Living with *P* Values

Resurrecting a Bayesian Perspective on Frequentist Statistics

Sander Greenland[a,b] *and Charles Poole*[c]

Abstract: In response to the widespread abuse and misinterpretation of significance tests of null hypotheses, some editors and authors have strongly discouraged *P* values. However, null *P* values still thrive in most journals and are routinely misinterpreted as probabilities of a "chance finding" or of the null, when they are no such thing. This misuse may be lessened by recognizing correct Bayesian interpretations. For example, under weak priors, 95% confidence intervals approximate 95% posterior probability intervals, one-sided *P* values approximate directional posterior probabilities, and point estimates approximate posterior medians. Furthermore, under certain conditions, a one-sided *P* value for a prior median provides an approximate lower bound on the posterior probability that the point estimate is on the wrong side of that median. More generally, *P* values can be incorporated into a modern analysis framework that emphasizes measurement of fit, distance, and posterior probability in place of "statistical significance" and accept/reject decisions.

(*Epidemiology* 2013;24: 62–68)

P values for null hypotheses of no association or no effect (null *P* values, which we denote by P_0) still thrive in most of the literature. Despite being data frequencies under a hypothetical sampling model, they are routinely misinterpreted as probabilities of a "chance finding" or of the null, which are Bayesian probabilities. To quote one biostatistics textbook: "The *P* value is the probability of being wrong when asserting that a true difference exists."[1] Similar misinterpretations remain in guides for researchers. For example, Cordova[2] states the null *P* value "represents the likelihood that groups differ after a treatment due to chance."

It is thus unsurprising that many call for improved education regarding *P* values.[3–11] Others discourage specific uses of *P* values[12–15] or, occasionally, all uses (as in the EPIDEMIOLOGY instructions for authors: "We strongly discourage the use of *P* values and language referring to statistical significance . . ."), regarding them as a confounded mix of estimate size and precision.[16,17]

Many advocates of educational reform specifically decry use of *P* values in Neyman-Pearson hypothesis testing, in which results are placed in bins labeled "association" and "no association" based on whether the *P* value is below or above a prespecified alpha level (the maximum tolerable type-I or false-positive error probability), usually 0.05. Neyman-Pearson testing has often been criticized as groundless and arbitrary[9,17–20] and even mindless[3,5]—aptly in our view. Unfortunately, Neyman-Pearson testing has also been called "significance testing," and the alpha level has been called the "significance level" of the test. This secondary terminology has fed the oft-lamented confusion of Neyman-Pearson testing with the significance-testing approach of Fisher.[21,22] In the latter approach, significance level refers to the *P* value itself, which is treated as a continuous measure of evidence against the tested hypothesis, with lower *P* values corresponding to greater evidence.[6,9,23–26]

This article is in the reform tradition. We first consider some arguments against *P* values. We then review ways in which *P* values and other ordinary frequentist statistics can be given correct Bayesian interpretations. We focus on how a one-sided *P* value can provide a lower bound on the posterior probability that the point estimate is on the wrong side of the prior median (the prior odds that the true parameter is above vs. below this median is 50:50).[27] In particular, when one sees a two-sided null *P* value P_0 for a difference or log ratio measure, one can immediately say that if one starts with a normal or uniform prior centered on the null, the posterior probability that the point estimate is in the wrong direction (ie, on the wrong side of the null) is no less than $P_0/2$.

ARE *P* VALUES SUPERFLUOUS GIVEN CONFIDENCE INTERVALS?

It is sometimes said that confidence intervals render *P* values redundant because the two give "equivalent information," with no indication of what "equivalent" means. True, for any parameter value, its *P* value is under 0.05 if and only if the parameter value is outside the 95% confidence interval computed by the same method. But this equivalence argument seems to restore 0.05-level testing to the primacy it retains in most medical journals.[28]

Submitted 2 May 2012; accepted 9 October 2012.
From the [a]Department of Epidemiology and [b]Department of Statistics, University of California, Los Angeles, CA; and the [c]Department of Epidemiology, University of North Carolina, Chapel Hill, NC.
The authors report no conflicts of interest.
Editors' note: *Related articles appear on pages 69 and 73.*
Correspondence: Sander Greenland, 22333 Swenson Drive, Topanga, CA. E-mail: lesdomes@ucla.edu.

ISSN: 1044-3983/12/2401–0062
DOI: 10.1097/EDE.0b013e3182785741

One can compute a *P* value from a confidence interval; therefore, the abstract (mathematical) information in a *P* value is indeed redundant given the confidence interval. But, unless the *P* value is 1, we can compute a confidence interval from the point estimate and *P* value, rendering the confidence interval redundant in the same technical sense. Furthermore, just as the estimate and *P* value do not convey a confidence interval well, a confidence interval does not convey desired *P* values well.[10] For instance, we have found colleagues who believe that if a set of risk ratios (*RR*s) all have lower confidence limits above 1, the *RR* with the highest lower limit must have the smallest null *P* value, P_0, especially if that *RR* has the largest point estimate. Others seem to think that if the lower limits are the same, the null *P* values are the same, or the one with the higher point estimate has the lower null *P* value. These beliefs are often the opposite of reality; for example, if the estimate $\widehat{RR}$ is 1.6 with lower limit 1.2, P_0 is 0.001, whereas for $\widehat{RR} = 4.8$ with lower limit 1.3, P_0 is 0.020.

