The Role of Informative Priors in Zero-Numerator Problems: Being Conservative Versus Being Candid

Robert L. WINKLER, James E. SMITH, and Dennis G. FRYBACK

The "Rule of Three" gives an approximation for an upper 95% confidence bound for a proportion in a zero-numerator problem, which occurs when the observed relative frequency is zero. We compare the results from the Rule of Three with those from a Bayesian approach with noninformative and informative priors. Informative priors are especially valuable in zero-numerator problems because they can represent the available information and because different noninformative priors can give conicting advice. Moreover, the use of upper 95% bounds and noninformative priors in an effort to be conservative may backfire when the values are used in further predictive or decision-theoretic calculations. It is better to be candid than conservative, using all of the information available in forming the prior and considering the uncertainty represented by the full posterior distribution.

KEY WORDS: Bayesian inference; Noninformative priors; Rare events; Rule of Three.

1. INTRODUCTION

A zero-numerator problem is a situation in which we estimate the probability for an event that is conceivably possible but has not yet occurred in the data that are available. Examples (Hanley and Lippman-Hand 1983) include "a still-perfect surgical record, a field trial of a vaccine that uncovered no major side effects, an ophthalmology practice in which no patient with glaucoma was younger than 23 years, an airline that has never had a fatality." Our interest in this question was motivated by considering the false-positive rate for a medical test with a history of no positive results. Having observed no occurrences of the event, hence an observed relative frequency of zero, is an indication of a low probability, but it clearly does not imply a probability of zero.

Robert L. Winkler is James B. Duke Professor in the Fuqua School of Business, Duke University, Durham, NC 27708-0120, and in the Institute of Statistics and Decision Sciences at Duke (E-mail: rwinkler@mail.duke.edu). James E. Smith is Associate Professor in the Fuqua School of Business, Duke University, Durham, NC 27708-0120 (E-mail: jes9@mail.duke.edu). Dennis G. Fryback is Professor in the Department of Population Health Sciences, University of Wisconsin, 610 North Walnut Street, Madison, WI 53705-2397 (E-mail: dfryback@facstaff.wisc.edu). This work was supported in part by the National Science Foundation under grants SES 98-188855 (Winkler) and SBR 98-09176 (Smith). The authors are grateful to the editor, an associate editor, and a referee for helpful comments.

The standard formula for confidence bounds for a Bernoulli parameter p breaks down in a zero-numerator situation. The Rule of Three states that $3/n$ is a good approximation for an upper 95% confidence bound for p when we have seen n independent trials with no occurrences of the event of interest (Louis 1981; Hanley and Lippman-Hand 1983). Exact confidence bounds are not difficult to compute, but the Rule of Three approximation is quite simple and very accurate for large n. Jovanovic and Levy (1997) adopted a Bayesian approach to generate modified Rules of Three, which approximate the 0.95 fractile of the posterior distribution of p under a uniform prior distribution and under a certain class of informative prior distributions.

This article extends the Bayesian analysis of zero-numerator problems. In the current literature, applied Bayesian analyses often use noninformative priors, presumably to avoid injecting subjectivity in the conclusion. In zero-numerator problems we find that the posterior distribution of p and its 0.95 fractile are quite sensitive to the particular choice of noninformative prior from among those usually deemed reasonable for inference about a Bernoulli parameter. This, along with the fact that noninformative priors are likely to be highly unrealistic in zero-numerator problems, suggests that informative priors should play an especially important role in such problems. We also demonstrate how the focus on an upper 95% bound, often justified by a desire to be conservative, can wind up having the opposite effect in some situations. The general Bayesian approach has the advantage of providing an entire posterior distribution about p rather than a single bound. This distribution fully describes our uncertainty about p and can be used in a variety of different ways in different kinds of analyses.

