Letters to the Editor

WINKLER, R. L., SMITH, J. E., AND FRYBACK, D. G. (2002), "THE ROLE OF INFORMATIVE PRIORS IN ZERO-NUMERATOR PROBLEMS: BEING CONSERVATIVE VERSUS BEING CANDID," *THE AMERICAN STATISTICIAN*, 56, 1–4: COMMENTS BY BROWNE AND EDDINGS AND REPLY

BROWNE

Winkler, Smith, and Fryback (2002) provided an interesting look at Bayesian solutions to estimating the posterior 95% fractile of the binomial parameter (p), when no successes have been seen in n trials. They show that this approach provides much smaller upper bounds on p when the investigator is favorably disposed toward using Bayesian methods.

However, if the investigator is unwilling to consider a prior distribution on p, the investigator will want a 95% upper confidence limit on p (p_{95}). The exact of p_{95} is $1 - (.95)^{1/n}$, but relatively few nonstatisticians know this to be the correct calculation, or can carry it out without help. For ease of use, the Rule of Three (Hanley and Lippman-Hand 1983) was published, which says that $3/n$ provides an estimate of p_{95}. Howver, $3/n$ is a poor estimator of p_{95} for $n < 20$, overestimating p_{95} by 20% or more for $n < 9$, decreasing to 7.8% for $n = 20$. Even for $n = 80$, $3/n$ overestimates p_{95} by more than 2%. In parallel with the article by Winkler, Smith, and Fryback, we need improvements on $3/n$ for those who wish to use the frequentist approach.

To avoid confusion, we want a simple enhancement of the $3/n$ formula, and one that could be easily remembered and used away from a computer. The Bayesian Rule of Three of $3/(n + b)$ (Jovanovic and Levy 1997) seems a reasonable alternative form for an enhanced rule. By simple search methods, I found that $3/(n + 1.7)$ underestimates p_{95} by less than 1% for all $n > 3$ and overestimates p_{95} by 1.1% for $n = 3$. I would hope that this Modified Rule of Three could be disseminated as a preferred alternative to $3/n$.

Richard H. BROWNE
Texas Scottish Rite Hospital for Children

REFERENCES

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.

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

EDDINGS

In their discussion of interval estimates for a Bernoulli parameter p when no successes have been observed, Winkler, Smith, and Fryback neglect a simple, useful solution: support intervals based directly on the likelihood function. A $1/k$ support interval contains all parameter values whose relative likelihood is at least $1/k$; 8 and 32 are typical choices for k (Royall 1997). Given n independent Bernoulli trials, all failures, with common success probability p, the likelihood function is proportional to $(1 - p)^n$, and the $1/k$ support interval is $(0, 1 - k^{-1/n})$. The exact $100(1 - \alpha)\%$ confidence interval, $(0, 1 - \alpha^{1/n})$, always corresponds to a support interval with $k = 1/\alpha$ ($k = 20$ for a 95% confidence interval), but the Rule of Three confidence interval, $(0, 3/n)$, does not—it is asymptotically a support interval with $k = \exp(3) \approx 20.1$ but for small n (less than ten) can include points of very low likelihood relative to points near zero. Support intervals, unlike confidence intervals, are consistent with the likelihood principle and, unlike Bayesian intervals based on noninformative priors, do not involve the dubious representation of ignorance by probability distributions. Presentation of the entire likelihood function eliminates the somewhat arbitrary choice of k and allows readers to select their own intervals. I agree with the authors' pleas for realistic prior distributions when decisions must be made.

Wesley D. EDDINGS
The Johns Hopkins University

REFERENCES

Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*, Boca Raton, FL: Chapman & Hall/CRC.

REPLY TO BROWNE: STATISTICAL LITERACY, NOT SIMPLE RULES

We're glad that Browne is interested in zero-numerator problems. He inadvertently provides evidence for his point about a few people knowing the exact formula. The exact value of p_{95} is $1 - (0.05)^{1/n}$ for a one-sided interval, not $1 - (0.95)^{1/n}$ as he claims.