Equivalence and redundancy are thus not good arguments for suppressing *P* values as supplements to point and interval estimates. There are, however, good psychological arguments for that suppression. They are based on the observation that null *P* values are almost always interpreted badly, even by statisticians. For example, a small P_0 is routinely taken as refuting the null and a large P_0 is routinely taken as supporting the null; however, both interpretations are usually far from correct.[8,11–13,25,26,28–37] We will explain how these problems could be avoided if *P* values were interpreted as distance measures and posterior probabilities rather than as tests of hypotheses.

Assumptions and Notation

Throughout, we assume that an analysis goal is inference about the true value θ_t of a parameter or some simple function of θ_t (eg, its antilog) in an analysis using standard models and approximations. Usually, θ_t is a coefficient in a linear, logistic, Poisson, or proportional hazards regression and analysis and thus represents a difference or log ratio; however, when is a product-term coefficient, it measures variation (heterogeneity) in a difference or log ratio. $\hat{\theta}$ will denote one of the usual efficient estimators of θ_t (eg, least-squares, maximum likelihood), with $\hat{\sigma}$ its estimated standard error. θ will denote the parameter name or a particular value for the parameter, possibly incorrect, but of special contextual interest. In almost all software and reports of *P* values, the null value 0 for θ is automatically taken to be of key interest, but other values may be as or more important to analyze. For example, in certain legal situations, a risk or rate ratio *RR* of 2 translates to a 50% lower bound for a causation probability,[38] which leads to examining $\theta = \ln(RR) = \ln(2)$.

CORRECT FREQUENCY INTERPRETATIONS OF *P* VALUES AND CONFIDENCE INTERVALS

We illustrate concepts using familiar Wald methods derived from the *Z* score $\left(\hat{\theta} - \theta\right)/\hat{\sigma}$ and its absolute value $|\hat{\theta} - \theta|/\hat{\sigma}$ which is the distance in standard errors from the observation $\hat{\theta}$ to the possibility θ. The two-sided *P* value P_θ for θ is, then, the relative frequency (over study repetitions) that the absolute *Z* statistic $\left|\hat{\theta} - \theta\right|/\hat{\sigma}$ would exceed its observed value if indeed $\theta_t = \theta$ and all other assumptions used to compute this frequency are correct. In particular, P_0 is the usual two-sided null *P* value for the hypothesis $\theta_t = 0$, derived from $\left|\hat{\theta}\right|/\hat{\sigma}$, and $P_{\ln(2)}$ is the two-sided *P* value for $\theta_t = \ln(2)$, derived from $\left|\hat{\theta} - \ln(2)\right|/\hat{\sigma}$. The usual 95% confidence interval with limits $\hat{\theta} \mp 1.96\hat{\sigma}$ can be derived by solving $\left|\hat{\theta} - \theta\right|/\hat{\sigma} = 1.96$ for θ and thus can be defined as all θ for which $P_\theta \geq 0.05$. Conversely, given a 95% confidence interval confidence interval $(\underline{\theta}, \overline{\theta})$, we can deduce $\hat{\theta} = \left(\overline{\theta} + \underline{\theta}\right)/2$, $\hat{\sigma} = \left(\overline{\theta} - \underline{\theta}\right)/3.92$ and thus compute a *Z* score and *P* value for any θ of interest.

P values such as P_θ are often described as measures of goodness of fit, distance, consistency, or compatibility between the observed data and the data-generating model, or between a parameter estimate $\hat{\theta}$ and a hypothesized parameter constraint such as $\theta_t = \theta$ or $\theta_t \leq \theta$.[23,39] P_θ is the probability transform of the distance from θ to $\hat{\theta}$; in particular, P_0 is the transformed distance of $\hat{\theta}$ to the null. A small P_θ is taken to indicate a problem with the assumptions used for its computation (eg, perhaps $\theta_t \neq \theta$, or there is some uncontrolled validity problem, or both).[21,22] This interpretation is popular among those seeking to avoid both hypothesis testing and Bayesian interpretations of *P* values.[26]

To illustrate, consider a disease indicator $\mathrm{Y} = 1,0$ and a logistic regression model for disease frequency given two binary factors X and Z,

$$\Pr(\mathrm{Y} = 1 \mid \mathrm{X} = \mathrm{x}, \mathrm{Z} = \mathrm{z}) = \mathrm{expit}(\alpha + \beta\mathrm{x} + \gamma\mathrm{z}) \qquad (1)$$

This model asserts that only three parameters (α, β, γ) are needed to perfectly specify (or encode) the disease frequencies at every combination of $\mathrm{X} = 1,0$ and $\mathrm{Z} = 1,0$. There is rarely any justification for this assumption; however, it is routine and usually unmentioned, or else unquestioned if the *P* value for the test of model fit is "big enough" (usually meaning at least 0.05 or 0.10).

