Letters to the Editor

MCCULLOUGH, B. D. (1999), "ASSESSING THE RELIABILITY OF STATISTICAL SOFTWARE: PART II," *THE AMERICAN STATISTICIAN*, 53, 149–159: COMMENT BY DIELMAN

McCullough (1998) provided a framework for evaluating the numerical accuracy of statistical software routines. This evaluation was based on a number of datasets with certified parameter values provided by the National Institute of Standards. McCullough (1999) then used this framework and the datasets to compare the numerical accuracy of SAS, SPSS, and S-Plus. I would like to point out a problem in the latter article with the interpretation of the ANOVA results for S-Plus.

Table 1 (p. 155) of the article should appear as follows:

Table 1. S-Plus LRE's for SmnLsg Problems

	d		
n	*1*	*7*	*13*
189	SmnLsg01 (14.5)	SmnLsg04 (10.4)	SmnLsg07 (4.6)
1809	SmnLsg02 (14.3)	SmnLsg05 (10.2)	SmnLsg08 (2.7)
18009	SmnLsg03 (12.9)	SmnLsg06 (10.2)	SmnLsg09 (0)

NOTE: *d*: number of constant leading digits; *n*: number of observations

The description of the datasets used in the table is also incorrect. The dataset SmnLsg04 has the same treatment variable as SmnLsg01, and the response variable is the same except six zeros (not nine as stated in the article) have been inserted between the 1 and decimal, so its mean is 1,000,000.4 exactly.

Because of the incorrect table used in the article, some of the conclusions do not follow. McCullough stated "Some readers may be surprised to see that in the present example, the number of observations has a more deleterious effect than the number of leading digits. This is shown clearly in Table 1." As can be seen from the corrected Table 1, the increase in sample size has a rather minor effect on the reported LRE's. It is the number of leading digits that causes the deterioration in the quality of the computed values. To correct McCullough's conclusion, the effect of increasing the number of leading digits (look across any row), is much more serious than increasing the number of observations (look down any column). This degradation is likely due to cancellation error. Obviously, McCullough inadvertently transposed the rows and columns, and so reached erroneous conclusions.

I enjoyed the articles by McCullough and believe that this type of examination of statistical packages is of importance. However, the interpretation of the results discussed in this note is important in judging what contributes to the degradation of ANOVA estimates in various computer packages.

Terry E. DIELMAN

M. J. Neeley School of Business
Texas Christian University
Fort Worth, TX 76129

REFERENCES

McCullough, B. D. (1998), "Assessing the Reliability of Statistical Software: Part I," *The American Statistician*, 52, 358–366.

——— (1999), "Assessing the Reliability of Statistical Software: Part II," *The American Statistician*, 53, 149–159.

LEHR, R. G. (2000), LETTER TO THE EDITOR, *THE AMERICAN STATISTICIAN*, 54, 325: COMMENT BY BERGER AND DOI AND REPLY

Lehr (2000) proposed an "exact conditional" confidence interval for the difference between two binomial proportions when the data are collected as matched pairs. Unfortunately, his proposed interval is not "exact." Its actual coverage probability can be below 50% when the nominal confidence coefficient is 95%. However, an easily computed interval that does guarantee the nominal $1-\alpha$ coverage probability is available for both matched pairs and independent data.

Lehr did not carefully distinguish between statistics and parameters. For example, he discussed a "CI for $p_1 - p_2$" even though $p_1 - p_2$ is defined to be a statistic. To clarify this, we will use Lehr's notation for the observed count data (statistics), namely,

		Pretreatment		
		Success	Failure	
Post-	Success	a	b	n_1
treatment	Failure	c	d	
		n_2		n,

and for the observed proportions, $p_1 = n_1/n = (a+b)/n$ and $p_2 = n_2/n = (a+c)/n$.

We will use the following notation for the parameters in this four-cell multinomial model:

		Pretreatment		
		Success	Failure	
Post-	Success	π_a	π_b	π_1
treatment	Failure	π_c	π_d	
		π_2		1.