Section 2 discusses the Rule of Three and some Bayesian results with noninformative priors, revisiting an example from Hanley and Lippman-Hand (1983). Section 3 looks at the role of informative priors in zero-numerator problems. Then, Section 4 examines the motivating example concerning the false-positive rate and considers what it means to be conservative in such settings. Some concluding comments are presented in the final section.

2. ZERO-NUMERATOR PROBLEMS WITH NONINFORMATIVE PRIORS

Following Hanley and Lippman-Hand (1983), suppose that the standard contrast agent used by radiologists over a long period has been shown to cause a serious reaction in about 15 of every 10,000 patients exposed to it. That is, the known risk with the old agent is 1.5 per 1,000. Suppose further that a new con-

trast agent is introduced. Soon afterward, a report of its use in 167 patients appears: no patient has had this reaction. What can we say about the risk associated with the new agent?

Using the Rule of Three, Hanley and Lippman-Hand (1983) would say that we should be 95% confident that the probability of a serious reaction with the new agent is at most $3/167 = 0.018$, or 1.8%. On the other hand, the probability that is most consistent with the data (the maximum likelihood estimate) is zero. If we had to use a number for this probability in making a decision about whether to use the old agent or the new agent, should we use either of these numbers?

A Bayesian approach to this problem requires the specification of a prior distribution. A convenient and reasonable prior used often in Bayesian analysis for Bernoulli processes is the family of beta distributions, with density

$$f(p) = p^{a-1}(1-p)^{b-1}/B(a,b) \quad \text{for} \quad 0 \le p \le 1, \quad (1)$$

where B represents the beta function. The mean is $a/(a+b)$. Beta distributions are quite exible (Johnson and Kotz 1970), capable of representing a wide range of prior distributions. They are easy to use for Bernoulli parameters because they allow closed-form updating with Bayes' rule. If we start with a Beta(a, b) prior as in (1) and observe r occurrences of the event in n trials, the posterior distribution is Beta$(a + r, b + n - r)$. The information about p represented by the Beta(a, b) distribution in (1) can be interpreted as equivalent to having seen a patients with a serious reaction in $a + b$ patients exposed to the new contrast agent.

A challenge in the Bayesian approach is to identify an appropriate prior. In many cases, Bayesians will suggest a noninformative, or diffuse, prior that is intended to represent little or no information about the parameter of interest and to have no material impact on the resulting posterior. This is often done in order to be conservative, and a noninformative prior is viewed by some as providing a more "objective" analysis than an informative prior. For example, Jovanovic and Levy (1997) suggested using beta priors with $a = 1$ and $b \ge 1$. The limiting case of $b = 1$ corresponds to a uniform distribution, which is often used as a noninformative prior for a Bernoulli parameter and which

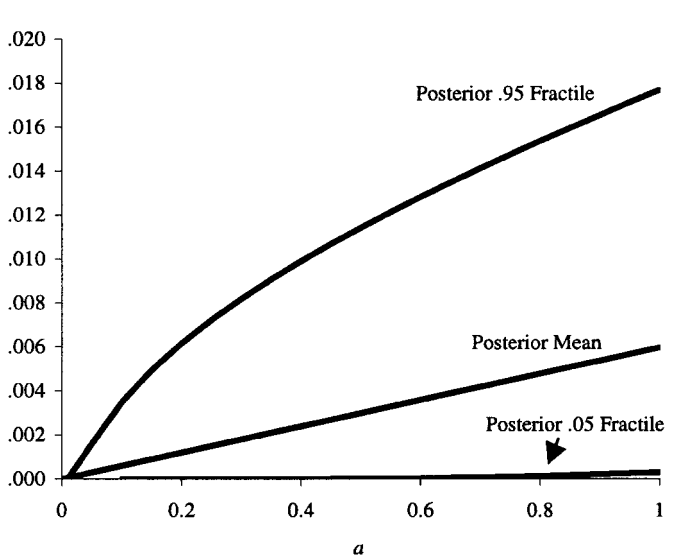

Figure 1. Analysis of p, the risk associated with the new contrast agent with Beta(a,a) priors.

