Toward Evidence-Based Medical Statistics. 2: The Bayes Factor

Steven N. Goodman, MD, PhD

Bayesian inference is usually presented as a method for determining how scientific belief should be modified by data. Although Bayesian methodology has been one of the most active areas of statistical development in the past 20 years, medical researchers have been reluctant to embrace what they perceive as a subjective approach to data analysis. It is little understood that Bayesian methods have a data-based core, which can be used as a calculus of evidence. This core is the Bayes factor, which in its simplest form is also called a *likelihood ratio*. The minimum Bayes factor is objective and can be used in lieu of the *P* value as a measure of the evidential strength. Unlike *P* values, Bayes factors have a sound theoretical foundation and an interpretation that allows their use in both inference and decision making. Bayes factors show that *P* values greatly overstate the evidence against the null hypothesis. Most important, Bayes factors require the addition of background knowledge to be transformed into inferences—probabilities that a given conclusion is right or wrong. They make the distinction clear between experimental evidence and inferential conclusions while providing a framework in which to combine prior with current evidence.

This paper is also available at http://www.acponline.org.

In the first of two articles on evidence-based statistics (1), I outlined the inherent difficulties of the standard frequentist statistical approach to inference: problems with using the *P* value as a measure of evidence, internal inconsistencies of the combined hypothesis test–*P* value method, and how that method inhibits combining experimental results with background information. Here, I explore, as nonmathematically as possible, the Bayesian approach to measuring evidence and combining information and epistemologic uncertainties that affect all statistical approaches to inference. Some of this presentation may be new to clinical researchers, but most of it is based on ideas that have existed at least since the 1920s and, to some extent, centuries earlier (2).

The Bayes Factor Alternative

Bayesian inference is often described as a method of showing how belief is altered by data. Because of this, many researchers regard it as nonscientific; that is, they want to know what the data say, not what our belief should be after observing them (3). Comments such as the following, which appeared in response to an article proposing a Bayesian analysis of the GUSTO (Global Utilization of Streptokinase and tPA for Occluded Coronary Arteries) trial (4), are typical.

> When modern Bayesians include a "prior probability distribution for the belief in the truth of a hypothesis," they are actually creating a metaphysical model of attitude change . . . The result . . . cannot be field-tested for its validity, other than that it "feels" reasonable to the consumer. . . .
>
> The real problem is that neither classical nor Bayesian methods are able to provide the kind of answers clinicians want. That classical methods are flawed is undeniable—I wish I had an alternative . . . . (5)

This comment reflects the widespread misperception that the only utility of the Bayesian approach is as a belief calculus. What is not appreciated is that Bayesian methods can instead be viewed as an evidential calculus. Bayes theorem has two components—one that summarizes the data and one that represents belief. Here, I focus on the component related to the data: the Bayes factor, which in its simplest form is also called a *likelihood ratio*. In Bayes theorem, the Bayes factor is the index through which the data speak, and it is separate from the purely subjective part of the equation. It has also been called the relative betting odds, and its logarithm is sometimes referred to as the *weight of the evidence* (6, 7). The distinction between evidence and error is clear when it is recognized that the Bayes factor (evidence) is a measure of how much the probability of truth (that is, $1 - \text{prob(error)}$, where *prob* is probability) is altered by the data. The equation is as follows:

$$\begin{array}{c} \textit{Prior Odds} \\ \textit{of Null Hypothesis} \end{array} \times \begin{array}{c} \textit{Bayes} \\ \textit{Factor} \end{array} = \begin{array}{c} \textit{Posterior Odds} \\ \textit{of Null Hypothesis} \end{array}$$

where Bayes factor =

$$\frac{\text{Prob}(\textit{Data, given the null hypothesis})}{\text{Prob}(\textit{Data, given the alternative hypothesis})}$$

The Bayes factor is a comparison of how well two hypotheses predict the data. The hypothesis that predicts the observed data better is the one that is said to have more evidence supporting it. Unlike the *P* value, the Bayes factor has a sound theoretical foundation and an interpretation that

See related article on pp 995-1004 and editorial comment on pp 1019-1021.

Table 1. Final (Posterior) Probability of the Null Hypothesis after Observing Various Bayes Factors, as a Function of the Prior Probability of the Null Hypothesis

Strength of Evidence	Bayes Factor	Decrease in Probability of the Null Hypothesis	
		From	To No Less Than
			%
Weak	1/5	90	64*
		50	17
		25	6
Moderate	1/10	90	47
		50	9
		25	3
Moderate to strong	1/20	90	31
		50	5
		25	2
Strong to very strong	1/100	90	8
		50	1
		25	0.3

* Calculations were performed as follows:
A probability (Prob) of 90% is equivalent to an odds of 9, calculated as Prob/(1 − Prob).
Posterior odds = Bayes factor × prior odds; thus, (1/5) × 9 = 1.8.
Probability = odds/(1 + odds); thus, 1.8/2.8 = 0.64.

allows it to be used in both inference and decision making. It links notions of objective probability, evidence, and subjective probability into a coherent package and is interpretable from all three perspectives. For example, if the Bayes factor for the null hypothesis compared with another hypothesis is 1/2, the meaning can be expressed in three ways.

1. *Objective probability:* The observed results are half as probable under the null hypothesis as they are under the alternative.

2. *Inductive evidence:* The evidence supports the null hypothesis half as strongly as it does the alternative.

3. *Subjective probability:* The odds of the null hypothesis relative to the alternative hypothesis after the experiment are half what they were before the experiment.