So, in this notation, Lehr discussed a CI for $\pi_1 - \pi_2 = \pi_b - \pi_c$.

Lehr first described an exact $100(1-\alpha)\%$ conditional CI for $\pi_b/(\pi_b+\pi_c)$. This interval, denoted by [LL, UL] {(2) in Lehr's letter}, is the well-known Clopper–Pearson (1934) CI. It is exact because the conditional distribution of b given $b+c$ is binomial $(b+c, \pi_b/(\pi_b+\pi_c))$. Then, without further explanation, Lehr stated that an exact conditional CI for $\pi_1 - \pi_2$ is given by $[(2\text{LL}-1)(b+c)/n, (2\text{UL}-1)(b+c)/n]$.

Unfortunately, the Lehr interval does not have confidence coefficient $100(1-\alpha)\%$. We will compute the coverage probability at a certain type of parameter point and show it is much less than $1-\alpha$. Consider a parameter value with $\pi_c = 0$. Then, $P(c=0) = 1$ and, for the Clopper–Pearson interval, if $c = 0$, then $\text{UL} = 1$. The upper endpoint of the Lehr interval with $\text{UL} = 1$ and $c = 0$ is b/n. In order for the Lehr interval to include the true value of

$$\pi_1 - \pi_2 = \pi_b - \pi_c = \pi_b - 0 = \pi_b$$

the upper endpoint b/n must be at least π_b. Suppose, for example, $n = 6$ and $\pi_b = .7$. Then, the probability that the Lehr interval includes the true parameter value is bounded above by

$$P(\pi_b \le b/n) = P(b = 5 \quad \text{or} \quad 6) = .42$$

because, $b \sim \text{binomial}(6, \pi_b)$. This is, of course, much less than the usual confidence levels of 95% or 99%. In fact, for any value of $0 < \pi_b < 1$, application of the central limit theorem yields $\lim_{n\to\infty} P(\pi_b \le b/n = 1/2)$. Thus, for any parameter point with $\pi_c = 0$ and $0 < \pi_b < 1$, the coverage probability of the Lehr interval is bounded above by $1/2$ as $n \to \infty$. The same is true if $\pi_b = 0$ and $0 < \pi_c < 1$.

The purpose of Lehr's letter was to demonstrate that in the case of "correlated proportions," it is possible to construct an exact CI for $\pi_1 - \pi_2$. Unfortunately, Lehr's interval does not achieve this goal. Lehr also stated that it is not possible to construct such an exact CI for the difference of independent proportions. But, actually, it is possible to construct exact CI's in both cases, as we shall now explain.

First, consider the case of correlated proportions considered above. Let $[\mathrm{LL}_1, \mathrm{UL}_1]$ be a $100(1-\alpha/2)\%$ Clopper–Pearson CI for π_1 based on $n_1 \sim \mathrm{binomial}(n, \pi_1)$. Similarly, let $[\mathrm{LL}_2, \mathrm{UL}_2]$ be a $100(1-\alpha/2)\%$ Clopper–Pearson CI for π_2 based on $n_2 \sim \mathrm{binomial}(n, \pi_2)$. Let R denote the rectangle $[\mathrm{LL}_1, \mathrm{UL}_1] \times [\mathrm{LL}_2, \mathrm{UL}_2]$. By the Bonferroni Inequality, $P((\pi_1, \pi_2) \in R) \geq 1 - \alpha$. Whenever $(\pi_1, \pi_2) \in R$, then

$$\mathrm{LL}_1 - \mathrm{UL}_2 \leq \pi_1 - \pi_2 \leq \mathrm{UL}_1 - \mathrm{LL}_2.$$

Thus, $[\mathrm{LL}_1 - \mathrm{UL}_2, \mathrm{UL}_1 - \mathrm{LL}_2]$ is a CI for $\pi_1 - \pi_2$ with coverage probability bounded below by $1 - \alpha$.