Everything we say in our article about the dangers of using the Rule of Three applies equally well to Browne's $3/(n + 1.7)$ rule. Browne is motivated by the desire to improve on the performance of the Rule of Three for small sample sizes, as small as $n = 3$. But as the sample size becomes smaller, prior information becomes even more important since there are so few data points to "speak for themselves." Indeed, small sample sizes provide not only the most compelling opportunity to think hard about the prior, but an obligation to do so.

More generally, we would like to take this opportunity to speak out against the mindless, uncritical use of simple formulas or rules. Instead of disseminating simple rules, we need to disseminate knowledge about statistical concepts and processes to help people think carefully and wisely about their statistical problems. We feel that the Bayesian framework is the best way to structure this thinking. As for calculations, many Bayesian calculations are not difficult; the calculations in our article were performed using simple spreadsheet formulas. Other cases can be much harder, but rapid advances in the use of simulation and other computer-based techniques to find posterior and predictive distributions are making Bayesian methods more accessible. With today's ready access to computers and user-friendly software, we do not need simple formulas that can be "easily remembered and used away from a computer." Instead, we need to join in working toward a new computer-enabled literacy in statistics, educating people to ask the right questions and providing processes and tools to answer them.

REPLY TO EDDINGS: DECISIONS USUALLY MUST BE MADE

We agree with Eddings that reporting the entire likelihood function is a good idea. However, that only goes part of the way toward addressing the question that is relevant in most applications: What is the posterior probability? Eddings agrees that realistic prior distributions are needed when decisions must be made, and they usually must be made. In principle, anyone can find a posterior distribution by assessing their own prior distribution and combining it with the likelihood function appropriately. But in practice, this may be a difficult assessment, and even given the prior, many people may not be able to do the math or to juggle the figures in their heads to get a posterior distribution. Thus, while we would like to see experimenters report results and likelihood functions, we also feel they should take a stand on the prior, report posterior results, and do some sensitivity

  DOI: 10.1198/000313002209

analysis to show the impact of assuming alternative reasonable priors on the posterior distribution and the resulting decisions.

Robert L. WINKLER
Duke University

BARKER, L., ROLKA, H., ROLKA, D., AND BROWN, C. (2001), "EQUIVALENCE TESTING FOR BINOMIAL RANDOM VARIABLES: WHICH TEST TO USE?" *THE AMERICAN STATISTICIAN*, 55, 279–287: COMMENT BY MARTÍN ANDRÉS AND HERRANZ TEJEDOR AND REPLY

Barker et al. undertook the problem of performing an asymptotic test to prove that two binomial proportions are equivalent, that is, "practically equal." Specifically, the aim is to contrast $H_0 : |p_x - p_y| \geq \Delta$ versus $H_a : |p_x - p_y| < \Delta$ (where Δ is a positive number given previously) starting from the two independent random variables $X \sim \text{binomial}(n_x, p_x)$ and $Y \sim \text{binomial}(n_y, p_y)$. To this end, the authors propose two groups of procedures: one based on the two one-sided test (TOST)—with 6 versions—the other based on the two-tailed test—with two versions. Unfortunately, both groups of tests contain omissions and/or important defects.

The TOST test is based on performing two one-tailed tests $H_{01} : p_x - p_y \geq +\Delta$ versus $H_{a1} : p_x - p_y < +\Delta$ and $H_{02} : p_x - p_y \leq -\Delta$ versus $H_{a2} : p_x - p_y > -\Delta$. If the p value of each test is P_i, then the p value for the equivalence test is $P = \max\{P_1, P_2\}$. The authors prefer the (equivalent) approach of obtaining a two-tailed approximate confidence interval (to the error 2α) for $p_x - p_y$ and declaring the equivalence (to the error α) if this is completely contained in the interval $[-\Delta; +\Delta]$. The six versions of the TOST test that they propose are based on six different ways of obtaining this confidence interval. Surprisingly, in the article the most competitive methods and the historical origin of the problem are not mentioned. Thus:

A. The problem was originally posed by Dunnett and Gent (1977, 1988) (D&G) and was adequately solved by Johnson (1988) in the format of the TOST test. The possible solutions are based on statistics of type χ^2 or type z (score statistics). Thus, under the hypothesis that $p_x - p_y = \delta$, the statistic

$$Z(\delta) = \frac{\hat{p}_x - \hat{p}_y - \delta}{\sqrt{\frac{p(1-p)}{n_y} + \frac{(p+\delta)(1-p-\delta)}{n_x}}} \tag{1}$$

is distributed as a normal standard z. In this expression $\hat{p}_x = X/n_x, \hat{p}_y = Y/n_y$ and p is a nuisance parameter which must be estimated. D&G proposed the conditional estimator $\tilde{p} = (X + Y - n_x\delta)/(n_x + n_y)$ which produces the statistic $Z_1(\delta)$. The p value using the TOST test is therefore:

$$\max\left[P\{z \leq Z_1(+\Delta)\}; P\{z \geq Z_1(-\Delta)\}\right]. \tag{2}$$

B. It is better to substitute p for its estimator $\hat{p}$ of maximum likelihood (Roebruck and Kühn 1995) and this gives rise to the statistic $Z_2(\delta)$ which today is the basis of all the published literature on the equivalence of two proportions. The p value is obtained as in expression (2)—with Z_2 in place of Z_1—but this procedure is not mentioned by the authors. The estimator $\hat{p}$ was proposed by Miettinen and Nurminen (1985)—who gave its explicit solution—and it is the solution of the cubic equation $L'_p(p, \delta) = 0$, where

$$L'_p(p, \delta) = \frac{\partial L(p, \delta)}{\partial p} = \frac{n_y(\hat{p}_y - p)}{p(1-p)} + \frac{n_x(\hat{p}_x - p - \delta)}{(p+\delta)(1-p-\delta)}, \tag{3}$$

and $L(p, \delta)$ is the logarithm of the likelihood: $L(p, \delta) \propto Y \times \ln p + (n_y - Y) \times \ln(1-p) + X \times \ln(p+\delta) + (n_x - X) \times \ln(1-p-\delta)$. So, for the example given by the authors, ($X = 980, Y = 1{,}100, n_x = n_y = 2{,}000$, and $\Delta = 0.10$) $Z_2(+\Delta) = -10.178$ in $\hat{p} = 0.4694$ and $Z_2(-\Delta) = 2.545$ in $\hat{p} = 0.5698$, and so the p value of the TOST test is 0.55% using expression (2). (It is possible to make a correction for continuity to expression (1), not contemplated here.)

C. From the perspective of the confidence intervals (which is what the authors analyze), it is quite usual to obtain these by inverting the appropriate hypothesis test (Agresti and Min 2001). Indeed, Miettinen and Nurminen (1985) proposed inverting expression (1), and this solution has not been investigated by Barker et al. either. This procedure is presumably the most suitable, for Agresti and Min have shown that the ideal exact confidence interval is the one based on the order given by the statistic $Z_2(\delta)$. So, for the data in the previous example, the equation $Z_2(\delta) = \pm 1.645$ yields the interval $\delta \in [-0.086, -0.033] \subset [-0.10, +0.10]$ and so H_0 should be rejected to the error $\alpha = 5\%$.

With regard to the two-tailed tests proposed by Barker et al., the following observations should be made:

D. The test of maximum likelihood (LRT) requires the calculation of the maximum of $L(p, \delta)$ and of $L(p, \delta | |\delta| \geq \Delta)$. The first is always $L(\hat{p}_y, \hat{p}_x - \hat{p}_y)$ as the authors indicate. In order to obtain the second, the authors offer a computer program. But this is not necessary: for the values $|\hat{p}_x - \hat{p}_y| < \Delta$:

$$\max L(p, \delta | |\delta| \geq \Delta) = \max\{L(\hat{p}, +\Delta), L(\hat{p}, -\Delta)\} \tag{4}$$

and for the values $|\hat{p}_x - \hat{p}_y| \geq \Delta$:

$$\max L(p, \delta | |\delta| \geq \Delta) = L(\hat{p}_y, \hat{p}_x - \hat{p}_y) \tag{5}$$