Consider the null *P* value P_0 for $\theta_t = 0$ when θ is the product-term coefficient in

$$\Pr(\mathrm{Y} = 1 \mid \mathrm{X} = \mathrm{x}, \mathrm{Z} = \mathrm{z}) = \mathrm{expit}(\alpha + \beta\mathrm{x} + \gamma\mathrm{z} + \theta\mathrm{xz}) \qquad (2)$$

$\theta_t = 0$ yields model 1 and translates into constancy of the odds ratio relating X to Y as Z varies. P_0 is a measure of

(a). The distance from 0 to $\hat{\theta}$,
(b). The goodness of fit of model 1 when taking the more general model 2 as the referent, and
(c). The distance from the fitted model 1 to the fitted model 2 (where the fitted models are the above equations with the parameters replaced by their estimates).

For comparison, the non-null value $\theta_t = \ln(2)$ translates into a doubling of the odds ratio relating X to Y when we move from $\mathrm{Z} = 0$ to $\mathrm{Z} = 1$ and corresponds to the model:

$$\Pr(Y = 1 \mid X = x, Z = z) = \text{expit}(\alpha + \beta x + \gamma z + \ln(2)xz) \quad (3)$$

$P_{\ln(2)}$ is a measure of

(a). The distance from ln(2) to $\hat{\theta}$,
(b). The goodness of fit of model 3 when taking model 2 as the referent, and
(c). The distance from the fitted model 3 to the fitted model 2.

Finally, the 95% confidence interval for θ_t is a measure of the spread of the four-dimensional likelihood function derived from the data and model 2 along the θ dimension. Consequently, confidence interval width is sometimes promoted as a measure of study precision in estimating θ_t.[12] This interpretation, however, assumes model 2; using other models, we would expect all the confidence intervals and *P* values to change, demonstrating that statistical concepts of precision, fit, and distance are relative to a model rather than absolute properties of a study.

CORRECT BAYESIAN INTERPRETATIONS OF *P* VALUES AND CONFIDENCE INTERVALS

Typical misinterpretations of *P* values treat them as Bayesian posterior probabilities. For example, one common but extreme mistake interprets P_θ as the probability that $\theta_t = \theta$; however, these two probabilities are usually far apart.[29,30] Such misinterpretations may be recognized and avoided by examining situations in which P_θ is indeed a probability (bet) about the true value θ_t. For a review of basic Bayesian ideas such as prior and posterior probabilities, see for example, Greenland.[40–42]

Extreme-Prior Interpretation

Under maximum-likelihood analyses, the most direct Bayesian interpretations of *P* values and confidence intervals arise when only two extreme kinds of prior distributions are allowed: point priors, which express 100% certainty that a parameter is a given value; and equal-odds (uniform) priors on the usual normalizing ("natural parameter") scale for setting confidence intervals (eg, typically the log scale for ratio parameters, and the coefficient scale in multiplicative models). If τ denotes the prior standard deviation for θ, point priors correspond to $\tau = 0$ and equal-odds priors correspond to $\tau = \infty$.

A parameter θ with a point prior asserting it must be zero is usually inapparent (implicit) in the model because it is replaced in the model by zero and so disappears from sight. Model 1 is an example in which θ is invisible, yet the model is a special case of model 2 with the added constraint $\theta_t = 0$, which in Bayesian terms corresponds to imposing the point prior $\Pr(\theta_t = 0) = 1$ for θ and placing equal odds on all possible combinations of α, β, and γ. Similarly, model 3 corresponds to imposing the point prior $\Pr\{\theta_t = \ln(2)\} = 1$ and placing equal odds on all possible combinations of α, β, and γ, but θ now remains visible in the model because it is nonzero.

If instead we assume only the more general model 2 and place equal prior odds on all combinations of α, β, γ, and θ, we obtain the following interpretations of frequentist statistics as approximate Bayesian posterior statistics:

(a). P_θ is the posterior probability that $\hat{\theta}$ is closer to θ than to the truth θ_t (ie, P_θ is the probability that $|\hat{\theta} - \theta_t| > |\hat{\theta} - \theta|$).
(b). $P_\theta / 2$ is the posterior probability that $\hat{\theta}$ is on the wrong side of θ relative to the truth θ_t; in particular, $P_0/2$ is the probability that the observed association is in the wrong direction. Hence, if $\hat{\theta} > \theta$, $P_\theta / 2$ is the posterior probability that $\theta_t < \theta$; if $\hat{\theta} < \theta$, $P_\theta / 2$ is the probability that $\theta_t > \theta$; in particular, if $\hat{\theta}$ is positive, $P_0 / 2$ is the probability that θ_t is negative; if $\hat{\theta}$ is negative, $P_0 / 2$ is the probability that θ_t is positive.
(c). The 95% confidence interval $(\underline{\theta}, \overline{\theta})$ for θ becomes a 95% posterior probability interval for θ; hence, under model 2 and the prior, the probability that $\underline{\theta} \le \theta_t \le \overline{\theta}$ is 0.95 (parallel interpretations extend to other confidence levels).
(d). $\hat{\theta}$ is the posterior median (the odds of θ_t being above vs. below $\hat{\theta}$ are equal).