For the case of independent proportions, let $X_1 \sim \mathrm{binomial}(m_1, \pi_1)$ and $X_2 \sim \mathrm{binomial}(m_2, \pi_2)$ be independent observations. Let $[\mathrm{LL}_i, \mathrm{UL}_i]$ be a $\sqrt{1-\alpha}$ Clopper–Pearson CI for π_i based on X_i. Let R denote the rectangle $[\mathrm{LL}_1, \mathrm{UL}_1] \times [\mathrm{LL}_2, \mathrm{UL}_2]$. By the independence, $P((\pi_1, \pi_2) \in R) \geq (\sqrt{1-\alpha})^2 = 1 - \alpha$. So, by the same argument as in the previous paragraph, $[\mathrm{LL}_1 - \mathrm{UL}_2, \mathrm{UL}_1 - \mathrm{LL}_2]$ is a CI for $\pi_1 - \pi_2$ with coverage probability bounded below by $1 - \alpha$. Many authors, such as Santner and Snell (1980), Chan and Zhang (1999), and Agresti and Min (2001), have described exact CI's for the difference between two independent proportions. Many might be more precise than the one we just described, but they require much more computation. Some users may prefer CI's with closed-form expressions like the intervals in this and the previous paragraph.

Roger L. BERGER

Jimmy DOI

Department of Statistics
North Carolina State University
Box 8203
Raleigh, NC 27896-8203
berger@stat.ncsu.edu

REFERENCES

Agresti, A., and Min, Y. (2001), "On Small-Sample Confidence Intervals for Parameters in Discrete Distribution," *Biometrics*, 57, 963–971.

Chan, I. S. F., and Zhang, Z. (1999), "Test-Based Exact Confidence Intervals for the Difference of Two Binomial Proportions," *Biometrics*, 55, 1202–1209.

Clopper, C. J., and Pearson, E. S. (1934), "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26, 404–413.

Santner, T. J., and Snell, M. K. (1980), "Small-Sample Confidence Intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 Contingency Tables," *Journal of the American Statistical Association*, 75, 386–394.

REPLY

I thank Berger and Doi for their comments on my letter to the editor that appeared in *The American Statistician*. It was encouraging to hear that they agree with the main point of my letter (that "exact" confidence intervals for the given situation exist); they question this particular CI used as an example.

In response to their first criticism, I certainly agree that a differentiation between parameters and statistics is necessary, and while the context of the usage of $(p_1 - p_2)$ would make this clear to most readers, the use of different symbols would have been preferable even in this informal communication.

Apparently much of the remainder of the criticism by Berger and Doi hinges on the term "exact." It is interesting to note that much of the literature dealing with exact inference puts the term "exact" in quotes. There is a good reason for this. As often as the term is used, it is rarely if ever defined. Authors use it almost as an undefined term, and describe it mostly by implication. Many authors couch their statements by using phrases like "exact in a certain sense." Berger and Doi do not do this in their response, but write as if there is a single, universally accepted definition (which they never state explicitly). They also do not differentiate between conditional, unconditional, and test-based confidence intervals.

My statements in the letter referred to conditional exact inference, which I think is generally assumed by most when the term "exact" is used without the qualifiers "unconditional" or "test-based." Berger and Doi question the statement that *no exact intervals exist for the difference of two independent binomial proportions*. This statement was based on an excellent survey article by Agresti (1992). The following excerpt is a direct quote:

> It is not possible to construct "exact" confidence intervals for association measures that are not functions of the odds ratio . . . . For instance, consider estimation of the difference of probabilities $\delta = \pi_1 - \pi_2$ for independent binomials. The joint sampling distribution can be expressed in terms δ and π_1, for instance, but conditioning on the marginal totals does not eliminate π_1

To support their claim that the disputed CI is not exact, Berger and Doi choose a particularly pathological example: they force a structural zero into the distribution. In the context of the example, this causes one treatment to never succeed unless the other succeeds.