The Bayes factor differs in many ways from a *P* value. First, the Bayes factor is not a probability itself but a ratio of probabilities, and it can vary from zero to infinity. It requires two hypotheses, making it clear that for evidence to be *against* the null hypothesis, it must be *for* some alternative. Second, the Bayes factor depends on the probability of the observed data alone, not including unobserved "long run" results that are part of the *P* value calculation. Thus, factors unrelated to the data that affect the *P* value, such as why an experiment was stopped, do not affect the Bayes factor (8, 9).

Because we are so accustomed to thinking of "evidence" and the probability of "error" as synonymous, it may be difficult to know how to deal with a measure of evidence that is not a probability. It is helpful to think of it as analogous to the concept of

energy. We know that energy is real, but because it is not directly observable, we infer the meaning of a given amount from how much it heats water, lifts a weight, lights a city, or cools a house. We begin to understand what "a lot" and "a little" mean through its effects. So it is with the Bayes factor: It modifies prior probabilities, and after seeing how much Bayes factors of certain sizes change various prior probabilities, we begin to understand what represents strong evidence, and weak evidence.

Table 1 shows us how far various Bayes factors move prior probabilities, on the null hypothesis, of 90%, 50%, and 25%. These correspond, respectively, to high initial confidence in the null hypothesis, equivocal confidence, and moderate suspicion that the null hypothesis is not true. If one is highly convinced of no effect (90% prior probability of the null hypothesis) before starting the experiment, a Bayes factor of 1/10 will move one to being equivocal (47% probability on the null hypothesis), but if one is equivocal at the start (50% prior probability), that same amount of evidence will be moderately convincing that the null hypothesis is not true (9% posterior probability). A Bayes factor of 1/100 is strong enough to move one from being 90% sure of the null hypothesis to being only 8% sure.

As the strength of the evidence increases, the data are more able to convert a skeptic into a believer or a tentative suggestion into an accepted truth. This means that as the experimental evidence gets stronger, the amount of external evidence needed to support a scientific claim decreases. Conversely, when there is little outside evidence supporting a claim, much stronger experimental evidence is required for it to be credible. This phenomenon can be observed empirically, in the medical community's reluctance to accept the results of clinical trials that run counter to strong prior beliefs (10, 11).

Bayes Factors and Meta-Analysis

There are two dimensions to the "evidence-based" properties of Bayes factors. One is that they are a proper measure of quantitative evidence; this issue will be further explored shortly. The other is that they allow us to combine evidence from different experiments in a natural and intuitive way. To understand this, we must understand a little more of the theory underlying Bayes factors (12–14).

Every hypothesis under which the observed data are not impossible can be said to have some evidence for it. The strength of this evidence is proportional to the probability of the data under that hypothesis and is called the *likelihood* of the hypothesis. This use of the term "likelihood" must not be confused with its common language meaning of

probability (12, 13). Mathematical likelihoods have meaning only when compared to each other in the form of a ratio (hence, the *likelihood ratio*), a ratio that represents the comparative evidential support given to two hypotheses by the data. The likelihood ratio is the simplest form of Bayes factor.

The hypothesis with the most evidence for it has the maximum mathematical likelihood, which means that it predicts the observed data best. If we observe a 10% difference between the cure rates of two treatments, the hypothesis with the maximum likelihood would be that the true difference was 10%. In other words, whatever effect we are measuring, the best-supported hypothesis is always that the unknown true effect is equal to the observed effect. Even when a true difference of 10% gets more support than any other hypothesis, a 10% observed difference also gives a true difference of 15% some support, albeit less than the maximum (**Figure**).

This idea—that each experiment provides a certain amount of evidence for every underlying hypothesis—is what makes meta-analysis straightforward under the Bayesian paradigm, and conceptually different than under standard methods. One merely combines the evidence provided by each experiment for each hypothesis. With log Bayes factors (or log likelihoods), this evidence can simply be added up (15–17).

With standard methods, quantitative meta-analysis consists of taking a weighted average of the observed effects, with weights related to their precision. For example, if one experiment finds a 10% difference and another finds a 20% difference, we would average the numbers 10% and 20%, pool their standard errors, and calculate a new *P* value based on the average effect and pooled standard error. The cumulative evidence (*P* value) for the meta-analytic average has little relation to the *P* values for the individual effects, and averaging the numbers 10% and 20% obscures the fact that both experiments actually provide evidence for the *same* hypotheses, such as a true 15% difference. Although it might be noted that a 15% difference falls within the confidence intervals of both experiments, little can be done quantitatively or conceptually with that fact. So while meta-analysts say they are combining evidence from similar studies, standard methods do not have a measure of evidence that is directly combined.

