Of *P*-Values and Bayes: A Modest Proposal

Steven N. Goodman

I am delighted to be invited to comment on the use of *P*-values, but at the same time, it depresses me. Why? So much brainpower, ink, and passion have been expended on this subject for so long, yet *plus ca change, plus c'ést le meme chose* – the more things change, the more they stay the same. The references on this topic encompass innumerable disciplines, going back almost to the moment that *P*-values were introduced (by R.A. Fisher in the 1920s). The introduction of hypothesis testing in 1933 precipitated more intense engagement, caused by the subsuming of Fisher's "significance test" into the hypothesis test machinery.[1–9] The discussion has continued ever since. I have been foolish enough to think I could whistle into this hurricane and be heard.[10–12] But we (and I) still use *P*-values. And when a journal like EPIDEMIOLOGY takes a principled stand against them,[13] epidemiologists who may recognize the limitations of *P*-values still feel as if they are being forced to walk on one leg.[14]

So why do those of us who criticize the use of *P*-values bother to continue doing so? Isn't the "real world" telling us something – that we are wrong, that the effort is quixotic, or that this is too trivial an issue for epidemiologists to spend time on? Admittedly, this is not the most pressing methodologic issue facing epidemiologists. Still, I will try to argue that the topic is worthy of serious consideration.

Let me begin with an observation. When epidemiologists informally communicate their results (in talks, meeting presentations, or policy discussions), the balance between biology, methodology, data, and context is often appropriate. There is an emphasis on presenting a coherent epidemiologic or pathophysiologic "story," with comparatively little talk of statistical "rejection" or other related tomfoolery. But this same sensibility is often not reflected in published papers. Here, the structure of presentation is more rigid, and statistical summaries seem to have more power. Within these confines, the narrative flow becomes secondary to the distillation of complex data, and inferences seem to flow from the data almost automatically. It is this automaticity of inference that is most distressing, and for which the elimination of *P*-values has been attempted as a curative.

Although I applaud the motivation of attempts to eliminate *P*-values, they have failed in the past and I predict that they will continue to fail. This is because they treat the symptoms and not the underlying mindset, which must be our target. We must change how we think about science itself.

I and others have discussed the connections between statistics and scientific philosophy elsewhere,[11,12,15–22] so I will cut to the chase here. The root cause of our problem is a philosophy of scientific inference that is supported by the statistical methodology in dominant use. This philosophy might best be described as a form of "naïve inductivism,"[23] a belief that all scientists seeing the same data should come to the same conclusions. By implication, anyone who draws a different conclusion must be doing so for nonscientific reasons. It takes as given the statistical models we impose on data, and treats the estimated parameters of such models as direct mirrors of reality rather than as highly filtered and potentially distorted views. It is a belief that scientific reasoning requires little more than statistical model fitting, or in our case, reporting odds ratios, *P*-values and the like, to arrive at the truth.

How is this philosophy manifest in research reports? One merely has to look at their organization. Traditionally, the findings of a paper are stated at the beginning of the discussion section. It is as if the finding is something derived directly from the results section. Reasoning and external facts come afterward, if at all. That is, in essence, naïve inductivism. This view of the scientific enterprise is aided and abetted by the *P*-value in a variety of ways, some obvious, some subtle. The obvious way is in its role in the reject/accept hypothesis test machinery. The more subtle way is in the fact that the *P*-value is a probability – something absolute, with nothing external needed for its interpretation.

Now let us imagine another world – a world in which we use an inferential index that does not tell us where we stand, but how much distance we have covered. Imagine a number that does not tell us what we know, but how much we have learned. Such a number could lead us to think very differently about the role of data in making inferences, and in turn lead us to write about our data in a profoundly different manner.

This is not an imaginary world; such a number exists. It is called the Bayes factor.[15,17,25] It is the data compo-

Department of Oncology, Division of Biostatistics, Johns Hopkins School of Medicine, Baltimore, MD.

Address correspondence to: Seven Goodman, Department of Oncology, Division of Biostatistics, Johns Hopkins School of Medicine, 550 N. Broadway, Suite 1103, Baltimore, MD 21205.

Submitted and accepted January 19, 2001.