Thus, the equal-odds prior renders correct the usual misinterpretation of confidence intervals and provides Bayesian interpretations of *P* values.

As an example, suppose $\hat{\theta} = 1.40$ and $\hat{\sigma} = 0.60$. Then, the following Bayesian posterior probability statements follow from model 2, the data, and an equal-odds prior:

(i). $|\hat{\theta} - 0| / \hat{\sigma} = 1.40 / 0.60 = 2.33$, giving $P_0 = 0.02$ as the probability that 1.40 is closer to 0 than to θ_t and $P_0 / 2 = 0.01$ as the probability that θ_t is negative or that $e^{\theta_t} < 1$.
(ii). $|\hat{\theta} - \ln(2)| / \hat{\sigma} = |1.40 - 0.693| / 0.60 = 1.18$, giving $P_{\ln(2)} = 0.24$ as the probability that 1.40 is closer to $\ln(2)$ than to θ_t, and $P_{\ln(2)}/2 = 0.12$ as the probability that $\theta_t < \ln(2)$ or that $e^{\theta_t} < 2$.
(iii). $\hat{\theta} \mp 1.96\,\hat{\sigma} = 1.40 \mp 1.96(0.60) = 0.22,\ 2.58$ has 95% probability of containing θ_t, so that the probability of $e^{0.22} = 1.25 < e^{\theta_t} < e^{2.58} = 13.2$ is 95%.
(iv). The odds of θ_t being above versus below 1.40 is 1, meaning equal probabilities that $e^{\theta_t} < e^{1.40} = 4.06$ and $e^{\theta_t} > 4.06$.

Taking $P_0 / 2$ as the posterior probability that the observed association is in the wrong direction is found among pre-Fisherian writers, including Gossett, the inventor of the *t* test,[4,34,43] and follows from Bayesian 2 × 2 table results published in 1877.[44,45] More generally, (a)–(c) follow from taking the posterior distribution of the model parameters as proportional to the likelihood function, approximating the marginal posterior for θ_t with a normal $(\hat{\theta}, \hat{\sigma})$ distribution (hence likelihood-ratio statistics will render a better approximation than *Z* scores).

Objections to Equal-Odds Priors and Responses

One objection to the equal-odds prior underlying (a)–(c) is that it is improper (it has infinite instead of 100% total probability), which leads to undefined prior probabilities even though prior odds are defined. Although some Bayesians defend direct use of improper priors,[46] one can instead replace the equal-odds prior by proper priors that are weak enough to leave properties (a)–(c) approximately correct in practice, as is done in "objective" and "reference" Bayesian methods.[47] We will call all these priors "weak priors," whether proper or improper.

A more serious objection is that weak priors are usually contextually absurd.[40,41] Suppose θ_t is a log death-rate ratio for an Food and Drug Administration–approved glioma chemotherapy versus radiation therapy. Then, an equal-odds prior says $\ln(10^{-100})$, 0, and $\ln(10^{100})$ are all equally credible values for θ_t. But $\ln(10^{-100})$ implies complete prevention and is disproved by any death with chemotherapy, whereas $\ln(10^{100})$ implies complete causation and is disproved by any survivor of chemotherapy.

One response to these criticisms is to take interpretations (a)–(c) as Bayesian measures of the information content of the data relative to the model, before prior information about the model parameters is added. This approach provides a contrast against Bayesian results derived from informative priors and is supported by noting that (a)–(c) are precisely the Bayesian results one obtains when no prior data are added to the actual data.[40,41] Reference-Bayes analyses produce similar interpretations in that they correspond to adding almost no data.

Bounding Results from More Informative Priors

Weak priors are most absurd when all the values of θ being debated (and thus all serious candidates for a prior median) are so close to the null that it is difficult to distinguish among them or distinguish them from the null.[40,41,48,49] Many, if not most, modern controversies (eg, long-term nutrient effects and drug side effects) involve debates within ranges like $¼ < RR < 4$. In these settings, confidence intervals (and thus posterior intervals under weak priors) typically cover a substantial part of the range under debate, and no one would take seriously claims of (say) $RR > 100$.

Even without controversy, the effort and discomfort of specifying detailed priors is daunting. Not only must we rationalize all choices to ourselves and the reader, but we may have to allow for the fact that readers may have different priors from ours. Anticipating the full spectrum of these differences with sensitivity analyses of various priors is even more demanding. Thus, in a world of sharply constrained time resources, demands for informative priors leads instead to avoidance of Bayesian analysis.

An alternative, however, is to offer bounds on posterior probabilities when the prior is restricted to a given class. These bounds show the range that a sensitivity analysis over that class would produce and thus may address concerns of those wary of detailed priors. In particular, suppose one would regard it reasonable to consider as a possible prior for θ_t a distribution that is symmetric around a single mode (maximum) at its median θ_m and weak for the remaining model parameters. This class includes the most common coefficient priors, including normal, *t*, logistic, and certain reference distributions with median θ_m, and implies that $\exp(\theta_t)$ has prior median $\exp(\theta_m)$. Of special interest, normal (Gaussian) priors arise when the prior information about θ_t is derived from a number of sources or studies, none of which contributes a dominant fraction of the information; this prior produces a lognormal prior for $\exp(\theta_t)$.