Of Bayes Factors and *P* Values

If we are to move away from *P* values and toward Bayes factors, it is helpful to have an "exchange rate"—a relation between the new unit of measurement and the old. With a few assumptions, we can make this connection. First, to compare like
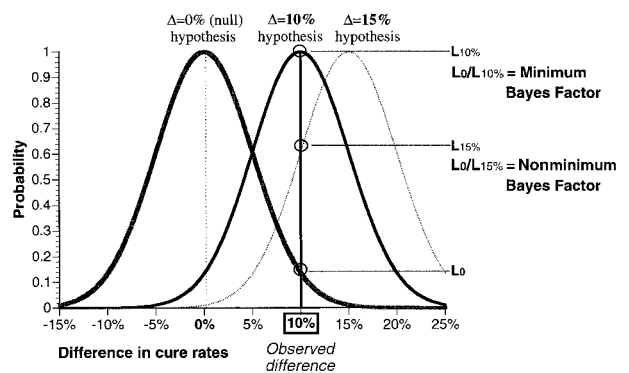

Figure. Calculation of a Bayes factor (likelihood ratio) for the null hypothesis versus two other hypotheses: the maximally supported alternative hypothesis (change $\Delta = 10\%$) and an alternative hypothesis with less than the maximum support ($\Delta = 15\%$). The likelihood of the null hypothesis (L_0) divided by the likelihood of the best supported hypothesis ($L_{10\%}$), is the minimum likelihood ratio or minimum Bayes factor, the strongest evidence against the null hypothesis. The corresponding ratio for the hypothesis $\Delta = 15\%$ results in a larger ratio, which means that the evidence against the null hypothesis is weaker.

with like, we must calculate the Bayes factor for the same hypothesis for which the *P* value is being calculated. The *P* value is always calculated by using the observed difference, so we must calculate the Bayes factor for the hypothesis that corresponds to the observed difference, which we showed earlier was the best-supported hypothesis. Second, because a smaller *P* value means less support for the null hypothesis (or more evidence against it), we must structure the Bayes factor the same way, so that a smaller Bayes factor also means less support for the null hypothesis. This means putting the likelihood of the null hypothesis in the numerator and the likelihood of an alternative hypothesis in the denominator. (Whether the null hypothesis likelihood is in the top or bottom of the ratio depends on the context of use.) If we put the evidence for the best-supported hypothesis in the denominator, the resulting ratio will be the smallest possible Bayes factor with respect to the null hypothesis. This reciprocal of the maximum likelihood ratio is also called the *standardized likelihood*. The minimum Bayes factor (or minimum likelihood ratio) is the smallest amount of evidence that can be claimed for the null hypothesis (or the strongest evidence against it) on the basis of the data. This is an excellent benchmark against which to compare the *P* value.

The simplest relation between *P* values and Bayes factors exists when statistical tests are based on a Gaussian approximation, which is the case for most statistical procedures found in medical journals. In that situation, the minimum Bayes factor (the minimum likelihood ratio) is calculated with the same numbers used to calculate a *P* value (13, 18, 19). The formula is as follows (see Appendix I for derivation):

$$\text{Minimum Bayes factor} = e^{-Z^2/2}$$

Table 2. Relation between Fixed Sample Size *P* Values and Minimum Bayes Factors and the Effect of Such Evidence on the Probability of the Null Hypothesis

P Value (Z Score)	Minimum Bayes Factor	Decrease in Probability of the Null Hypothesis, %		Strength of Evidence
		From	To No Less Than	
0.10 (1.64)	**0.26** (1/3.8)	75	44	Weak
		50	21	
		17	5	
0.05 (1.96)	**0.15** (1/6.8)	75	31	Moderate
		50	13	
		26	5	
0.03 (2.17)	**0.095** (1/11)	75	22	Moderate
		50	9	
		33	5	
0.01 (2.58)	**0.036** (1/28)	75	10	Moderate to strong
		50	3.5	
		60	5	
0.001 (3.28)	**0.005** (1/216)	75	1	Strong to very strong
		50	0.5	
		92	5	

where *z* is the number of standard errors from the null effect. This formula can also be used if a *t*-test (substituting *t* for *Z*) or a chi-square test (substituting the chi-square value for Z^2) is done. The data are treated as though they came from an experiment with a fixed sample size.

This formula allows us to establish an exchange rate between minimum Bayes factors and *P* values in the Gaussian case. **Table 2** shows the minimum Bayes factor and the standard *P* value for any given *Z* score. For example, when a result is 1.96 standard errors from its null value (that is, *P* = 0.05), the minimum Bayes factor is 0.15, meaning that the null hypothesis gets 15% as much support as the best-supported hypothesis. This is threefold higher than the *P* value of 0.05, indicating that the evidence against the null hypothesis is not nearly as strong as "*P* = 0.05" suggests.

Even when researchers describe results with a *P* value of 0.05 as being of borderline significance, the number "0.05" speaks louder than words, and most readers interpret such evidence as much stronger than it is. These calculations show that *P* values of 0.05 (corresponding to a minimum Bayes factor of 0.15) represent, at best, moderate evidence against the null hypothesis; those between 0.001 and 0.01 represent, at best, moderate to strong evidence; and those less than 0.001 represent strong to very strong evidence. When the *P* value becomes very small, the disparity between it and the minimum Bayes factor becomes negligible, confirming that strong evidence will look strong regardless of how it is measured.

The right-hand part of **Table 2** uses this relation between *P* values and Bayes factors to show the maximum effect that data with various *P* values would have on the plausibility of the null hypothesis. If one starts with a chance of no effect of 50%, a result with a minimum Bayes factor of 0.15 (corresponding to a *P* value of 0.05) can reduce confidence in the null hypothesis to no lower than 13%. The last row in each entry turns the calculation around, showing how low initial confidence in the null hypothesis must be to result in 5% confidence after seeing the data (that is, 95% confidence in a non-null effect). With a *P* value of 0.05 (Bayes factor ≥ 0.15), the prior probability of the null hypothesis must be 26% or less to allow one to conclude with 95% confidence that the null hypothesis is false. This calculation is not meant to sanctify the number "95%" in the Bayesian approach but rather to show what happens when similar benchmarks are used in the two approaches.

