Intervals for Post-Test Probabilities: A Comparison of Five Methods

DOUGLAS MOSSMAN, M.D.
Professor and Director, Division of Forensic Psychiatry,
Wright State University School of Medicine;
Adjunct Professor, University of Dayton School of Law

JAMES O. BERGER, PH.D.
Arts and Sciences Professor of Statistics
Institute of Statistics and Decision Sciences, Duke University

Submitted March 2001 to Medical Decision Making.

Address for correspondence:
W. S. U. Department of Psychiatry, P. O. Box 927, Dayton, Ohio 45401-0927
Telephone: (937) 276-8325 — Fax: (937) 275-2817 — e-mail: dmossman@pol.net

---

Abstract

Background. Several medical articles discuss methods of constructing confidence intervals for single proportions and the likelihood ratio, but scant attention has been given to the systematic study of intervals for the posterior odds or the "positive predictive value" of a test. ***Methods***. We describe five methods of constructing confidence intervals for post-test probabilities when estimates of sensitivity, specificity, and the pre-test probability of a disorder are derived from empirical data. We then evaluate each method to see how well the intervals' coverage properties correspond to their nominal value. ***Results***. When the estimates of pre-test probabilities, sensitivity, and specificity are derived from more than 80 subjects and are not close to zero or 1, all methods generate intervals with appropriate coverage properties. When these conditions are not met, however, the best-performing method is an "objective Bayesian" approach implemented by a simple simulation using a spreadsheet. ***Conclusions***. Physicians and investigators can generate accurate confidence intervals for post-test probabilities in small-sample situations using the objective Bayesian approach.

* * * * *

Key words: positive predictive value, Bayes's Theorem, posterior probability, posterior odds, post-test probability, confidence interval, objective Bayesian

Introduction

Mr. Smith is worried that he has disorder *D*, and consults Dr. Jones. Dr. Jones orders a test *T* that has only two outcomes, "positive" and "negative," and Mr. Smith's test result comes back "positive." Dr. Jones tells his patient that he has a 69% chance of having the disorder, given this test result.

Mr. Smith asks his doctor how he came to this conclusion. Dr. Jones — an unusual physician who uses Bayes's Theorem and discusses it with his patients — answers, "Twenty percent of men who are your age have *D*. Test *T* has a 90% sensitivity and a 90% specificity. Applying Bayes's Theorem, your positive test result means your chance of having *D* is 69%."

Mr. Smith likes his doctor's answer, because he is a *very* unusual patient who wants his care-givers to formulate clinical decisions with explicit mathematical rigor. He expects his physicians to use Bayes's Theorem, but is troubled by one aspect of Dr. Jones's calculation.

"I understand how you reached your conclusion," Mr. Smith responds, "but your information doesn't allow you to say that my chance of having *D* is *exactly* 69%. Your figure for the 20% prevalence came from a published medical study that evaluated 50 men in my age group. Ten of these men had *D*, but this finding only lets doctors *estimate* the prevalence of *D* as being 20% in men my age." Dr. Jones agrees with his patient's summary of the literature on the prevalence of *D*.

"Likewise," Mr. Smith continues, "your figures for sensitivity and specificity came from another published medical study of the diagnostic test *T*. That article reported that 36 of 40 people who had *D* also had positive tests, but only four of 40 people who did not have *D* had positive tests. These results just let the researchers *estimate* the test's sensitivity and specificity as 90%.

"The published estimates of prevalence, sensitivity, and specificity are subject to random sampling error," continues Mr. Smith. "So what I want to know is this: What is the 95 percent confidence interval for my probability of having *D*, given my positive test result and the imprecision in the estimates? Knowing whether the interval is narrow or broad might affect my decisions about getting other tests or choosing treatment."

Dr. Jones is not sure how to compute the answer to his patient's question, and the existing medical literature provides little help. Several writers have discussed methods of constructing confidence intervals for single proportions,[1] for the likelihood ratio,[2-4] and the odds ratio.[5] To the best of our knowledge, however, the medical literature does not prescribe or discuss methods for constructing intervals when proportions are combined in Bayes's Theorem.

This article attempts to fill this gap. The following section describes five methods of constructing intervals, and applies these methods to post-test probabilities calculated from

random samples and data-based estimates of prevalence (or pre-test probability of a disorder), sensitivity, and specificity. Subsequent sections describe calculations and simulation studies designed to evaluate how well the coverage of each method's intervals corresponds to its nominal value.

Description of Five Methods for Constructing Intervals

Symbols and Assumptions

Let p_0 represent a population's prevalence (or pre-test probability) of the disorder D. Assume the diagnostic test has just two outcomes, "positive" and "negative." Let p_1 represent the probability of a positive test when a patient from the population has D, or the *true positive rate* or *sensitivity* of test T. Let p_2 represent the probability of a positive test when a patient from the population does not have D, or the *false positive rate* of test T; the test's specificity will equal $1-p_2$. Bayes's Theorem then says that the probability ϕ of having D given a positive test result (sometimes called the "positive predictive value") is

$$\begin{aligned}\phi &= \frac{prevalence \cdot sensitivity}{prevalence \cdot sensitivity + [1 - prevalence]\cdot[1 - specificity]} \\ &= \frac{p_0 p_1}{p_0 p_1 + (1-p_0)p_2}\end{aligned} \qquad \textbf{(1)}$$