Let P_m denote the two-sided *P* value for the prior median θ_m. Then, $P_m / 2$ approximates the smallest possible posterior probability that the observed estimate $\hat{\theta}$ is on the wrong side of θ_t[27,50]; in addition, $\hat{\theta}$ is the furthest the posterior median could be from the prior median θ_m. These interpretations also apply using uniform priors on θ_t with median θ_m. Thus, $P_m / 2$ provides a statistic with a correct Bayesian interpretation, without demanding detailed specification of a prior or computation beyond $P_m / 2$. Again, suppose $\hat{\theta} = 1.40$ and $\hat{\sigma} = 0.60$, with weak priors for all parameters but θ. Then, $P_0 / 2 = 0.01$ is the smallest possible posterior probability that θ_t is negative when the θ_t prior is unimodal and symmetric (eg, normal) around 0; $P_{\ln(2)} / 2 = 0.12$ is the smallest possible posterior probability that $\theta_t < \ln(2)$ when the θ_t prior is unimodal and symmetric around $\ln(2)$; and 1.40 is the furthest the posterior median could be from any prior median.

Some Limitations

Within the above class of priors, the discrepancy between the lower bound $P_m / 2$ and the posterior probability increases as the spread of the prior decreases, albeit the discrepancy is not large until the prior becomes fairly informative relative to the data and model. To illustrate, suppose we have a normal prior with mean 0 and standard deviation τ for a log rate ratio $\theta_t = \ln(RR_t)$. This is a symmetric, unimodal prior with median 0 for θ_t and median 1 for $RR_t = \exp(\theta_t)$. The Table shows the resulting posterior probabilities of $RR_t < 1$ for various estimates $\widehat{RR} = \exp(\hat{\theta})$, standard errors $\hat{\sigma}$ for $\hat{\theta}$, and prior standard deviations τ for θ_t expressed in multiples of $\hat{\sigma}$. The final column uses an infinite τ, which gives back $P_0 / 2$.

The Table shows a general property of normal priors: in practical terms, when the prior mean is θ_m, $P_m / 2$ will not be far below the posterior probability that $\hat{\theta}$ is on the wrong side of θ_m if τ is over twice $\hat{\sigma}$ (so that the weight $1/\tau^2$ for the prior mean θ is less than a quarter the weight $1/\hat{\sigma}^2$ for the ordinary estimate $\hat{\theta}$). Of course, if one focuses on a particular normal prior, it can be easily used to calculate directly all desired posterior intervals and probabilities[40,41,51]; $P_m / 2$ then remains a reference point showing what weaker normal priors with the same mean could yield.

The bounding interpretation also assumes weak priors on the rest of the parameters. Such priors are usually unrealistic. Nonetheless, provided care is taken in specifying the

TABLE. Posterior Probabilities of $\theta_t < 0$, Given a Lognormal Ratio Estimate $\widehat{RR} = \exp(\hat{\theta})$ and Lognormal Prior Distribution on $\exp(\theta_t)$ with Median 1, Under Various Scenarios

	Prior Standard Deviation τ of Log Ratio, θ_t:			
$\widehat{RR}$	$\hat{\sigma}$	$2\hat{\sigma}$	$4\hat{\sigma}$	$\infty\hat{\sigma}$
When standard error $\hat{\sigma} = 0.5$ and τ =				
	0.5	1	2	∞
1	0.50	0.50	0.50	*0.50*
1.5	0.28	0.23	0.22	*0.21*
2	0.163	0.107	0.089	*0.083*
3	0.060	0.025	0.017	*0.014*
5	0.0114	0.0020	0.0009	*0.0006*
8	0.00164	0.00010	0.00003	*0.00002*
When standard error $\hat{\sigma} = 1$ and τ =				
	1	2	4	∞
1	0.50	0.50	0.50	*0.50*
1.5	0.39	0.36	0.35	*0.34*
2	0.31	0.27	0.25	*0.24*
3	0.22	0.16	0.14	*0.14*
5	0.127	0.075	0.059	*0.054*
8	0.071	0.031	0.022	*0.019*

When prior standard deviation τ is infinite (last column, in italics), these are $P_0/2$.

regression model (in particular, by centering the regressors[51]), posterior probabilities for the target parameter tend to be much less sensitive to reasonably imprecise priors on the remaining parameters than they are to the target (θ_t) prior. Regardless, stronger priors on all parameters require much more specification effort than most are willing to provide; but again, if provided, one can use them directly in Bayesian procedures,[51] and *P* values can serve as reference values for readers wary of more informative priors.

All the above approximations assume use of efficient methods such as maximum likelihood under standard asymptotics[52] and that the data distribution satisfies certain conditions met by ordinary epidemiologic models.[27] For less efficient methods (like inverse-probability weighting), the interpretations have to be qualified technically by stating they are approximate posterior probabilities if one is given only the estimating function, rather than the complete likelihood function.