These tables show us what many researchers learn from experience and what statisticians have long known; that the weight of evidence against the null hypothesis is not nearly as strong as the magnitude of the *P* value suggests. This is the main reason that many Bayesian reanalyses of clinical trials conclude that the observed differences are not likely to be true (4, 20, 21). They conclude this not always because contradictory prior evidence outweighed the trial evidence but because the trial evidence, when measured properly, was not very strong in the first place. It also provides justification for the judgment of many experienced meta-analysts who have suggested that the threshold for significance in a meta-analysis should be a result more than two standard errors from the null effect rather than two (22, 23).

The theory underlying these ideas has a long history. Edwards (2) traces the concept of mathematical likelihood into the 18th century, although the name and full theoretical development of likelihood didn't occur until around 1920, as part of R.A. Fisher's theory of *maximum likelihood*. This was a frequentist theory, however, and Fisher did not acknowledge the value of using the likelihood directly for inference until many years later (24). Edwards (14) and Royall (13) have built on some of Fisher's ideas, exploring the use of likelihood-based measures of evidence outside of the Bayesian paradigm. In the Bayesian realm, Jeffreys (25) and Good (6) were among the first to develop the theory behind Bayes factors, with the most comprehensive recent summary being that of Kass (26). The suggestion that the minimum Bayes factor (or minimum likelihood ratio) could be used as a reportable index appeared in the biomedical literature at least as early as 1963 (19). The settings in which Bayes factors differ from likelihood ratios are discussed in the following section.

Bayes Factors for Composite Hypotheses

Bayes factors larger than the minimum values cited in the preceding section can be calculated (20, 25–27). This is a difficult technical area, but it is important to understand in at least a qualitative way what these nonminimum Bayes factors measure and how they differ from simple likelihood ratios.

The definition of the Bayes factor is the probability of the observed data under one hypothesis divided by its probability under another hypothesis. Typically, one hypothesis is the null hypothesis of no difference. The other hypothesis can be stated in many ways, such as "the cure rates differ by 15%." That is called a *simple hypothesis* because the difference (15%) is specified exactly. The null hypothesis and best-supported hypothesis are both simple hypotheses.

Things get more difficult when we state the alternative hypothesis the way it is usually posed: for example, "the true difference is not zero" or "the treatment is beneficial." This hypothesis is called a *composite hypothesis* because it is composed of many simple hypotheses ("The true difference is 1%, 2%, 3%... ,"). This introduces a problem when we want to calculate a Bayes factor, because it requires calculating the probability of those data under the hypothesis, "The true difference is 1%, 2%, 3%...." This is where Bayes factors differ from likelihood ratios; the latter are generally restricted to comparisons of simple hypotheses, but Bayes factors use the machinery of Bayes theorem to allow measurement of the evidence for composite hypotheses.

Bayes theorem for composite hypotheses involves calculating the probability of the data under each simple hypothesis separately (difference = 1%, difference = 2%, and so on) and then taking an average. In taking an average, we can weight the components in many ways. Bayes theorem tells us to use weights defined by a prior probability curve. A prior probability curve represents the plausibility of every possible underlying hypothesis, on the basis of evidence from sources other than the current study. But because prior probabilities can differ between individual persons, different Bayes factors can be calculated from the same data.

Different Questions, Different Answers

It may seem that the fact that the same data can produce different Bayes factors undermines the initial claim that Bayesian methods offer an objective way to measure evidence. But deeper examination shows that this fact is really a surrogate for the more general problem of how to draw scientific conclusions from the totality of evidence. Applying different weights to the hypotheses that make up a composite hypothesis does not mean that different answers are being produced for the same evidential question; it means that different questions are being asked. For example, in the extreme, if we put all of the weight on treatment differences near 5%, the question about evidence for a nonzero treatment difference becomes a question about evidence for a 5% treatment difference alone. An equal weighting of all hypotheses between 5% and 20% would provide the average evidence for a difference in that range, an answer that would differ from the average evidence for all hypotheses between 1% and 25%, even though all of these are nonzero differences.

Thus, the problem in defining a unique Bayes factor (and therefore a unique strength of evidence) is not with the Bayesian approach but with the fuzziness of the questions we ask. The question "How much evidence is there for a nonzero difference?" is too vague. A single nonzero difference does not exist. There are many nonzero differences, and our background knowledge is usually not detailed enough to uniquely specify their prior plausibility. In practical terms, this means that we usually do not know precisely how big a difference to expect if a treatment or intervention "works." We may have an educated guess, but this guess is typically diffuse and can differ among individuals on the basis of the different background information they bring to the problem or the different weight that they put on shared information. If we could come up with generally accepted reasons that justify a unique plausibility for each underlying truth, these reasons would constitute a form of explanation. Thus, the most fundamental of statistical questions—what is the strength of the evidence?—is related to the fundamental yet most uncertain of scientific questions—how do we explain what we observe?

This fundamental problem—how to interpret and learn from data in the face of gaps in our substantive knowledge—bedevils all technological approaches to the problem of quantitative reasoning. The approaches range from evasion of the problem by considering results in aggregate (as in hypothesis testing), solutions that leave background information unquantified (Fisher's idea for *P* values), or representation of external knowledge in an idealized and imperfect way (Bayesian methods).