Clinicians and researchers cannot know the true values of p_0, p_1, p_2, but they often can estimate these values — $\hat{p}_0$, $\hat{p}_1$, $\hat{p}_2$ — from empirical data on disorder prevalence and test accuracy. The prevalence is estimated by $\hat{p}_0 = x_0/n_0$, where n_0 is the number of subjects evaluated in a "random" sample from the relevant population, and x_0 is the number of subjects who had D; test sensitivity is estimated by $\hat{p}_1 = x_1/n_1$, where n_1 equals the number of randomly chosen subjects known to have D who underwent testing with T, and x_1 is the number of these subjects who had positive test results; and the false positive rate is estimated by $\hat{p}_2 = x_2/n_2$, where n_2 is the number of randomly chosen subjects known not to have D who were tested, and x_2 is the number of these subjects who had positive test results. Dr. Jones estimated $\hat{\phi}$, Mr. Smith's chance of having disorder D, as follows:

$$\hat{\phi} = \frac{\hat{p}_0\hat{p}_1}{\hat{p}_0\hat{p}_1 + (1-\hat{p}_0)\hat{p}_2} = \frac{\frac{x_0}{n_0}\frac{x_1}{n_1}}{\frac{x_0}{n_0}\frac{x_1}{n_1} + \left(1-\frac{x_0}{n_0}\right)\frac{x_2}{n_2}} = \frac{\frac{10}{50}\frac{36}{40}}{\frac{10}{50}\frac{36}{40} + \left(1-\frac{10}{50}\right)\frac{4}{40}} = 0.692 \qquad (2)$$

The following subsections describe five ways to produce intervals for ϕ using x_0/n_0, x_1/n_1, and x_2/n_2. The first of these procedures — the objective Bayesian method — is the one that we ultimately recommend.

Objective Bayesian Method

In general, the Bayesian approach assigns prior distributions to p_0, p_1, and p_2, from which one determines a posterior distribution and interval for ϕ. The structure of the prior distributions might reflect physicians' or investigators' knowledge of (or "subjective" beliefs about) the nature of the population or the test data, in which case the analysis would be termed a "subjective Bayesian" analysis. Often, however, intervals for ϕ will be derived solely from the data, in which case the appropriate Bayesian approach is "objective Bayesian" analysis. Objective Bayesians choose "noninformative" or default prior distributions for the parameters, prior distributions that depend only on the data model under consideration, and not subjective beliefs.

The intervals for ϕ that result from the Bayesian approach are usually called "credible intervals," the distinction in terminology reflecting a distinction in interpretation: for a 95% credible interval computed in a specific application, one can assert that ϕ has a 95% chance of being in the interval. In contrast, 95% frequentist coverage means only that, in repeated use, a confidence procedure will result in intervals that contain ϕ 95% of the time; one cannot assert confidence for specific instances. The difference can be quite dramatic and have important practical consequences.[6] However, to avoid engaging in a debate concerning the Bayesian versus frequentist approaches, we will not further discuss this here. Instead, we will simply view the credible intervals developed from the objective Bayesian approach as frequentist confidence intervals to be considered for their frequentist coverage properties. A substantial literature[7] suggests that the objective Bayesian approach is the best way to develop intervals with good frequentist coverage in complex situations.

The beta distribution is useful[1,8–10] for describing the *a priori* distribution of the parameter p, arising as the probability of a success on an individual trial, of the binomial distribution. The density of a $Beta(a,b)$ distribution is given by

$$\beta_{(a,b)}(p) = C\,p^{a-1}(1-p)^{b-1} \tag{3}$$

for $0 < p < 1$, and zero otherwise.[10] The constant C is chosen so the total area under the distribution equals 1.

Objective Bayesians have primarily used two possible "noninformative" prior distributions for p. Laplace[11] utilized the "uniform" prior distribution, $Beta(1,1)$, while Jeffreys[12] proposed using the $Beta(½,½)$ prior distribution — now called the "Jeffreys prior" — for a proportion p. The Jeffreys prior is known to give confidence sets with very good coverage properties for a single proportion p.[13,14] In this article, we study its use for multiple proportions; in particular, we use it as the prior distribution for each of the

proportions p_i found in Equation 1.

After observing (x_i,n_i), the posterior distribution of p_i is the $Beta(x_i + ½, n_i - x_i + ½)$ distribution. Kleiter has described a complicated formula that uses a weighted-sum of beta distributions to produce the posterior distribution for the Bayes parameter ϕ.[9] However, it is much easier to obtain the desired confidence interval for ϕ by working directly with the $Beta(x_i + ½, n_i - x_i + ½)$ distributions via simulation to produce what we will henceforth call "ϕ_β intervals." Indeed, the following simple Monte Carlo procedure will yield an estimate of the "equal-tailed" 100(1-α)% confidence set for ϕ. One uses a computer to:

(1) draw, at random, values from each of the $Beta(x_i + ½, n_i - x_i + ½)$ distributions, i=0,1,2;

(2) combine these three values in Equation 1 to obtain a value of $\hat{\phi}$;

(3) repeat this process N times (where N is a large number, say, 10,000) to generate N values of $\hat{\phi}$, and sort these values from lowest to highest;

(4) find the integers nearest to $N\alpha/2$ and $N(1-\alpha/2)$, and choose the corresponding sorted values of $\hat{\phi}$ as the lower and upper limits of the confidence set for ϕ.