TABLE 1 Bayesian Interpretations of *P*-Values

P-value (*Z-score*)	Minimum Bayes factor	-e p ln(p)	Decrease in probability of the null hypothesis.. . . From 50%, to no less than	From 75%, to no less than	Strength of evidence
0.10 (*1.64*)	0.26	0.6	21%	44%	Weak
0.05 (*1.96*)	0.15	0.4	13%	31%	Moderate
0.03 (*2.17*)	0.1	0.3	9%	22%	Moderate
0.01 (*2.58*)	0.04	0.13	3.5%	10%	Moderate to strong
0.001 (*3.28*)	0.005	0.02	0.5%	1%	Strong to very strong

nent of Bayes Theorem. The odds we put on the null hypothesis (relative to others) using data external to a study is called the "prior odds," and the odds after seeing the data is the "posterior odds." The Bayes factor tells us how far apart those odds are, *ie*, the degree to which the data from a study move us from our initial position. It is quite literally an epistemic odds ratio, the ratio of posterior to prior odds, although it is calculable from the data, without those odds. It is the ratio of the data's probability under two competing hypotheses.[15,17]

If we have a Bayes factor equal to 1/10 for the null hypothesis relative to the alternative hypothesis, it means that these study results have decreased the relative odds of the null hypothesis by 10-fold. For example, if the initial odds of the null were 1 (*ie*, a probability of 50%), then the odds after the study would be 1/10 (a probability of 9%). Suppose that the probability of the null hypothesis is high to begin with (as they typically are in data dredging settings), say an odds of 9 (90%). Then a 10-fold decrease would change the odds of the null hypothesis to 9/10 (a probability of 47%), still quite probable. The Bayes factor is a measure of evidence in the same way evidence is viewed in a legal setting, or informally by scientists. Evidence moves us in the direction of greater or lesser doubt, but except in extreme cases it does not dictate guilt or innocence, truth or falsity.

I should warn readers knowledgeable in Bayesian methods to stop here. They may be severely disappointed (or even horrified) by the proposal I am about to make. I suggest that the Bayes factor does not necessarily have to be derived from a standard Bayesian analysis, although I would prefer that it were. As a simple alternative, it is possible instead to use the minimum Bayes factor (for the null hypothesis).[26] The appeal of the minimum Bayes factor is that it is calculated from the same information that goes into the *P*-value, and can easily be derived from standard analytic results, as described below. Quantitatively, it is only a small step from the *P*-value (and shares the liability of confounding the effect size with its precision). But conceptually, it is a huge leap. I recommend it not as a cure-all, but as a practical first step toward methodologic sanity.

The calculation goes like this. If a statistical test is based on a Gaussian approximation (as they are in many epidemiologic analyses), the strongest Bayes factor against the null hypothesis is $\exp(-Z^2/2)$, where Z is the number of standard errors from the null value. Thus it can be applied to most regression coefficients (whose significance is typically based on some form of normal approximation) and contingency tables. (When the *t*-statistic is used, it can substitute for Z.) If the log-likelihood of a model is reported, the minimum Bayes factor is simply the exponential of the difference between the log-likelihoods of two competing models (*ie*, the ratio of their maximum likelihoods). This likelihood-ratio (the minimum Bayes factor) is the basis for most frequentist analyses. While it is invariably converted into a *P*-value, it has inferential meaning without such conversion.

The minimum Bayes factor described above does not involve a prior probability distribution over non-null hypotheses; it is a global minimum for all prior distributions. However, there is also a simple formula for the minimum Bayes factor in the situation where the prior probability distribution is symmetric and descending around the null value. This is $-e\ p\ \ln(p)$,[27,28] where p is the fixed-sample size *P*-value. The table shows the correspondence between *P*-values, Z- (or *t*-) scores, and the two forms of minimum Bayes factors described above. Note that even the strongest evidence against the null hypothesis does not lower its odds as much as the *P*-value magnitude might lead people to believe. More importantly, the minimum Bayes factor makes it clear that we cannot estimate the credibility of the null hypothesis without considering evidence outside the study.

This translation from *P*-value to minimum Bayes factor is not merely a recalibration of our evidential measure, like converting from Fahrenheit to Celsius. By assessing the result with a minimum Bayes factor, we bring into play a different conceptual framework, which requires us to separate statistical results from inductive inferences. Reading from Table 1, a *P*-value of 0.01 represents a "weight of evidence" for the null hypothesis of somewhere between 1/25 (0.04)) and 1/8 (0.13). In other words, the relative odds of the null hypothesis *vs* any alternative are at most 8–25 times lower than they were before the study. If I am going to make a claim that a null effect is highly unlikely (*eg*, less than 5%), it follows that I should have evidence outside the study that the prior probability of the null was no greater than 60%. If the relationship being studied is far-fetched (*eg*, the probability of the null was greater than 60%), the evidence may still be too weak to make a strong knowledge claim. Conversely, even weak evidence in support of a highly plausible relationship may be enough for an author to make a convincing case.[15,17]

The use of the Bayes factor could give us a different view of results and discussion sections. In the results section, both the data and model-based data summaries are presented. (The choice of a mathematical model can be regarded as an inferential step, but I will not explore that here.) This can be followed by an index like the Bayes factor if two hypotheses are to be contrasted. The discussion section should then serve as a bridge between these indices and the conclusions. The components of

this bridge are the plausibility of the proposed mechanisms, (drawing on laboratory, other experimental evidence and patterns within this data), other empirical results related to this hypothesis and the qualitative strength of the current study's design and execution.