Proposed Solutions

Acknowledging the need for a usable measure of evidence even when background knowledge is incomplete, Bayesian statisticians have proposed many approaches. Perhaps the simplest is to conduct a sensitivity analysis; that is, to report the Bayes factors produced by a range of prior distributions, representing the attitudes of enthusiasts to skeptics (28, 29). Another solution, closely related, is to report the smallest Bayes factor for a broad class of prior distributions (30), which can have a one-to-one re-

lation with the P value, just as the minimum Bayes factor does in the Gaussian case (31). Another approach is to use prior distributions that give roughly equal weight to each of the simple hypotheses that make up the composite hypothesis (25, 26, 32), allowing the data to speak with a minimal effect of a prior distribution. One such index, the Bayesian information criterion, for which Kass (26) makes a strong case, is closely related to the minimum Bayes factor, with a modification for the sample size. Finally, there is the approach outlined here: not to average at all, but to report the strongest Bayes factor against the null hypothesis.

Beyond the Null Hypothesis

Many statisticians and scientists have noted that testing a hypothesis of exact equivalence (the null hypothesis) is artificial because it is unlikely to be exactly true and because other scientific questions may be of more interest. The Bayesian approach gives us the flexibility to expand the scope of our questions to, for example, "What is the evidence that the treatment is harmful?" instead of "What is the evidence that the treatment has no effect?" These questions have different evidential answers because the question about harm includes all treatment differences that are not beneficial. This changes the null hypothesis from a simple hypothesis (a difference of 0) into a composite hypothesis (a difference of zero or less). When this is done, under certain conditions, the one-sided P value can reasonably approximate the Bayes factor (33, 34). That is, if we observe a one-sided P value of 0.03 for a treatment benefit and give all degrees of harm the same initial credibility as all degrees of benefit, the Bayes factor for treatment harm compared with benefit is approximately 0.03. The minimum Bayes factor for no treatment effect compared with benefit would still be 0.095 (**Table 2**).

Objectivity of the Minimum Bayes Factor

The minimum Bayes factor is a unique function of the data that is at least as objective as the P value. In fact, it is more objective because it is unaffected by the hypothetical long-run results that can make the P value uncertain. In the first article (1), I presented an example in which two different P values (0.11 and 0.03) were calculated from the same data by virtue of the different mental models of the long run held by two researchers. The minimum Bayes factor would be 0.23, identical for both scientists' approaches (Appendix 2). This shows us again how P values can overstate the evidence, but more important, it vindicates our intuition that the identical data should produce identical evidence.

This example is important in understanding two problems that plague frequentist inference: multiple comparisons and multiple looks, or, as they are more commonly called, *data dredging* and peeking at the data. The frequentist solution to both problems involves adjusting the P value for having looked at the data more than once or in multiple ways. But adjusting the measure of evidence because of considerations that have nothing to do with the data defies scientific sense (8, 35–41), belies the claim of "objectivity" that is often made for the P value, and produces an undesirable rigidity in standard trial design. From a Bayesian perspective, these problems and their solutions are viewed differently; they are caused not by the reason an experiment was stopped but by the uncertainty in our background knowledge. The practical result is that experimental design and analysis is far more flexible with Bayesian than with standard approaches (42).

External Evidence

Prior probability distributions, the Bayesian method for representing background knowledge, are sometimes derided as representing opinion, but ideally this opinion should be evidence-based. The body of evidence used can include almost all of the factors that are typically presented in a discussion section but are not often formally integrated with the quantitative results. It is not essential that an investigator know of all of this evidence before an experiment. This evidence can include the following:

1. The results of similar experiments.
2. Experiments studying associations with similar underlying mechanisms.
3. Laboratory experiments directly studying the mechanism of the purported association.
4. Phenomena seen in other experiments that would be explained by this proposed mechanism.
5. Patterns of intermediate or surrogate end points in the current experiment that are consistent with the proposed mechanism.
6. Clinical knowledge based on other patients with the same disease or on other interventions with the same proposed mechanism.

Only the first of these types of evidence involves a simple comparison or summation of results from similar experiments, as in a meta-analysis. All of the others involve some form of extrapolation based on causal reasoning. The use of Bayes factors makes it clear that this is necessary in order to draw conclusions from the statistical evidence.

Use of the Bayes Factor

We will now use two statements from the results sections of hypothetical reports to show the mini-

mum Bayes factor can be used to report and interpret evidence.

Hypothetical Statement 1

The difference in migraine relief rates between the experimental herbal remedy and placebo groups (54% compared with 40% [CI for difference, −2% to 30%]) was not significant ($P = 0.09$).

Bayesian data interpretation 1: The P value of 0.09 ($Z = 1.7$) for the difference in migraine relief rates corresponds to a minimum Bayes factor of $e^{-1.7^2/2} =$ ¼ for the null hypothesis. This means that these data reduce the odds of the null hypothesis by at most a factor of 4, fairly modest evidence for the efficacy of this treatment. For these data to produce a final null hypothesis probability of 5%, the external evidence supporting equivalence must justify a prior probability of equivalence less than 17%. But no mechanism has been proposed yet for this herbal migraine remedy, and all previous reports have consisted of case studies or anecdotal reports of relief. This a priori support is weak and does not justify a prior probability less than 50%. The evidence from this study is therefore insufficient for us to conclude that the proposed remedy is effective.