Commonly available spreadsheet programs contain random number generators and inverse beta distribution commands, and with a little practice, it takes just a few seconds to create a spreadsheet that performs the just-described procedure. For the Smith–Jones example from the introduction, the 95% ϕ_β interval is [0.431, 0.887].

Two Log-Odds Methods

If one divides Equation 1 by $1-\phi$ and rearranges terms, the resulting version of Bayes's Theorem expresses O_{post} — the post-test *odds* of having D, given a positive test result — as the product of the pre-test odds, O_{ante}, and the likelihood ratio, LR:

$$O_{post} = \frac{\phi}{1-\phi} = \frac{p_0}{1-p_0} \times \frac{p_1}{p_2} = O_{ante} \times LR \qquad (4)$$

The likelihood ratio, O_{ante}, and O_{post} can range from zero to infinity. Symmetric confidence intervals for LR constructed from $\hat{p}_1$ and $\hat{p}_2$ are unlikely to be accurate and may include values less than zero.[2,3,15] One commonly-invoked remedy[3,16] is to construct confidence intervals for LR using its natural logarithm, ln LR. Extending this reasoning to O_{post}, one would obtain upper and lower bounds of the 100(1-α)% interval, $\hat{O}_{post}^{\,U}$ and $\hat{O}_{post}^{\,L}$, from

$$\hat{O}_{post}^{\,U,L} \approx e^{\ln\left(\frac{\hat{\phi}}{1-\hat{\phi}}\right) \pm z_{\alpha/2}\sqrt{\mathrm{var}(\ln \hat{O}_{post})}} \tag{5}$$

where $z_{\alpha/2}$ is the normal deviate of the upper $\alpha/2$ percentile of the normal distribution, and

$$\mathrm{var}(\ln \hat{O}_{post}) \approx \frac{n_0}{x_0(n_0 - x_0)} + \frac{1}{x_1} - \frac{1}{n_1} + \frac{1}{x_2} - \frac{1}{n_2} \tag{6}$$

Equation 6 arises from application of the standard "delta method" together with some algebra.[17] One uses the relationship $\phi = \frac{O_{post}}{1+O_{post}}$ to convert $[\hat{O}_{post}^{\,U}, \hat{O}_{post}^{\,L}]$ to $[\hat{\phi}^{U}, \hat{\phi}^{L}]$, the interval for ϕ.

Equations 5 and 6 are undefined if x_0, x_1, or x_2 equal zero, or if $x_0=n_0$; also, if $x_1=n_1$ and $x_2=n_2$, the likelihood ratio makes no contribution to var($\ln \hat{O}_{post}$). Here are two simple ways to address this.

(1) When calculating $\hat{\phi}$ in Equation 5 and $\hat{O}_{post}^{\,U}$ and $\hat{O}_{post}^{\,L}$ in Equation 6, substitute ½ for x_0, x_1, or x_2 whenever these terms equal zero, and substitute $(n_i - ½)$ for x_i whenever $x_i = n_i$. Hereinafter, these "substituted" results will be designated "ϕ_S intervals." For the Smith–Jones example, the 95% ϕ_S interval is [0.413, 0.878].

(2) Adapt the *LR* variance-approximation procedure of Walter[18] and by Pettigrew and colleagues[19] to the present problem; that is, estimate O_{post} from

$$\ln(\hat{O}_{post}) \approx \ln\left(\frac{x_0 + ½}{n_0 - x_0 + ½}\right) + \ln\left(\frac{x_1 + ½}{n_1 + ½}\right) - \ln\left(\frac{x_2 + ½}{n_2 + ½}\right) \tag{7}$$

calculate var($\ln \hat{O}_{post}$) from

$$\mathrm{var}(\ln \hat{O}_{post}) \approx \frac{n_0 + ½}{(x_0 + ½)(n_0 - x_0 + ½)} + \frac{1}{x_1 + ½} - \frac{1}{n_1 + ½} + \frac{1}{x_2 + ½} - \frac{1}{n_2 + ½} \tag{8}$$

and use these results in Equation 5. Intervals obtained using this approach are henceforth referred to as "$\phi_{½}$ intervals." For the Smith–Jones example, the 95% $\phi_{½}$ interval is [0.410, 0.864].

Notice that one can quickly compute ϕ_S intervals and $\phi_{½}$ intervals using an inexpensive pocket calculator.

A Method Using the Gart-Nam Approximation for the Likelihood Ratio

Gart and Nam[2] describe an iterative procedure for calculating the $100(1-\alpha)\%$ confidence interval for *LR* using likelihood-based estimates supplemented by skewness correction derived from Bartlett's score method.[20,21] Centor has discussed the advantages of the Gart-Nam procedure and provides computer code that performs it,[22] and the present authors have developed a spreadsheet that accomplishes the same task.