P VALUES INTO BAYES FACTORS

There is a huge literature on Bayesian interpretation of *P* values, most of which demands nonlinear transforms and concepts unfamiliar in basic statistics. We describe the most common example,[29,30,33,34,52–55] although others exist.[56]

Your Bayes factor for a hypothesis is the ratio of your posterior odds on the hypothesis to your prior odds on the hypothesis; thus, it is how much your odds change (in multiplicative terms) in light of the data.[49,57,58] Suppose your prior odds on the null versus all other possibilities combined was 1 and your posterior odds on the null was ¼ (you would bet only 20% on it after seeing the data); then, your Bayes factor for the null would be ¼/1 = ¼. When the odds are comparing simple point hypotheses, such as $\theta_t = 0$ versus $\theta_t = \ln(2)$, the Bayes factor simplifies to the likelihood ratio.

Consider the class of "spike and smear" priors for which there is a point probability q that $\theta_t = \theta_m$ (ie, a spike or point mass of probability at θ_m of size q), and for the remaining possibilities $\theta_t \neq \theta_m$, the prior follows a symmetric unimodal distribution around θ_m. This makes $q/(1-q)$ the prior odds on $\theta_t = \theta$ versus $\theta_t \neq \theta$. Sellke et al[34] show that, when $P_m < 1/e = 0.37$, $-eP_m\ln(P_m)$ is an approximate lower bound on the Bayes factor; in particular, for $P_m = 0.10$, 0.05, and 0.01, lower bounds on the Bayes factor are 0.63, 0.41, and 0.12, respectively. They further show that for normal priors on $\theta_t \neq \theta_m$, these bounds are even higher: 0.70, 0.47, and 0.15. Thus, if P_0 = 0.05 a Bayesian analysis with $q = ½ = 50\%$ prior probability on the null and mean-zero normal otherwise leaves a posterior probability for the null of at least $0.47/(1+0.47) = 32\%$; this maximum drop from 50% to 32% probability for $\theta_t = 0$ is nowhere near as impressive as the usual "borderline significant" description of $P_0 = 0.05$. We can also find the highest value of q for which P_0 would deserve its common misinterpretation as the posterior probability of the null. For $P_0 = 0.05$, this value is a prior probability of $q = 10\%$ for $\theta_t = 0$.

Transforming P_0 into a Bayes factor thus reveals a serious weakness of conventional interpretations of null *P* values. Bayes factors have several disadvantages, however. They cannot be read directly off tabulated estimates or *P* values; they require the conceptual steps of thinking in terms of Bayes factors; and they do not provide any posterior probabilities unless we specify q, an arbitrary analysis constant that represents commitment of an appreciable fraction of prior probability to a single specific value of θ. In contrast, the interpretations given earlier use only confidence intervals and *P* values, with no need for Bayes factors or probability spikes.

There are serious objections to the use of probability spikes in population research because of their contextual meaning.[27,59] For example, a null spike represents an assertion that, with prior probability q, we have background data that prove $\theta_t = 0$ with absolute certainty; $q = ½$ thus represents a 50–50 bet that there is decisive information literally proving the null. Without such information (such as a point prediction by a highly plausible physical law), a probability spike at the null is an example of "spinning knowledge out of ignorance."[31] This prejudice favoring the null characterizes common misinterpretations of frequentist tests[37] and yet can lead to Bayesian conflicts with those tests.[60] In contrast, weak priors have no spike.

DISCUSSION

For more than 70 years, null *P* values have been the most common and yet most controversial inferential statistic,

primarily because of their use for hypothesis testing against fixed alpha (significance) levels such as 0.05. A point often lost in statistical training, however, is that *P* values can be divorced from decision rules, testing, arbitrary cutoffs, and null hypotheses. In ordinary epidemiologic analyses, two-sided *P* values can instead be used as probability measures (with both frequentist and Bayesian meaning) of the distance from (or fit of) the entire set assumptions (including model and validity assumptions and the explicit assumption of $\theta_t = \theta$) to the data. One-sided *P* values such as $P_0/2$ can be used to measure support for or probability of their one-sided hypotheses, given the remaining assumptions. Large *P* values are then as telling as small ones; for example, if $P_0/2$ is large, then, considered in isolation, the study seems vague about the direction of an association when viewed through the same model lens used to compute the estimate $\hat{\theta}$ and the confidence interval. Furthermore, if one uses an informative prior to derive the posterior probability of the point estimate being in the wrong direction, $P_0/2$ provides a reference point indicating how much the prior information influenced that posterior probability.

Confidence intervals provide valuable information about the precision implied by the data[12,17] given the assumed model; hence, we would certainly not wish to see them replaced by *P* values. However, confidence intervals are rarely precise about the relation of observations to prespecified parameter values, which in part explains why null *P* values can be viewed as defensible supplements to confidence intervals when the null has legitimate claim to our attention (as in exploratory and many highly controversial settings).[4,10] Furthermore, the frequency behavior and posterior interpretation of P_0 are more robust to model misspecification than are confidence intervals and non-null *P* values, which is worth noting given that we are always uncertain about the model form.[5] They are also more robust to validity problems that produce bias toward the null, especially certain simple types of measurement error or misclassification of the exposure or disease,[61] where, under a weak prior, $P_0/2$ remains the posterior probability that the estimate is in the wrong direction,[62] whereas confidence intervals become invalid.