Bayesian data interpretation 2: . . . For these data to produce a final null hypothesis probability of 5%, the external evidence supporting equivalence must justify a prior probability of equivalence less than 17%. However, the active agent in this remedy is in the same class of drugs that have proven efficacy in migraine treatment, and this agent has been shown to have similar vasoactive effects both in animal models and in preclinical studies in humans. Three uncontrolled studies have all shown relief rates in the range seen here (50% to 60%), and the first small randomized trial of this agent showed a significant effect (60% compared with 32%; $P = 0.01$). The biological mechanism and observed empirical evidence seem to justify a prior probability of ineffectiveness of 15% to 25%, which this evidence is strong enough to reduce to 4% to 8%. Thus, the evidence in this trial, in conjunction with prior findings, is strong enough for us to conclude that this herbal agent is likely to be effective in relieving migraine.

Hypothetical Statement 2

Among the 50 outcomes examined for their relation with blood transfusions, only nasopharyngeal cancer had a significantly elevated rate (relative risk, 3.0; $P = 0.01$).

Bayesian data interpretation: The minimum Bayes factor for relative risk of 1.0 compared with a relative risk not equal to 1.0 for nasopharyngeal cancer is 0.036. This is strong enough to reduce a starting probability on the null hypothesis from at most 59% to 5%. However, there is no previous evidence for such an association or of a biological mechanism to explain it. In addition, rates of cancers with similar risk factor profiles and molecular mechanisms were not elevated, meaning that blood transfusion would have to produce its effect by means of a mechanism that differs from any other previously identified causes of this cancer. Previous studies of blood transfusions have not reported this association, and there have been no reports of increased incidence of nasopharyngeal cancer among populations who undergo repeated transfusions. Therefore, prior evidence suggests that the probability of the null hypothesis is substantially higher than 60%. A minimum Bayes factor of 0.036 means that this result can reduce a 85% prior probability to no lower than 17% and a 95% prior probability to no lower than 41%. Therefore, more evidence than that provided by this study is needed to justify a reliable conclusion that blood transfusion increases the risk for nasopharyngeal cancer. However, future studies should explore this relation and its potential mechanisms.

Discussion

The above examples do not nearly represent full Bayesian interpretation sections, which might use a range of prior distributions to define a range of Bayes factors, or use priors that have been elicited from experts (29, 43, 44). These scenarios do, however, illustrate a few essential facts. First, this measure of evidence can usually be easily calculated from the same information used to calculate a P value or confidence interval and thus can be implemented without specialized software or extensive statistical expertise. Some expertise is needed to assure that the Gaussian approximation underlying the formula applies in a particular situation. When it doesn't apply, many standard software programs report some function of the exact likelihood (typically, its logarithm), from which it is not hard for a statistician to calculate the minimum Bayes factor. Its independence from prior probabilities can also help overcome the reluctance of many investigators to abandon what they regard as objective statistical summaries.

More important, these examples highlight how this index can help keep the statistical evidence distinct from the conclusions, while being part of a calculus that formally links them. The first example showed how the same quantitative results could be included in discussions that came to different conclusions. The explicitness of this process encourages debate about the strength of the supporting evidence. As outlined in the first article, standard methods discourage this because they offer no way to combine supporting evidence with a study's P values or confidence intervals.

These examples demonstrate how the minimum

Bayes factor enables simple threshold Bayesian analyses to be performed without a formal elicitation of prior probability distributions. One merely has to argue that the prior probability of the null hypothesis is above or below a threshold value, on the basis of the evidence from outside the study. If the strongest evidence against the null hypothesis (the minimum Bayes factor) is not strong enough to sufficiently justify a conclusion, then the weaker evidence derived from a Bayes factor from a full Bayesian analysis will not be either.

The use of the minimum Bayes factor does not preclude a formal Bayesian analysis and indeed might be an entrée to one. Recent reviews and books outline how full Bayesian analyses can be conducted and reported (21, 29, 45–50). Bayesian results can also be extended into formal decision analyses (51). The availability of user-friendly software for Bayesian calculations (52) makes implementation of this method more practicable now than in the past.

In not using a proper Bayesian prior probability distribution, the minimum Bayes factor represents a compromise between Bayesian and frequentist perspectives, which can be criticized from both camps. Some statisticians might deride the minimum Bayes factor as nothing more than a relabelled P value. But as I have tried to show, P values and Bayes factors are far more than just numbers, and moving to Bayes factors of any kind frees us from the flawed conceptual framework and improper view of the scientific method that travels with the P value.

The Bottom Line: Both Perspectives Are Necessary, but *P* Values Are Not

Standard frequentist methods are most problematic when used to draw conclusions in a single experiment. Their denial of a formal role for external information in inference poses serious practical and logical problems. But Bayesian methods, designed for inductive inference in single experiments, do not guarantee that in the long run, conclusions in which we have 95% confidence will turn out to be true 95% of the time (53). This is because Bayesian prior probability distributions are not ideal quantitative descriptors of what we know (or what we don't know) (54, 55), and Bayes theorem is an imperfect model for human learning (54, 56). This means that the frequentist, long-run perspective cannot be completely ignored, leading many statisticians to emphasize the importance of using frequentist criteria in the evaluation of Bayesian and likelihood methods (6, 13, 32, 53), which these methods typically fulfill quite well.

In the end, we must recognize that there is no automatic method in statistics, as there is not in life, that allows us both to evaluate individual situations and know exactly what the long-run consequences of that evaluation will be. The connection between inference in individual experiments and the number of errors we make over time is not found in the P value or in hypothesis tests. It is found only in properly assessing the strength of evidence from an experiment with Bayes factors and uniting this with a synthesis of all of the other scientific information that bears on the question at hand. There is no formula for performing that synthesis, nor is there a formula for assigning a unique number to it. That is where room for meaningful scientific discourse lies.