One can include Gart and Nam's interval-estimation procedure in a Monte Carlo simulation similar to that described for the objective Bayesian method to obtain what we shall term "ϕ_{GN} intervals" for ϕ. Instead of sampling from the beta distributions for p_1, and p_2, one uses the formulation of Bayes's Theorem in Equation 4, and samples from the distribution for *LR* implied by the Gart-Nam algorithm. One obtains a set of *N* values of $\hat{O}_{post}$ that are readily converted to *N* values of $\hat{\phi}$, the central $N(1-\alpha)$ of which will represent the $100(1-\alpha)\%$ confidence interval for ϕ. For the Smith–Jones example, the 95% ϕ_{GN} interval is [0.439, 0.891].

A Delta Method Derived from Kleiter

In addition to providing an analytic expression for the posterior distribution of ϕ, Kleiter suggests approximating this posterior using a secondary beta distribution, with parameters *a* and *b* estimated from the data utilizing the delta method. Equal-tailed $100(1-\alpha)\%$ confidence intervals can then be found using the inverse beta distribution function available on many spreadsheets. The Appendix describes the derivation and details of this interval estimation method. Hereinafter, the results obtained with this delta method will be referred to as "ϕ_δ values." For the Smith–Jones example, the 95% ϕ_δ interval is [0.419, 0.888].

Comparing Intervals

The first issue of interest to practitioners is the typical size of the confidence intervals and the magnitude of the differences that can be expected from the five just-described procedures. Table 1 gives the 95% confidence intervals for these procedures under nine possible data sets. These data sets utilize values for $\hat{p}_0$, $\hat{p}_1$, and $\hat{p}_2$ and sample sizes that physicians typically encounter in clinical medicine. The most important fact to note from Table 1 is simply that the 95% confidence intervals for [illegible] are quite large, even for the moderately large sample sizes (n_i=80). The impact of the uncertainty in [illegible] can thus be quite significant.

[place Table 1 about here]

The top portion of Table 1 shows that, with moderately large samples and when $\hat{p}_0$, $\hat{p}_1$, and $\hat{p}_2$ are not close to zero or 1, the methods produce very similar intervals; the

differences usually would be too small to influence a practical decision-making task. When $\hat{p}_0$, $\hat{p}_1$, and/or $\hat{p}_2$ equal $^1/_{10}$ or $^9/_{10}$, however, the intervals differ; their dissimilarities also become more salient when the data samples are small (middle portion of Table 1, n_i=20). The bottom portion of Table 1 shows how the five methods behave when the sample size is small and either $\hat{p}_0$ or $\hat{p}_2$ equals 0; here, the differences in calculated intervals become pronounced. In particular, the objective Bayesian intervals are considerably shorter, a desirable property that might have significant clinical consequences.

The other key feature of the confidence intervals that needs to be ascertained is the extent to which they meet their nominal coverage goals. If one asserts, for example, that $[\hat{\phi}^L, \hat{\phi}^U]$ is the central or "equal-tailed" 95% confidence interval for ϕ, one is saying that there is a 2.5% chance that ϕ falls below the interval and a 2.5% chance that ϕ falls above the interval.[12] So, one can evaluate a particular interval-generating procedure by computing the actual percentage of the time that, in repeated use, the procedure produces a confidence interval that lies above or below the actual ϕ.

To evaluate the coverage of the ϕ_β and ϕ_{GN} interval estimation procedures, we choose triads of p_0, p_1, and p_2 which yield a known value of ϕ, and choose values of n_0, n_1, and n_2 to represent sample sizes from the population and subpopulations relevant to our inferences about p_0, p_1, and p_2. We then "simulate" research studies with a computer, making a large number J of random drawings of sizes n_0, n_1, and n_2 from the populations. Each of the J random drawings yields data from which we can construct confidence intervals for ϕ, and we then see how often the J intervals obtained with each method undershoot or exceed the true value of ϕ. A "good" interval construction method will produce $100(1-\alpha)$% intervals that correspond to their nominal value, so that roughly $J\alpha/2$ intervals will undershoot ϕ and roughly $J\alpha/2$ intervals will exceed ϕ.[23]

For the $\phi_{½}$, ϕ_S, and ϕ_δ methods, which use explicit formulae to calculate intervals, one can use an analytic evaluative approach based on the binomial probabilities of the data that correspond to the given values of p_0, p_1, and p_2. The probability of each possible triad of data is simply the product of the three respective binomial probabilities. Thus, for each possible data triad, one can simply compute the confidence interval and note whether the interval exceeds or lies below the true value of ϕ (computed by Equation 1). Summing the data triad probabilities over all those data for which the interval exceeds or lies below ϕ yields the desired information about interval coverages properties.

Table 2 provides coverage results for three "scenarios" involving hypothetical, known populations for which the values of the triplet (p_0, p_1, p_2) were (¼,¾,¼), ($^1/_{10}$,$^9/_{10}$,$^1/_{10}$), or (½,$^9/_{10}$,$^1/_{10}$), the same values used in the upper and middle parts of Table 1. For the simulation studies of the ϕ_β and ϕ_{GN} methods, J=200,000 random draws of 20, 40, 80, or 160 were made from each of the three "populations" whose characteristics are represented by p_0, p_1, and p_2, so that n_0=n_1=n_2 always equaled 20, 40, 80, or 160, but x_{0j}, x_{1j}, and x_{2j} varied randomly with each draw. The values chosen for n_i allowed us to explore the small- and larger-sample performance behavior of each interval construction method. For the $\phi_{½}$,

ϕ_S, and ϕ_δ methods, we assumed the same "population" characteristics, but made exact computations according to the above-discussed analytic calculation procedure.