which simplifies the problem enormously. In order to see this, bear in mind that because $\hat{p}$ is the estimator of maximum likelihood, then, for each fixed value δ, L reaches the maximum in $L(\hat{p}, \delta)$. Because $dL(\hat{p}, \delta)/d\delta = \partial L(\hat{p}, \delta)/\partial\delta$—since $\partial L(p, \delta)/\partial p = 0$ in $p = \hat{p}$—then $dL(\hat{p}, \delta)/d\delta = n_x(\hat{p}_x - \delta - \hat{p})/[(\hat{p} + \delta)(1 - \hat{p} - \delta)]$. This indicates that L increases (decreases) in δ when $\hat{p}_x - \delta - p > 0$ ($\hat{p}_x - \delta - p < 0$). But, using the expression (3), $\hat{p}_x - \hat{p} - \delta$ and $\hat{p}_y - \hat{p}$ have to have opposing signs in order for $L'_p(\hat{p}_x, \delta) = 0$, and so L increases (decreases) in δ when $\hat{p}_x - \hat{p}_y > \delta$ ($\hat{p}_x - \hat{p}_y < \delta$). Hence the expressions (4) and (5). Expression (5) implies that when $|\hat{p}_x - \hat{p}_y| \geq \Delta$, the test LRT can never have significance, and this is in keeping with the inferential logic: if $|\hat{p}_x - \hat{p}_y| \geq \Delta$, one cannot conclude that $|p_x - p_y| < \Delta$. Consequently, the conclusion which the authors obtain in their final example is wrong. In it (in error) they worked with the values $X = 980$ and $Y = 1{,}200$, with the result that $|\hat{p}_x - \hat{p}_y| = 0.11 > 0.10$ and the test should not be significant ($-2\ln\lambda = 0$ and not -4.00 as the authors indicated). The authors commited the same inferential error in several of the examples proposed. Moreover, the statistic LRT is $2\ln\lambda$, not $-2\ln\lambda$ (the way the authors defined λ). Based on all the above, the test LRT consists in comparing with $\chi^2_{1\text{df}}$ the value of

$$\chi^2 = -2 \times \ln\frac{\max\{f(\hat{p}, \hat{p} - \Delta), f(\hat{p}, \hat{p} + \Delta)\}}{f(\hat{p}_y, \hat{p}_x)}$$
$$\text{where} \quad f(p_1, p_2) = p_1^y(1-p_1)^{n_y - y} p_2^x(1-p_2)^{n_x - x} \tag{6}$$

if $|\hat{p}_x - \hat{p}_y| < \Delta$. Otherwise $\chi^2 = 0$.

E. Other reasonable two-tailed asymptotic tests are those based on the statistics $Z_1(\delta)$ and $Z_2(\delta)$, and the authors also failed to give these. Mau (1988) obtained the p value using the criterion $Z_1(\delta)$:

$$|P\{z \leq Z_1(-\Delta)\} - P\{z \leq -Z_1(+\Delta)\}| \tag{7}$$

and the same can be done using the criterion $Z_2(\delta)$ (Mau did not take into consideration the need to put the absolute value into the above expression). The authors of this letter have tested (in an article submitted to press) the following more suitable expression

$$\max_{i=1,2} = P\left\{-\frac{|\hat{p}_x - \hat{p}_y| - \Delta}{s_i} \leq z \leq +\frac{|\hat{p}_x - \hat{p}_y| + \Delta}{s_i}\right\}, \tag{8}$$

where s_1 and s_2 are the denominator of the expression (1) for $\delta = -\Delta$ and $\delta = +\Delta$, respectively. So, for the example cited in B, $|\hat{p}_x - \hat{p}_y| = 0.06$ and $s_1 = s_2 = 0.01572$. In this way the p value will be $P(2.545 \leq z \leq 10.178) = 0.55\%$ just as in section B (although generally this is not so). The last two expressions assume that $|\hat{p}_x - \hat{p}_y| < \Delta$ and they are capable of being assigned a correction for continuity.

A. MARTÍN ANDRÉS
Universidad de Granada

I. HERRANZ TEJEDOR
Universidad Complutense

ACKNOWLEDGMENT

This research was supported by the Dirección General de Investigación, Spain, grant BFM2000-1472.