P-values need not be banned, although I would be happy to see them go. (When I see them, I translate them into approximate Bayes factors.) But we should certainly ban inferential reasoning based on the naïve use of *P*-values and hypothesis tests, and their various partners in crime, *eg*, stepwise regression (which chooses regression terms based exclusively on statistical significance, widely recognized as egregiously biased and misleading).[29,30] Even without formal Bayesian analysis, the use of minimum Bayes factors (along with, or in lieu of, *P*-values) might provide an antidote for the worst inferential misdeeds. More broadly, we should incorporate a Bayesian framework into our writing, and not just our speaking. We should describe our data as one source of information among many that make a relationship either plausible or unlikely. The use of summaries such as the Bayes factor encourages that, while use of the *P*-value makes it nearly impossible.

Changing the *P*-value culture is just a beginning. We utilize powerful tools to organize data and to guess at the reality which gave rise to them. We need to remember that these tools can create their own virtual reality.[17,30,31] The object of our study must be nature itself, not artifacts of the tools we use to probe its secrets. If we approach our data with respect for their complexity, with humility about our ability to sort that out, and with detailed knowledge of the phenomena under study, we will serve our science and the public health well. From that perspective, whether or not we use *P*-values seems, well, insignificant.

References

1. Berkson J. Tests of significance considered as evidence. J Am Stat Assoc 1942;37:325–335.
2. Fisher R. Statistical Methods and Scientific Inference. 3rd ed. New York: Macmillan, 1973.
3. Nunnally J. The place of statistics in psychology. Educ Psychol Meas 1960; 20(4):641–650.
4. Morrison D, Henkel R. The Significance Test Controversy: A Reader. Chicago: Aldine Publishing, 1970.
5. Buchanan-Wollaston H. The philosophic basis of statistical analysis. J Int Council Explor Sea 1935;10:249–263.
6. Rothman K. Significance questing. Ann Int Med 1986;105:445–447.
7. Rozeboom W. The fallacy of the null hypothesis significance test. Psychol Bull 1960;57(5):416–428.
8. Pearson E. Some thoughts on statistical inference. Ann Math Stat 1962;33: 394–403.
9. Cohen J. The earth is round ($p < .05$). Am Psychol 1994;49:997–1003.
10. Goodman SN, Royall R. Evidence and scientific research. Am J Public Health 1988;78:1568–1574.
11. Goodman SN. P-values, hypothesis tests and likelihood: implications for epidemiology of a neglected historical debate (with commentary and response). Am J Epidemiol 1993;137:485–496.
12. Goodman SN. Toward evidence-based medical statistics I. The P value fallacy. Ann Intern Med 1999;130:995–1004.
13. Rothman K. Writing for Epidemiology. Epidemiology 1998;9:333–337.
14. Lang J, Rothman K, Cann C. That Confounded P-value. Epidemiology 1998;9:7–8.
15. Goodman SN. Towards evidence-based medical statistics. II. The Bayes Factor. Ann Intern Med 1999;130:1005–1013.
16. Poole C. Beyond the confidence interval. Am J Public Health 1987;77:195–199.
17. Greenland S. Probability logic and probabilistic induction [see comments]. Epidemiology 1998;9:322–332.
18. Lindley D. The philosophy of statistics (with discussion). The Statistician 2000;49:293–337.
19. Howson C, Urbach P. Scientific Reasoning: The Bayesian Approach. 2nd ed. La Salle, IL: Open Court, 1993.
20. Rothman K, Greenland S. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott-Raven, 1998.
21. Oakes M. Statistical Inference. Chestnut Hill, MA: Epidemiology Resources Inc, 1990.
22. Greenland S. Summarization, smoothing, and inference in epidemiologic analysis. Scand J Soc Med 1993;21:227–232.
23. Chalmers A. What is this thing called science? 3rd ed. Indianapolis: Hackett, 1999.
24. Deleted in proof.
25. Kass R, Raftery A. Bayes factors. JASA 1995;90:773–795.
26. Edwards W, Lindman H, Savage L. Bayesian statistical inference for psychological research. Psychol Rev 1963;70:193–242.
27. Bayarri MJ, Berger J. Quantifying Surprise in the Data and Model Verification. In: Bernado, et al, eds. Bayesian Statistics. Oxford, Oxford University Press, 1998; 53-82.
28. Berger JO, Sellke T. Testing a point null hypothesis: The irreconcilability of *P*-values and evidence. J Am Stat Assoc 1987;82:112-122.
29. Harrell F, Lee K, Mark D. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–387.
30. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989;79:340–349.
31. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 1986;123:392–402.