[place Table 2 about here]

The results from evaluating the five methods appear in Table 2, the interpretation of which is most easily explained through an example. The six numbers in the first row of the first column describe the $\phi_{½}$ method's performance when 20 "subjects" are randomly sampled from a population for which $p_0=¼$, $p_1=¾$, and $p_2=¼$. If $n_0=n_1=n_2=20$, there are $(20+1)^3 = 9{,}261$ possible combinations of samples, and for each possible combination, the 90%, 95%, and 99% confidence intervals for ϕ (*i.e.*, intervals with $\alpha=0.10$, $\alpha=0.05$, and $\alpha=0.01$) were calculated, along with the probability of obtaining each of those intervals.*

Because $\phi = \dfrac{(¼)(¾)}{(¼)(¾) + (1-¼)(¼)} = ½$, we hope to find that the lower limit of the 99% intervals is *greater* than ½ about 0.5% of the time, and that the upper limit of the 99% intervals is *less* than ½ about 0.5% of the time. The results in Table 2 show, however, that only 0.12% of the $\phi_{½}$ method's 99% intervals would be expected to lie above ½ and that only 0.03% of the intervals would lie below ½. That is, when $n_0=n_1=n_2=20$, the $\phi_{½}$ method's 99% intervals are too wide, as are the 95% and 90% intervals.

The next three rows show that as the n_i increase, the $\phi_{½}$ intervals get closer to the nominal values. In the fifth row, however, which contains results involving reference populations where $p_0={}^1/_{10}$, $p_1={}^9/_{10}$, and $p_2={}^1/_{10}$, the $\phi_{½}$ method did poorly again. Especially when the populations are small, the $\phi_{½}$ method produced confidence intervals that were too broad or badly skewed.

Why would these problems occur? The answer comes from using Table 2 to compare the results for the $\phi_{½}$ method with those for the ϕ_β method, and then examining the lower portions of Table 1:

- Table 2 shows that intervals using the ϕ_β method come fairly close to what we would desire given our earlier interpretation of the assertion "$(\hat{\phi}^L, \hat{\phi}^U)$ is the central $100(1-\alpha)$% confidence interval for ϕ," even when sample sizes are small and the reference populations' values of p_0, p_1, or p_2 are ${}^1/_{10}$ or ${}^9/_{10}$. When, for example, $p_0={}^1/_{10}$ and $n_0=20$, approximately 12.2% of the 20-member draws from the population would contain *no* members with the disorder, so that $\hat{p}_0=x_0/n_0=0$. When random 20-member drawings are made simultaneously from populations in

---

* To illustrate: in random independent samplings from a population in which $p_0=¼$, $p_1=¾$, and $p_2=¼$, the probability of obtaining $\hat{p}_0=x_0/n_0=4/20$ is 0.190, the probability of obtaining $\hat{p}_1=x_1/n_1=17/20$ is 0.134, and the probability of obtaining $\hat{p}_2=x_2/n_2=4/20$ is 0.202; the joint probability of these three results is 0.00514. Using the $\phi_{½}$ method, these results yield $\hat{\phi}=0.465$, with a 95% interval of [0.1965, 0.7548].

which $p_0 = {}^1/_{10}$, $p_1 = {}^9/_{10}$, and $p_2 = {}^1/_{10}$, the probability of getting a draw in which either $\hat{p}_0 = x_0/n_0 = 0$, $\hat{p}_1 = x_1/n_1 = 1$, or $\hat{p}_2 = x_2/n_2 = 0$, or some combination thereof occurs, is $1 - [1 - (0.9)^{20}]^3 = 0.322$. This means that, for small samples, the coverage properties are very strongly influenced by the difficult-to-handle cases where $\hat{p}_0 = 0$, $\hat{p}_1 = 1$, and/or $\hat{p}_2 = 0$.[†]

- The lower portion of Table 1 shows that when $x_0/n_0 = 0/20$ or $x_2/n_2 = 0/20$, the ϕ_β intervals are quite different from — and, as was mentioned earlier, narrower than — the $\phi_{½}$ intervals. Thus, the difference in how the two methods handle these types of cases probably accounts for the major differences in the two procedures' coverage properties.

Taken as a whole, the results in Table 2 show that the objective Bayesian method is the best performer. Even with the smaller samples, the ϕ_β intervals are consistently near the desired frequency of under- and over-shooting ϕ, and there is little systematic tendency for ϕ_β intervals to be biased downward or upward. The ϕ_{GN} method is the "runner up": although it yields intervals that usually are as close to nominal values as the ϕ_β intervals, it is much more difficult to implement and also typically gives larger intervals. Of the three methods that make use of explicit formulae, the ϕ_δ method does best, though it sometimes does poorly with small samples. Both log-odds methods do well in the larger-sample evaluations when reference probabilities do not lie close to 0 or 1. When these conditions do not obtain, however, the ϕ_S and $\phi_{½}$ intervals usually are too broad and tend to be biased downward.