Jovanovic and Levy (p. 138) called "a limiting and most conservative prior in this context." The posterior distribution is then Beta$(1, n + 1)$, and Jovanovic and Levy showed that approximating the 0.95 fractile of this distribution gives a Bayesian Rule of Three as $3/(n + 1)$. For the example involving the contrast agent, this gives $3/168 = 0.017857$, which differs from the Rule of Three's $3/167 = 0.017964$ only beyond the third decimal place.

There are, however, different noninformative priors that are used in Bernoulli settings. Unfortunately, they lead to very different results in zero-numerator problems. Commonly used noninformative priors include Beta$(0.5, 0.5)$ (a Jeffreys prior and a reference prior) and the improper Beta$(0, 0)$ in addition to the uniform Beta$(1, 1)$; see Geisser (1984) for discussion and review of the rationale for these priors. For $0 < a < 1$, the Beta(a, a) distribution is symmetric and U-shaped, with a mean of 0.5 and modes at zero and one. As a moves toward zero, more of the prior density is concentrated near zero and one. Varying a in the Beta(a, a) prior with $0 < a \le 1$ in our example involving the contrast agent $(n = 167)$, we see from Figure 1 that posterior 0.95 fractiles are arbitrarily close to zero at the low end and as high as 0.018 when $a = 1$; the posterior means vary similarly. Intuitively, as a decreases towards 0, more of the prior mass is concentrated near 0 and 1. Observing no occurrences of the reaction effectively wipes out the upper end of the U-shaped distribution and leaves a posterior that is more concentrated near zero. For smaller values of n the range of posterior 0.95 fractiles is much wider, with the low end remaining arbitrarily close to zero and the high end increasing (to 0.133 when $n = 20$ and 0.238 when $n = 10$, for example).

While in many applications the choice of a noninformative prior does not have a material impact on the results, this is clearly not the case in zero-numerator problems. When different schools of thought on what is an appropriate noninformative prior lead to very different results, it is clear that the analysis cannot be viewed as objective. With even a little bit of thought about the prior, in many zero-numerator problems it is clear that any reasonable set of prior judgments will *not* be consistent with commonly encountered noninformative prior distributions, which by symmetry place half of their probability on values above $p = 0.5$. For example, in the contrast-agent example, such values of p (and even somewhat lower values of p) would be extremely unlikely a priori.

3. ZERO-NUMERATOR PROBLEMS WITH INFORMATIVE PRIORS

For a Bernoulli process with a single, easy-to-interpret parameter p, it should be feasible to assess a prior distribution that reects an individual's prior judgments about p. Jovanovic and Levy (1997, p. 138) considered Beta(a, b) priors and recognized the need "to choose a and b in such a way as to ascertain agreement with the a priori knowledge about p." Then they go on to restrict their class of priors by setting $a = 1$ and requiring $b \ge 1$ on the grounds that the uniform prior is then a limiting case and that the prior mode is zero when we move away from the limiting case. This leads to another Bayesian Rule of Three (an approximation to the posterior 0.95 fractile of p) of $3/(n + b)$. The assumption that $a = 1$ is restrictive and we could argue just as easily for a different noninformative prior,