Even when the null is an arbitrary hypothesis unworthy of special attention, null-related questions are encouraged by automatic software emission of P_0. Again, incorrect Bayesian interpretations of P_0 remain the norm, encouraged by some instructional sources.[1,2,63] It thus seems important to displace these misinterpretations with correct ones. With correct interpretation of P_0 in mind, extensions of *P* values to θ of greater contextual interest is simple, and extensions to more realistic priors can also be obtained by translating the priors to data and adding them to the conventional analysis.[41,51]

After one separates *P* values from arbitrary testing cutoffs such as 0.05 and goes beyond the null, one may see confidence intervals as problematic for encouraging degraded binary inferences.[28] Asking whether a value θ is in or out of the confidence interval turns the confidence interval into a 0.05-level test of that value. If we try to be more precise, we hit failings of intuition. For example, if the 95% confidence limits for *RR* are 1.8 and 3.3, what is the one-sided *P* value for $RR \leq 2$? Few would recognize quickly that it is $P_{\ln(2)}/2 = 0.10$, giving a 10% posterior probability that $RR < 2$ under a weak prior for all model parameters, and at least 10% if the prior for *RR* was more sharply specified as lognormal with median at 2.

Failure to appreciate the complementarity of confidence intervals and *P* values may stem more from historical accidents in how the two statistics entered common usage.[22] The chief disaster is that for generations Bayesian perspectives were banished from introductory statistics. Much like what happened with other attempts at knowledge suppression, users have filled the gap with interpretations that miss crucial subtleties and are often simply false. A common false interpretation is that P_0 is the probability that chance alone produced the observed association. Unfortunately, the premise that chance alone operated is logically the same as the null hypothesis ("chance alone" implies there is no effect and no bias), and thus, this misinterpretation is the same as claiming P_0 is the probability of the null.[11,37] Correct Bayesian interpretations can address this problem.

Given the extensive misinterpretations of frequentist statistics and the enormous (some would say impossible)[64] demands made by fully Bayesian analyses, a serious argument can be made for de-emphasizing (if not eliminating) inferential statistics in favor of more data presentation such as tables of raw numbers.[65] After all, the only purpose of a single study is to contribute its description and data to a broad pool of relevant information.[66] Without an inferential ban, however, an improvement of practice will require re-education, not restriction.

REFERENCES

1. Glantz SA. *Primer of Biostatistics*. 5th ed. New York, NY: McGraw-Hill; 2001.
2. Cordova ML. Giving clinicians more to work with: let's incorporate confidence intervals into our data. *J Athl Train*. 2007;42:445.
3. Ware JH, Mosteller F, Delgado F, Donnelly C, Ingelfinger JA. P values. In: Bailar JC, Mosteller F, eds. *Medical Uses of Statistics*. Boston, MA: NEJM Books; 1992:181–200.
4. Senn S. Two cheers for P-values? *J Epidemiol Biostat*. 2001;6:193–204; discussion 205.
5. Weinberg CR. It's time to rehabilitate the P-value. *Epidemiology*. 2001;12:288–290.
6. Sterne JA. Teaching hypothesis tests–time for significant change? *Stat Med*. 2002;21:985–994; discussion 995.
7. Whitley E, Ball J. Statistics review 3: hypothesis testing and P values. *Crit Care*. 2002;6:222–225.
8. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med*. 1999;130:995–1004.
9. Pocock SJ, Ware JH. Translating statistical findings into plain English. *Lancet*. 2009;373:1926–1928.
10. VanderWeele TJ. Re: "The ongoing tyranny of statistical significance testing in biomedical research." *Eur J Epidemiol*. 2010;25:843–844.
11. Greenland S, Poole C. Problems in common interpretations of statistics in scientific articles, expert reports, and testimony. *Jurimetrics* 2011;51:113–129.
12. Poole C. Low P-values or narrow confidence intervals: which are more durable? *Epidemiology*. 2001;12:291–294.

13. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol*. 2008;45:135–140.
14. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. *Eur J Epidemiol*. 2010;25:225–230.
15. Poole C, Kuss O, Stang A. On a use of the null P-value (letter). *Eur J Epidemiol* 2010;25:844–845, erratum 899–900.
16. Lang JM, Rothman KJ, Cann CI. That confounded P-value. *Epidemiology*. 1998;9:7–8.
17. Rothman KJ. Curbing type I and type II errors. *Eur J Epidemiol*. 2010;25:223–224.
18. Rothman KJ. A show of confidence. *N Engl J Med*. 1978;299:1362–1363.
19. Rothman KJ. Significance questing. *Ann Intern Med*. 1986;105:445–447.
20. Cohen J. The earth is round (p <.05). *Amer Psychol*. 1994;49:997–1003.
21. Fisher RA. *Statistical Methods for Research Workers*. 14th ed. New York, NY: Oxford; 1970.
22. Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol*. 1993; 137:485–496.
23. Cox DR. The role of significance tests (with discussion). *Scand J Stat*. 1977;4:49–70.
24. Cox DR. Statistical significance tests. *Br J Clin Pharmacol*. 1982;14: 325–331.
25. Poole C. Beyond the confidence interval. *Am J Public Health*. 1987; 77:195–199.
26. Rothman KJ, Greenland S, Lash TL. Precision and statistics in epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology.* 3rd ed. Philadelphia, PA: Lippincott-Wolters-Kluwer; 2008:148–167.
27. Casella G, Berger RL. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J Am Stat Assoc*. 1987;82:106–111.
28. Poole C. Confidence intervals exclude nothing. *Am J Public Health*. 1987;77:492–493.
29. Berger JO, Delampady M. Testing precise hypotheses (with discussion). *Stat Sci*. 1987;2:317–352.
30. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of p values and evidence (with discussion). *J Am Stat Assoc*. 1987;82:112–139.
31. Poole C. Causal values. *Epidemiology*. 2001;12:139–141.
32. Schervish M. P-values: what they are and what they are not. *Am Stat* 1996;50:203–206.
33. Goodman SN. Of P-values and Bayes: a modest proposal. *Epidemiology*. 2001;12:295–297.
34. Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. *Am Stat* 2001;55:62–71.
35. Senn S. A comment on replication, p-values and evidence, S.N.Goodman, Statistics in Medicine 1992; 11:875-879. *Stat Med*. 2002;21:2437–2444; author reply 2445.
36. Hubbard R, Lindsay RL. Why P values are not a useful measure of evidence in statistical significance testing. *Theory Psychol.* 2008;18:69–88.
37. Greenland S. Null misinterpretation in statistical testing and its impact on health risk assessment. *Prev Med*. 2011;53:225–228.
38. Greenland S, Robins JM. Epidemiology, justice, and the probability of causation. *Jurimetrics.* 2000; 40:321–340.
39. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated cystem of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*. 1900;50:157–175.
40. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol*. 2006;35:765–775.
41. Greenland S. Introduction to Bayesian statistics. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology.* 3rd ed. Philadelphia, PA: Lippincott-Wolters-Kluwer; 2008:328–344.
42. Held L. An introduction to Bayesian methods with applications in epidemiology. In: Ahrens W, Pigeot I, eds. *Handbook of Epidemiology*. New York, NY: Springer; 2013;in press.
43. Student [Gosset WS]. The probable error of a mean. *Biometrika.* 1908;VI:1–25.
44. Seneta E. Carl Liebermeister's hypergeometric tails. *Historia Mathematica.* 1994;21:453–462.
45. Nurminen M, Mutanen P. Exact Bayesian analysis of two proportions. *Scand J Statistics.* 1987;14:67–77.
46. Taraldsen G, Lindqvist BH. Improper priors are not improper. *Am Stat.* 2010;64:154–158.
47. Berger JO. The case for objective Bayesian analysis (with discussion). *Bayesian Anal.* 2006;1:385–472.
48. DeGroot MH. Comment on Shafer. *J Am Stat Assoc.* 1982;77:336–339.
49. Greenland S. Probability logic and probabilistic induction. *Epidemiology*. 1998;9:322–332.
50. Greenland S. Re: "P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate" (letter). *Am J Epidemiol*. 1994;140:116–117.
51. Greenland S. Bayesian perspectives for epidemiologic research. II. Regression analysis. *Int J Epidemiol* 2007;36:195–202.
52. Lindley DV. *Introduction to Probability and Statistics from a Bayesian Viewpoint; Part 2, Inference.* London, UK: Cambridge University Press; 1965.
53. Edwards W, Lindman H, Savage L. Bayesian statistical inference for psychological research. *Psychol Rev.* 1963;70:193–242.
54. Pratt JW. Bayesian interpretation of standard inference statements. *J Royal Stat Soc B*. 1965;27:169–203.
55. DeGroot MH. Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. *J Am Stat Assoc*. 1973;68:966–969.
56. Hodges J. Who knows what alternative lurks in the heart of significance tests? In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, eds. *Bayesian Statistics 4*. New York, NY: Oxford University Press; 1992: 247–266.
57. Good IJ. *Good Thinking*. Minneapolis, MN: U Minn Press; 1983: 140–143.
58. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med*. 1999;130:1005–1013.
59. Casella G, Berger RL. Comment. *Stat Sci*. 1987;2:344−417.
60. Lindley DV. A statistical paradox. *Biometrika.* 1957;44:187–192.
61. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C. *Measurement Error in Nonlinear Models*. Boca Raton, FL: Chapman and Hall; 2006.
62. Greenland S, Gustafson P. Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. *Am J Epidemiol*. 2006;164:63–68.
63. Lobo I. Genetics and statistical analysis. *Nature Education.* Available at: http://www.nature.com/scitable/topicpage/genetics-and-statistical-analysis-34592. Accessed 6 April 2012.
64. Senn SJ. You may believe you are a Bayesian, but you are probably wrong. *RMM.* 2011;2:48–66.
65. Greenland S, Gago-Dominguez M, Castelao JE. The value of risk-factor ("black-box") epidemiology. *Epidemiology*. 2004;15:529–535.
66. Poole C, Peters U, Il'yasova D, Arab L. Commentary: this study failed? *Int J Epidemiol*. 2003;32:534–535.