How might the ϕ_β method work under "extreme" circumstances, *i.e.*, in situations where reference probabilities lie very close to 0 or 1? Of course, there are an infinite number of combinations of such circumstances, but the lowest portion of Table 2 gives us a flavor of the method's limitations. This portion summarizes findings from 200,000 simulations when $p_0 = {}^1/_{100}$, $p_1 = {}^{99}/_{100}$, and $p_2 = {}^1/_{100}$. Unless the sample size is nearly 200, the coverage of this procedure fails to attain the nominal level. Interpretation of this finding is far from clear, however. The problem could be the rather simplistic use, in this multivariate situation, of the *Beta*(½,½) priors, which were designed for binomial parameters alone. The objective Bayesian approach can, in principle, be used to determine that prior which would provide intervals having optimal frequentist coverage, but determination of this prior would be a formidable undertaking and would result in a much more complicated procedure. Furthermore, in extreme cases where the parameters are close to the boundary of the parameter space, the difference between the Bayesian and the frequentist

---

[†] When $p_0 = {}^1/_{10}$, $p_1 = {}^9/_{10}$, $p_2 = {}^1/_{10}$, and the sample size is 40, the chance that random drawing will lead to $\hat{p}_0 = 0$, $\hat{p}_1 = 1$, and/or $\hat{p}_2 = 0$ is $1 - [1 - (0.9)^{40}]^3 = 0.0437$. This suggests that the intervals constructed under these circumstances figure importantly in the results concerning 40-member reference populations. When the sample size is 80, however, the chance is only $1 - [1 - (0.9)^{80}]^3 = 6.55 \times 10^{-4}$.

interpretations of confidence intervals becomes highly significant and cannot be ignored in practice.

A Medicolegal Example

Malingering is a commonly-encountered phenomenon in medical evaluations that occur in legal contexts, where being (or appearing to be) ill may lead to financial compensation or permit an individual to avoid legal responsibility.[24] For example, studies and surveys suggest that roughly one-sixth of persons evaluated concerning their competence to stand trial (CST) feign incapacitating mental disorders.[25–27] Each year, an estimated 30,000–60,000 CST evaluations are performed each year in the United States,[28,29] which implies that U.S. mental health professionals evaluate 5,000–10,000 defendants each year who are feigning incompetence to stand trial.

Mossman and Hart have suggested that by using Bayes's Theorem, mental health professionals could improve their interpretations of data on malingering and make better presentations of evidence on malingering in court.[30] Rogers and Salekin[27] argue against doing this: they assert that imprecision in the estimated "base rate" of malingering (in this article's symbols, $\hat{p}_0$) and the estimated accuracy of malingering instruments (*i.e.*, $\hat{p}_1$ and $\hat{p}_2$) would result in confidence intervals for the probability of malingering (*i.e.*, ϕ) that were so broad as to make Bayesian estimates useless, misleading, and unfair to evaluees.

The preceding sections give us the tools to evaluate Rogers and Salekin's claim. One of the malingering measures discussed by Mossman and Hart is a simple "Symptom Validity Test" (SVT) for detecting feigned memory deficits. In one study describing the SVT, 17 of 20 malingering simulators made four or more errors, but only one of 40 genuine patients did this poorly;[31] based on this small study then, $\hat{p}_1 = x_1/n_1 = 17/20$ and $\hat{p}_2 = x_2/n_2 = 1/40$. As an example of how one might estimate p_0, we use results reported by Gothard and colleagues,[25] who examined findings in 55 CST evaluees and determined that 25 were competent, 23 were incompetent, and seven were malingering. Because the purpose of testing for malingered incompetence involves distinguishing defendants who are feigning illness from defendants who are truly incompetent, the relevant portion of the sample does not include the 25 defendants who were deemed competent; thus $\hat{p}_0 = x_0/n_0 = 7/30$.

We now can estimate ϕ and construct a confidence interval. Suppose a defendant-evaluee has a "positive" result on the SVT, i.e., has made four or more errors on the test. Using Equation 2, we find that $\hat{\phi} = 0.912$, and the ϕ_β method gives a 95% confidence interval for ϕ of [0.633, 0.990]. This interval is too broad to let us say that a positive SVT is "proof beyond a reasonable doubt" of malingering. In U.S. jurisdictions, however, adjudicatory competence typically is assumed unless a preponderance of the evidence shows that a defendant's mental condition makes him unable to proceed.[32,33] Thus, even with the small sample sizes used in this example, the SVT's confidence interval is narrow enough to support testimony that the defendant probably was malingering.

Discussion

Some version of Equation 1 is a commonplace in texts on medical decision-making[3,4] and Bayesian statistics.[6,8] Medical articles commonly report tests' positive and negative predictive values, but medical publications rarely (if ever) report confidence intervals for these values. Although the probabilities used to calculate positive and negative predictive values are themselves imprecise, little systematic attention has been devoted to the problem of "propagating" these imprecisions[9] to generate confidence intervals for post-test probabilities.