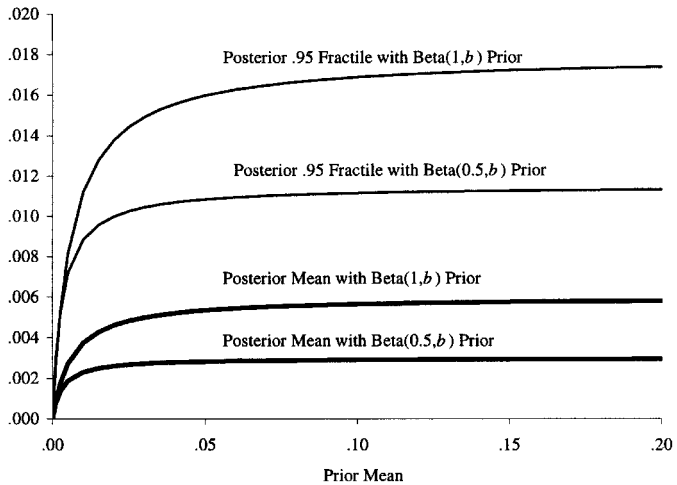

Figure 2. Analysis of p, the risk associated with the new contrast agent with Beta (1, b) and Beta(.5, b) priors.

such as Beta$(0.5, 0.5)$, as a limiting case. The assumption about a matters: in Figure 2, we see that Beta$(1, b)$ and Beta$(0.5, b)$ priors with the same prior mean can yield very different posterior means and 0.95 fractiles in the contrast-agent example.

Since the posterior is very sensitive to the choice of a in zero-numerator problems, the limitation to Beta$(1, b)$ priors seems arbitrary and unduly restrictive. If the goal is to allow the use of an informative prior, why should we restrict ourselves to Beta$(1, b)$ distributions? Given the ease of computing fractiles of beta distributions with today's computers (e.g., Microsoft Excel includes a function that generates any fractile of a beta distribution), simple approximations to 0.95 fractiles seem unnecessary. Allowing Beta(a, b) priors for any $a > 0$ and $b > 0$ provides a much richer class of possibilities with minimal effort. Moreover, with additional effort a full analysis is possible with *any* proper prior using modern Bayesian computational techniques. For a discussion of Monte Carlo methods in Bayesian computation, for example, see Chen, Shao, and Ibrahim (2000).

There are a variety of ways one might assess a prior distribution; see, for example, Spetzler and Staël von Holstein (1975), and Morgan and Henrion (1990). We have found it relatively easy to assess the prior mean (this is the prior probability that the event will occur on any given trial) together with fractiles (say, 5th and/or 95th fractiles) from the prior distribution. One can then fit a beta distribution or some other distribution to these assessments. For example, before the data from the 167 patients are seen in the example with a new contrast agent, a natural value to take for the prior mean might be 0.0015, the known risk with the old agent. Thus, we have $a/(a + b) = 0.0015$. Further suppose that we feel a priori that there is a 95% chance that p is less than 0.75%, five times the risk of the old agent. We can fit these assessments with a Beta$(0.042, 27.96)$ distribution, which implies that our prior uncertainty about p is equivalent to having seen .042 occurrences of the reaction in a sample of size $28\ (= a + b)$.

With this Beta$(0.042, 27.96)$ prior and a sample of $n = 167$ patients, of whom $r = 0$ had the reaction of concern, the posterior distribution is Beta$(.042, 194.96)$. Thus, after seeing the data the posterior mean of p is 0.022%; this is the predictive probability that the next patient will have a serious reaction to the contrast agent. The 0.95 fractile of the posterior distribution is 0.11%. Observing no patients with the reaction shifts the distribution of p to the left, as would be expected. What if we had chosen a different prior distribution? With a prior mean equal to the risk from the old agent, 0.0015, and prior strengths $(a + b)$ varying from 1 to 5,000, Figure 3 shows that the revised 0.95 fractile varies substantially but is never above $0.004(0.4\%)$; it is *always* much smaller than the Rule of Three value of 1.8%.