Sir Francis Bacon, the writer and philosopher who was one of the first inductivists, commented on the two attitudes with which one can approach nature. His comment could apply to the perspectives contrasted in these essays: "If we begin with certainties, we shall end in doubts; but if we begin with doubts, and are patient with them, we shall end with certainties" (57). Putting P values aside, Bayesian and frequentist approaches each provide an essential perspective that the other lacks. The way in which we balance their sometimes conflicting demands is what makes the process of learning from nature creative, exciting, uncertain, and, most of all, human.

Appendix I

Derivation of the minimum Bayes factor under a Gaussian distribution: The likelihood of a hypothesis given an observed effect, x, is proportional to the probability of x under that hypothesis. For a Gaussian distribution, the hypothesis typically concerns the mean. The probability of x under a Gaussian distribution with true mean $= \mu$ and standard error $= \sigma$, is (where the symbol "|" is read as "given"):

$$\Pr(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2 / 2}$$

Because the exponent is negative, the above probability is maximized when the exponent is zero, which occurs when $\mu = x$ (that is, the true mean μ equals the observed effect, x). The likelihood ratio for the null hypothesis ($\mu = 0$) versus the maximally supported hypothesis ($\mu = x$) is the minimum Bayes factor:

$$\frac{\Pr(x \mid \mu = 0, \sigma)}{\Pr(x \mid \mu = x, \sigma)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-0}{\sigma}\right)^2 / 2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-x}{\sigma}\right)^2 / 2}} = e^{-\left(\frac{x}{\sigma}\right)^2 / 2}$$

Because the Z-score is the observed effect, x, divided by its standard error, σ, the final term in the above equation is:

$$e^{-\left(\frac{x}{\sigma}\right)^2 / 2} = e^{-Z^2/2}$$

Appendix II

In the example posed in the first article (1), two treatments, called A and B, were compared in the same patients, and the preferred treatment in each patients was chosen. The two experimenters had different mindsets while conducting the experiment: one planned to study all six patients, whereas the other planned to stop as soon as treatment B was preferred. The first five patients preferred treatment A, and the sixth preferred treatment B.

The probability of the data under the two hypotheses is as follows.

Null hypothesis: Probability that treatment A is preferred = 1/2

Alternative hypothesis: Probability that treatment A is preferred = 5/6

In the "$n = 6$" experiment, this ratio is:

$$6\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^1 \bigg/ 6\left(\frac{5}{6}\right)^5\left(\frac{1}{6}\right)^1 = 0.23$$

The "6" appears above because the preference for treatment B could have occurred in any of the first five patients or in the sixth patient without a change in the inference.

In the "stop at first preference for treatment B" experiment, the ratio is:

$$\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^1 \bigg/ \left(\frac{5}{6}\right)^5\left(\frac{1}{6}\right)^1 = 0.23$$