This article suggests that investigators and clinicians can construct such intervals in several ways. Intervals based on the natural logarithm of the posterior odds can be calculated quickly using inexpensive hand-held calculators; these intervals are reasonably accurate when the estimates of pre-test probabilities, sensitivity, and specificity are based on large numbers of subjects and do not lie too close to either 0 or 1. When these conditions are not met, or when better accuracy is desired, investigators can produce useful confidence intervals quickly if they have access to one of the many available spreadsheets with a command that computes the inverse of the cumulative beta distribution function. The ϕ_β method should, under many (though not all) conditions, generate intervals whose coverage properties are close to the nominal value.

Acknowledgments

The authors thank the anonymous reviewers for their many helpful comments on an earlier version of this article, and German Molina for his computer assistance in performing the simulation studies. Professor Berger's research was partially supported by the National Science Foundation, Grants DMS-9802261 and DMS-0103265.

References

1. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine 1998;17:857-872.

2. Gart JJ, Nam J. Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. Biometrics 1988;44:323-338.

3. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. Journal of Clinical Epidemiology 1991;44:763-770.

4. Agresti A. Introduction to categorical data analysis. New York: Wiley, 1996.

5. Fleiss JL, Davies M. Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. American Journal of Epidemiology 1982;115:841-845.

6. Berger JO. Statistical decision theory and Bayesian analysis, 2nd edition. New York: Springer-Verlag, 1985.

7. Yang R, Berger JO: A catalog of noninformative priors. Viewable at www.stat.duke.edu/~berger/papers/catalog.html.

8. Iverson GR. Bayesian Statistical Inference. Newbury Park, CA: Sage Publications, 1984.

9. Kleiter GD. Propagating imprecise probabilities in Bayesian networks. Artificial Intelligence 1996;88:143-161.

10. Freund JE, Williams FJ. Dictionary/Outline of Basic Statistics. Mineola, NY: Dover, 1991.

11. Laplace PS. Théorie Analytique des Probabilités. Paris: V. Courcier, 1812.

12. Jeffreys H. Theory of probability (3rd edition). Oxford: Clarendon Press, 1961.

13. Brown L, Cai T, DasGupta A. Confidence intervals for a binomial proportion and edgeworth expansions. Annals of Statistics (in press, 2001).

14. Brown L, Cai T, DasGupta A. Interval estimation for a binomial proportion. Statistical Science (in press, 2001)

15. Peirce JC, Cornell RG. Integrating stratum-specific likelihood ratios with the analysis of ROC curves. Med Decis Making 1993;13:141-51.

16. Katz D, Baptista J, Azen SP, Pike MC. Obtaining confidence intervals for the risk ratio in cohort studies. Biometrics 1978;34:469-474.

17. Lehmann E, Casella G. Theory of Point Estimation. New York: Springer, 1998.

18. Walter SD. The distribution of Levin's measure of attributable risk. Biometrika 1975;62:371-375.

19. Pettigrew HM, Gart JJ, Thomas DG. The bias and higher cumulants of the logarithm of a binomial variate. Biometrika 1986;73:425-435.

20. Bartlett MS. Approximate confidence intervals, II. More than one unknown parameter. Biometrika 1953;40:306-317.

21. Bartlett MS. Approximate confidence intervals, III. A bias correction. Biometrika 1955;43:993-998.

22. Centor RM. Estimating confidence intervals of likelihood ratios. Med Decis Making 1992;12:229-233.

23. Sims CA, Zha T. Error Bands for Impulse Responses. Federal Reserve Bank of Atlanta Working Paper 95-6, September 1995.

24. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition. Washington, DC: American Psychiatric Association, 1994.

25. Gothard S, Rogers R, Sewell KW. Feigning incompetency to stand trial: An investigation of the Georgia Court Competency Test. Law and Human Behavior 1995;19:363-373

26. Rogers R, Sewell KW, Goldstein A. Explanatory models of malingering: A prototypical analysis. Law and Human Behavior 1994;18:543-552.

27. Rogers R, Salekin RT. Beguiled by Bayes: A reanalysis of Mossman and Hart's estimates of malingering. Behavioral Sciences and the Law. 1998;16:147-153.

28. Nicholson RA, Kugler KE. Competent and incompetent criminal defendants: a quantitative review of comparative research. Psychological Bulletin 1991;109:355-370.

29. Poythress NG, Nichholson R, Otto RK, Edens JF, Bonnie RJ, Monahan J, Hoge SK. The MacArthur Competence Assessment Tool—Criminal Adjudication. Odessa, FL: Psychological Assessment Resources, 1999.

30. Mossman D, Hart KJ. Presenting evidence of malingering to courts: insights from decision theory. Behavioral Sciences and the Law. 1996;14:271-291.

31. Guilmette TJ, Hart KJ, Guiliano AJ, Leininger BE. A comparison of the Fifteen Item Test and the Hiscock Forced-Choice Procedure in detecting simulated memory impairment. Clinical Neuropsychologist. 1994;8:283-294.

32. Cooper v. Oklahoma, 517 U.S. 348 (1996).

33. Ohio Revised Code § 2945.37(G).

34. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Boston: Butterworths, 1988.