4. BEING CONSERVATIVE VERSUS BEING CANDID

The Rule of Three focuses on the upper 95% confidence bound for p in an attempt to provide a conservative estimate of p. However, what seems to be conservative in terms of p may not be conservative for some probability or other value that is a function of p. Our interest in zero-numerator problems was stimulated when Casey, the newborn daughter of one of the authors (Smith), was the first to test positive (after approximately 13,000 correct negatives) in an experimental screening program to test blood for certain genetic metabolic disorders. Casey's situation is described and analyzed by Smith and Winkler (1999) and Smith, Winkler, and Fryback (2000). The relevant point for this article is that the probability of concern was the probability that Casey actually had the metabolic disorder indicated by the test result. This can be found by applying Bayes' rule:

$$P(\mathrm{D}|+) = P(\mathrm{D})P(+|\mathrm{D})/[P(\mathrm{D})P(+|\mathrm{D}) + P(\mathrm{noD})P(+|\mathrm{noD})], \quad (2)$$

where $P(\mathrm{D})$ is the prevalence of the disorder, $P(+|\mathrm{D})$ is the correct-positive rate (in medical terms, the sensitivity), and $P(+|\mathrm{noD})$ is the false-positive rate (in medical terms, one minus the specificity). The data from the screening program pertain directly to $p = P(+|\mathrm{noD})$, the false-positive rate.

What would the Rule of Three say with respect to the false positive rate? Thinking in terms of having seen no false positives in the 13,000 noD newborns, we could say that we are 95% confident that the chance of a false positive is at most $3/13{,}000 = 0.000231$. Using this as the false-positive rate in applying (2) to Casey's situation, with estimates of $1/250{,}000$ for

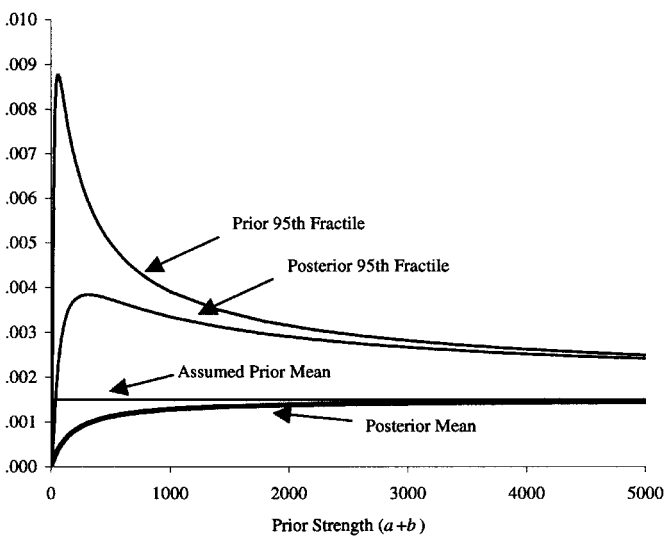

Figure 3. Analysis of p, the risk associated with the new contrast agent with Beta(a, b) priors and prior mean = .0015.

the prevalence and 0.999 for the sensitivity, yields a probability of 0.018 that Casey has the disorder given a positive test result. On the other hand, if we use the maximum likelihood estimate of zero for the false-positive rate, the probability that Casey has the disorder is one given a positive test result. The "conservative" estimate given by the upper 95% confidence bound of the Rule of Three thus gives a *lower* diagnostic probability than the maximum likelihood estimate. In what sense is this conservative?

To make a "conservative" estimate generally means to be cautious in the estimate, preferring to err on the high side rather than the low side for an estimate of a probability of a "bad" event. This situation does not allow an unambiguous way to hedge. The desire to be conservative in the estimate of the false-positive rate leads us to adopt a higher false-positive rate (e.g., the 95th percentile of the Rule of Three), which in turn leads us to *under*estimate the probability that Casey has the disorder. To be conservative about the diagnostic probability, we would want to hedge toward *lower* values of p, not higher values. Thus, the desire to be conservative may backfire in some scenarios.

The more important point, however, is that in order to make sound decisions we should be candid about what is known and report the entire distribution rather than an arbitrary "worst case" scenario. We should not try to be conservative with the inputs to the analysis, either in terms of the selection of a prior distribution or in the use of any measure(s) to summarize the posterior distribution. The aspects of the posterior distribution that are of interest may vary from problem to problem.

