathrm{Ho}: p = 0.5 \mid N, x) \frac{\Pr(Data \mid \mathrm{Ha})}{\Pr(Data \mid \mathrm{Ho})} = \frac{\mathrm{Bin}(N, x, p = p_a)}{\mathrm{Bin}(N, x, p = 0.5)}$$

Where Bin(N, x, p) = Binomial probability for sample size = N, successes = x and success probability = p.

For the first entry in the Table (LR = 1.04), these parameters are: N = 10, x = 6, $p_a = 0.51$

$$LR(\mathrm{Ha}: p = 0.51 \text{ vs. } \mathrm{Ho}: p = 0.50 \mid N = 10, x = 6) = \frac{\binom{10}{6}(0.51)^6(0.49)^4}{\binom{10}{6}(0.5)^6(0.5)^4} = 1.04$$

LR for the other entries are calculated similarly, using the reported values of N, x, and p_a.

References

[1] Berkson J. Tests of significance considered as evidence. *J Am Statist Assoc* 1942;**37:**325–35.

[2] Jeffreys H. *Theory of Probability.* Oxford: Oxford University Press, 1939, 1961.

[3] Good I. *Probability and the Weighing of Evidence*. New York: Charles Griffin & Co., 1950.

[4] Savage L. *The Foundations of Statistical Inference: A Discussion*. New York: Wiley, 1962.

[5] Savage L. On rereading RA Fisher. *American Statistician*. 1976;**4:**441–500.

[6] Birnbaum A. On the Foundations of Statistical Inference (with discussion). *J Am Statist Assoc* 1962;**53:**259–326.

[7] Hacking I. *The Logic of Statistical Inference*. Cambridge: Cambridge University Press, 1965.

[8] Cornfield J. A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *J Am Statist Assoc* 1966;**61:**577–94.

[9] Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *American Statistician* 1966;**20:**18–23.

[10] Edwards A. *Likelihood*. Cambridge: Cambridge University Press, 1972.

[11] Berger J, Sellke T. Testing a point null hypothesis: the irreconcilability of p-values and evidence. *J Am Statist Assoc* 1987;**82:**112–22.

[12] Royall R. The effect of sample size on the meaning of significance tests. *American Statistician* 1986;**40:**313–15.

[13] Royall R. *Statistical Evidence: A Likelihood Paradigm*. London: Chapman and Hall, 1997 Monographs on Statistics and Applied Probability, #71.

[14] Lang J, Rothman K, Cann C. That confounded P-value. *Epidemiology* 1998;**9:**7–8.

[15] Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Phil Trans Roy Soc A* 1933;**231:**289–337.

[16] Fisher R. *Statistical Methods and Scientific Inference*. *3rd Edn*. New York: Macmillan, 1973.

[17] Fisher RA. Statistical methods and scientific induction *J R Statist Soc* 1955;**17:**69–78.

[18] Goodman SN, Royall R. Evidence and scientific research. *Am J Public Health* 1988;**78:**1568–74.

[19] Kass R, Raftery A. Bayes Factors. *J Am Statist Assoc* 1995;**90:**773–95.

[20] Goodman SN. Towards evidence-based medical statistics, II: The Bayes Factor. *Ann Intern Med* 1999;**130:**1005–13.

[21] Greenland S. Probability logic and probabilistic induction. *Epidemiology* 1998;**9:**322–32.

[22] Kadane J, Schervish MJ, Seidenfeld T. *Rethinking the Foundations of Statistics*. Cambridge, UK: Cambridge University Press; 1999. (Cambridge Studies in probability, induction and decision theory.)

[23] Fisher RA. Note on Dr. Berkson's Criticism of Tests of Significance. *J Am Statist Assoc* 1943;**38:**103–04.