References

1. **Goodman SN.** Toward evidence-based medical statistics. 1: The *P* value fallacy. Ann Intern Med. 1999;130:995-1004.
2. **Edwards A.** A History of Likelihood. International Statistical Review. 1974; 42:9-15.
3. **Fisher LD.** Comments on Bayesian and frequentist analysis and interpretation of clinical trials. Control Clin Trials. 1996;17:423-34.
4. **Brophy JM, Joseph L.** Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. JAMA. 1995;273:871-5.
5. **Browne RH.** Bayesian analysis and the GUSTO trial. Global Utilization of Streptokinase and Tissue Plasminogen Activator in Occluded Coronary Arteries [Letter]. JAMA. 1995;274:873.
6. **Good I.** Probability and the Weighing of Evidence. New York: Charles Griffin; 1950.
7. **Cornfield J.** The Bayesian outlook and its application. Biometrics. 1969;25: 617-57.
8. **Berger JO, Berry DA.** Statistical analysis and the illusion of objectivity. American Scientist. 1988;76:159-65.
9. **Berry D.** Interim analyses in clinical trials: classical vs. Bayesian approaches. Stat Med. 1985;4:521-6.
10. **Belanger D, Moore M, Tannock I.** How American oncologists treat breast cancer: an assessment of the influence of clinical trials. J Clin Oncol. 1991;9:7-16.
11. **Omoigui NA, Silver MJ, Rybicki LA, Rosenthal M, Berdan LG, Pieper K, et al.** Influence of a randomized clinical trial on practice by participating investigators: lessons from the Coronary Angioplasty Versus Excisional Atherectomy Trial (CAVEAT). CAVEAT I and II Investigators. J Am Coll Cardiol. 1998;31:265-72.
12. **Goodman SN, Royall R.** Evidence and scientific research. Am J Public Health. 1988;78:1568-74.
13. **Royall R.** Statistical Evidence: A Likelihood Primer. Monographs on Statistics and Applied Probability, #71. London: Chapman and Hall; 1997.
14. **Edwards A.** Likelihood. Cambridge, UK: Cambridge Univ Pr; 1972.
15. **Goodman SN.** Meta-analysis and evidence. Control Clin Trials. 1989;10:188-204, 435.
16. **Efron B.** Empirical Bayes methods for combining likelihoods. Journal of the American Statistical Association. 1996;91:538-50.
17. **Hardy RJ, Thompson SG.** A likelihood approach to meta-analysis with random effects. Stat Med. 1996;15:619-29.
18. **Berger J.** Statistical Decision Theory and Bayesian Analysis. New York: Springer-Verlag; 1985.
19. **Edwards W, Lindman H, Savage L.** Bayesian statistical inference for psychological research. Psychol Rev. 1963;70:193-242.
20. **Diamond GA, Forrester JS.** Clinical trials and statistical verdicts: probable grounds for appeal. Ann Intern Med. 1983;98:385-94.
21. **Lilford R, Braunholtz D.** The statistical basis of public policy: a paradigm shift is overdue. BMJ. 1996;313:603-7.
22. **Peto R.** Why do we need systematic overviews of randomized trials? Stat Med. 1987;6:233-44.
23. **Pogue J, Yusuf S.** Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet. 1998;351:47-52.
24. **Fisher R.** Statistical Methods and Scientific Inference. 3d ed. New York: Macmillan; 1973.
25. **Jeffreys H.** Theory of Probability. 2d ed. Oxford: Oxford Univ Pr; 1961.
26. **Kass R, Raftery A.** Bayes Factors. Journal of the American Statistical Association. 1995;90:773-95.
27. **Cornfield J.** A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. Journal of the American Statistical Association. 1966;61:577-94.
28. **Kass R, Greenhouse J.** Comments on "Investigating therapies of potentially great benefit: ECMO" (by JH Ware). Statistical Science. 1989;4:310-7.
29. **Spiegelhalter D, Freedman L, Parmar M.** Bayesian approaches to randomized trials. Journal of the Royal Statistical Society, Series A. 1994;157:357- 87.
30. **Berger J, Sellke T.** Testing a point null hypothesis: the irreconcilability of p-values and evidence. Journal of the American Statistical Association. 1987; 82:112-39.
31. **Bayarri M, Berger J.** Quantifying surprise in the data and model verification. Proceedings of the 6th Valencia International Meeting on Bayesian Statistics, 1998. 1998:1-18.
32. **Carlin C, Louis T.** Bayes and Empirical Bayes Methods for Data Analysis. London: Chapman and Hall; 1996.
33. **Casella G, Berger R.** Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association. 1987;82:106-11.
34. **Howard J.** The 2×2 table: a discussion from a Bayesian viewpoint. Statistical Science. 1999;13:351-67.
35. **Cornfield J.** Sequential trials, sequential analysis and the likelihood principle. American Statistician. 1966;20:18-23.
36. **Savitz DA, Olshan AF.** Multiple comparisons and related issues in the interpretation of epidemiologic data. Am J Epidemiol. 1995;142:904-8.
37. **Perneger T.** What's wrong with Bonferroni adjustments. BMJ. 1998;316: 1236-8.
38. **Goodman SN.** Multiple comparisons, explained. Am J Epidemiol. 1998;147: 807-12.
39. **Thomas DC, Siemiatycki J, Dewar R, Robins J, Goldberg M, Armstrong BG.** The problem of multiple inference in studies designed to generate hypotheses. Am J Epidemiol. 1985;122:1080-95.
40. **Greenland S, Robins JM.** Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology. 1991;2:244-51.
41. **Rothman KJ.** No adjustments are needed for multiple comparisons. Epidemiology. 1990;11:43-6.
42. **Berry DA.** A case for Bayesianism in clinical trials. Stat Med. 1993;12:1377-93.
43. **Chaloner K, Church T, Louis T, Matts J.** Graphical elicitation of a prior distribution for a clinical trial. The Statistician. 1993;42:341-53.
44. **Chaloner K.** Elicitation of prior distributions. In: Berry D, Stangl D, eds. Bayesian Biostatistics. New York: Marcel Dekker; 1996.
45. **Freedman L.** Bayesian statistical methods [Editorial]. BMJ. 1996;313:569-70.
46. **Fayers PM, Ashby D, Parmar MK.** Tutorial in biostatistics: Bayesian data monitoring in clinical trials. Stat Med. 1997;16:1413-30.
47. **Etzioni RD, Kadane JB.** Bayesian statistical methods in public health and medicine. Annu Rev Public Health. 1995;16:23-41.
48. **Berry DA.** Benefits and risks of screening mammography for women in their forties: a statistical appraisal. J Natl Cancer Inst. 1998;90:1431-9.
49. **Hughes MD.** Reporting Bayesian analyses of clinical trials. Stat Med. 1993; 12:1651-64.
50. **Berry DA, Stangl D, eds.** Bayesian Biostatistics. New York: Marcel Dekker; 1996.
51. **Berry DA.** Decision analysis and Bayesian methods in clinical trials. Cancer Treat Res. 1995;75:125-54.
52. **Spiegelhalter D, Thomas A, Best N, Gilks W.** BUGS: Bayesian Inference Using Gibbs Sampling. Cambridge, UK: MRC Biostatistics Unit; 1998. Available at www.mrc-bsu.cam.ac.uk/bugs.
53. **Rubin D.** Bayesianly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics. 1984;12:1151-72.
54. **Shafer G.** Savage revisited. Statistical Science. 1986;1:463-501.
55. **Walley P.** Statistical Reasoning with Imprecise Probabilities. London: Chapman and Hall; 1991.
56. **Tversky A, Kahneman D.** Judgment under uncertainty: heuristics and biases. In: Slovic P, Tversky A, Kahneman D, eds. Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge Univ Pr; 1982:1-20.
57. **Bacon F.** De Augmentis Scientiarum, Book I (1605). In: Curtis C, Greenslet F, eds. The Practical Cogitator. Boston: Houghton Mifflin; 1962.