In Casey's situation, based on available information about the test and consultation with medical experts, we assessed a Beta$(1, 999)$ prior for p, implying a Beta$(1, 13{,}999)$ posterior. To calculate a diagnostic probability for Casey, the appropriate false-positive rate to use is the posterior mean of p, $1/14{,}000 = 0.0000714$. Based on all of the information available, this is the probability that Casey would have a false positive. Using this value in Bayes' rule gives a probability of 0.053 that Casey has the deficiency given a positive test result; this is almost three times as large as the probability of 0.018 based on a Rule-of-Three value for p. Because we are interested in the probability that Casey had a false-positive test result, the appropriate estimate of p is the posterior mean. In other contexts, other summary measures or the entire distribution may be required. For example, when deciding whether to gather more information by running more trials or whether to use the screening test on a particular population, we need the full distribution of p in order to find the distribution of the number of false positives in the additional trials or in the population. This latter distribution will be beta-binomial if p has a beta distribution.

5. CONCLUSIONS

The Bayesian approach represents uncertainty about a probability and allows us to update our information as new evidence becomes available. The importance of prior information is highlighted in the zero-numerator problem. For fixed n, different noninformative priors that are commonly used yield different results, implying that such priors are not really noninformative in this case. In zero-numerator situations, the assumptions implied by noninformative priors (e.g., half of the probability assigned to values of p greater than 0.5) seem inappropriate and give inappropriate results. Careful assessment of a prior distribution that reects all of the available information about p is always important, but it is especially important in zero-numerator situations, just as it is, for example, in multiparameter situations with identification problems.

The Bayesian approach also provides a full posterior distribution as well as any predictive distributions of interest, enabling us to understand how likely various risks are. If we want a particular summary measure (e.g., a point or interval estimate or a given fractile), we can easily find that from the distribution. In contrast, the Rule of Three gives only a single probability, taking an arbitrary cutoff value with 95% probability in an attempt to be conservative. The analysis of Casey's situation in Section 4 shows how such conservatism may backfire. When we plug the upper 95% bound from the Rule of Three into the formula for finding the risk of Casey having the disorder, the result is an unusually *low* risk because high false-positive rates translate into low risks of the deficiency. When probabilities such as the probability of a harmful side effect or a false-positive rate are used in further analysis for diagnostic or decision-making purposes, we should consider all of the available information and be candid about what is known about such probabilities.

[Received August 2000. Revised March 2001.]

REFERENCES

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000), *Monte Carlo Methods in Bayesian Computation*, New York: Springer-Verlag.

Geisser, S. (1984), "On Prior Distributions for Binary Trials" (with discussion), *The American Statistician*, 38, 244–251.

Hanley, J. A., and Lippman-Hand, A. (1983), "If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators," *Journal of the American Medical Association*, 249, 1743–1745.

Johnson, N. L., and Kotz, S. (1970), *Continuous Univariate Distributions—2*, New York: Wiley.

Jovanovic, B. D., and Levy, P. S. (1997), "A Look at the Rule of Three," *The American Statistician*, 51, 137–139.

Louis, T. A. (1981), "Confidence Intervals for a Binomial Parameter After Observing No Successes," *The American Statistician*, 35, 154.

Morgan, M. G., and Henrion, M. (1990), *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, New York: Cambridge University Press.

Smith, J. E., and Winkler, R. L. (1999), "Casey's Problem: Interpreting and Evaluating a New Test," *Interfaces*, 29, 63–76.

Smith, J. E., Winkler, R. L., and Fryback, D. G. (2000), "The First Positive: Computing Positive Predictive Value at the Extremes," *Annals of Internal Medicine*, 132, 804–809.

Spetzler, C. S., and Staël von Holstein, C.-A.S. (1975), "Probability Encoding in Decision Analysis," *Management Science*, 39, 176–190.