Appendix: Calculation of *a* and *b* for Kleiter's Delta Method

For a beta distribution with parameters a and b, the mean μ is $\frac{a}{a+b}$ and the variance σ^2 is $\frac{ab}{(a+b)^2(a+b+1)}$.[12] Some algebra gives these solutions for a and b:

$$a = \mu\left[\frac{\mu(1-\mu)}{\sigma^2} - 1\right]; \qquad b = (1-\mu)\left[\frac{\mu(1-\mu)}{\sigma^2} - 1\right] \tag{A1}$$

To obtain estimates of these values from the data, one can use $\hat{\phi}$ from Equation 2 as the estimate of the mean μ, and the variance σ^2 can be approximated using the delta method, namely:

$$\text{var}(\hat{\phi}) \approx \left(\frac{\partial\hat{\phi}}{\partial\hat{p}_0}\right)^2 \text{var}(\hat{p}_0) + \left(\frac{\partial\hat{\phi}}{\partial\hat{p}_1}\right)^2 \text{var}(\hat{p}_1) + \left(\frac{\partial\hat{\phi}}{\partial\hat{p}_2}\right)^2 \text{var}(\hat{p}_2) \tag{A1}$$

Differentiating Equation 1 with respect to p_0, p_1, and p_2 yields the partial derivatives

$$\left(\frac{\partial\hat{\phi}}{\partial\hat{p}_0}\right) = \frac{\hat{p}_1\hat{p}_2}{[\hat{p}_0\hat{p}_1 + (1-\hat{p}_0)\hat{p}_2]^2}$$

$$\left(\frac{\partial\hat{\phi}}{\partial\hat{p}_1}\right) = \frac{(1-\hat{p}_0)\hat{p}_0\hat{p}_2}{[\hat{p}_0\hat{p}_1 + (1-\hat{p}_0)\hat{p}_2]^2} \tag{A3}$$

$$\left(\frac{\partial\hat{\phi}}{\partial\hat{p}_2}\right) = \frac{-(1-\hat{p}_0)\hat{p}_0\hat{p}_1}{[\hat{p}_0\hat{p}_1 + (1-\hat{p}_0)\hat{p}_2]^2}$$

Three adjustments help one avoid undefined terms and improve this method's results:

- when $x_i=0$, change x_i to ½;
- when $x_i=n_i$, change x_i to $n_i - ½$;
- add ½ to the values of a and b calculated from Equations A1–A3, so that one calculates the $100(1-\alpha)\%$ interval from the inverse of the $Beta(a+½, b+½)$ distribution.

Table 1. — Results from five methods of constructing 95 percent intervals for ϕ. For explanation of symbols, see text.

	Data and sample sizes		
Method	x_0/n_0, x_1/n_1, x_2/n_2 8/80, 72/80, 8/80	x_0/n_0, x_1/n_1, x_2/n_2 20/80, 60/80, 20/80	x_0/n_0, x_1/n_1, x_2/n_2 40/80, 72/80, 8/80
$\phi_{½}$ (log odds, add ½)	[0.264, 0.714]	[0.341, 0.651]	[0.797, 0.949]
ϕ_{β} (objective Bayesian)	[0.264, 0.726]	[0.346, 0.658]	[0.808, 0.952]
ϕ_{S} (log odds, substituted)	[0.272, 0.728]	[0.344, 0.656]	[0.803, 0.952]
ϕ_{GN} (Gart-Nam)	[0.284, 0.740]	[0.349, 0.662]	[0.807, 0.955]
ϕ_{δ} (delta method)	[0.266, 0.734]	[0.343, 0.657]	[0.812, 0.955]
	x_0/n_0, x_1/n_1, x_2/n_2 2/20, 18/20, 2/20	x_0/n_0, x_1/n_1, x_2/n_2 5/20, 15/20, 5/20	x_0/n_0, x_1/n_1, x_2/n_2 10/20, 18/20, 2/20
$\phi_{½}$ (log odds, add ½)	[0.112, 0.842]	[0.210, 0.769]	[0.632, 0.970]
ϕ_{β} (objective Bayesian)	[0.107, 0.872]	[0.216, 0.780]	[0.679, 0.980]
ϕ_{S} (log odds, substituted)	[0.122, 0.878]	[0.216, 0.784]	[0.648, 0.978]
ϕ_{GN} (Gart-Nam)	[0.149, 0.902]	[0.241, 0.795]	[0.699, 0.982]
ϕ_{δ} (delta method)	[0.099, 0.901]	[0.208, 0.792]	[0.695, 0.982]
	x_0/n_0, x_1/n_1, x_2/n_2 0/20, 18/20, 2/20	x_0/n_0, x_1/n_1, x_2/n_2 0/20, 15/20, 5/20	x_0/n_0, x_1/n_1, x_2/n_2 10/20, 18/20, 0/20
$\phi_{½}$ (log odds, add ½)	[0.009, 0.799]	[0.004, 0.570]	[0.675, 0.998]
ϕ_{β} (objective Bayesian)	[0.000, 0.649]	[0.000, 0.326]	[0.857, 1.000]
ϕ_{S} (log odds, substituted)	[0.010, 0.837]	[0.004, 0.588]	[0.570, 1.000]
ϕ_{GN} (Gart-Nam)	[0.032, 0.824]	[0.014, 0.507]	[0.701, 0.984]
ϕ_{δ} (delta method)	[0.006, 0.833]	[0.003, 0.438]	[0.695, 0.982]