The confidence interval used as an illustration in my original letter to *TAS* may not be "exact" in every sense. The question is: *is it exact in any sense*? It is nonasymptotic, as it is based solely on a well-known exact confidence interval and a back-projection into the parameter space of interest. Inversion of this confidence interval results in a hypothesis test which is functionally equivalent to an "exact McNemar" test. It was on the basis of these two qualities that I represented the interval as "exact." I would be interested to hear opinions from other statisticians as to whether or not they feel the term "exact" is appropriate for this interval. At least one recently published survey article has classified an equivalent form of this CI as "exact" (Newcombe 1998).

Whichever term is applied to this particular confidence interval, it may be the most appropriate CI display in conjunction with the results of the "exact" McNemar test discussed in my original letter. I would be interested in hearing opinions on this, also.

Robert G. LEHR

Berlex Laboratories
340 Changebridge Road
PO Box 1000
Montville, NJ 07045-1000

REFERENCES

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 135.

Irony, T., Pereira, C., and Tiwari, R. (2000), "Analysis of Opinion Swing: Comparison of Two Correlated Proportions," *The American Statistician*, 54, 57–62.

Newcombe, R. G. (1998), "Improved Confidence Intervals for the Difference Between Binomial Proportions Based on Paired Data," *Statistics in Medicine*, 17, 2635–2650.

LONG, J. S., AND ERVIN, L. H. (2000), "USING HETEROSCEDASTICITY CONSISTENT STANDARD ERRORS IN THE LINEAR REGRESSION MODEL," *THE AMERICAN STATISTICIAN*, 54, 217–224: COMMENT BY WILCOX

Long and Ervin (2000) made the important point that the conventional method for computing a confidence interval for the parameters of the usual linear regression model can be highly inaccurate when the error term is heteroscedastic. They went on to recommend a particular method for dealing with this problem and presented simulations in support of its use. My goal is to point out situations where, when using the conventional method, problems with probability coverage are even worse than indicated. I also describe a situation where their recommended method can be unsatisfactory. An alternate method that gives better probability coverage in this special case is indicated.

The estimated standard error of the least squares estimators recommended by Long and Ervin is

$$\mathrm{HC3} = (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathrm{diag}\left[\frac{e_i^2}{(1-h_{ii})^2}\right]\mathbf{X}'(\mathbf{X}'\mathbf{X}),$$

where $e_i, i = 1, \ldots, N$ are the usual residuals and

$$h_{ii} = \mathbf{x}_i\mathbf{X}'\mathbf{X}\mathbf{x}_i'.$$

It is known that $1/N \leq h_{ii} \leq 1$. Also, a single unusual point among the rows of $\mathbf{X}$ can deate all h_{ii} values, causing them to be close to their lower bound. This suggests that in simple regression, for example, if the predictors $x_1, \ldots, x_n$ are sampled from a sufficiently heavy tailed distribution, probability coverage might be unsatisfactory for certain types of heteroscedasticity, even when N is moderately large, particularly when the error term is normal or light-tailed.

Following the notation used by Long and Ervin, and restricting attention to simple regression, consider the following four types of heteroscedasticity: $e_i = e_i^*, e_i = \sqrt{|x|}e_i^*, e_i = |x|e_i^*, e_i = (1 + 2/(|x| + 1))e_i^*$, where e_i^* is to be specified. These will be called variance structures (VS) 0, 1, 2, and 3. VS 0 and 1 were considered by Long and Ervin. Wilcox (1996) compared six methods for computing confidence intervals where e_i^* was taken and have a g-and-h distribution. That is,

$$e_i^* = \frac{\exp(gZ) - 1}{g}\exp(hZ^2/2),$$

where g and h determine skewness and kurtosis, respectively, and Z has a standard normal distribution (for details see Hoaglin 1985). So when $g = h = 0$, e_i^* is standard normal. Here, for brevity, attention is restricted to $g = 0$ and $h = .5$, in which case this last equation is taken to be

$$e_i^* = Z\exp(hZ^2/2),$$

which has a symmetric, heavy-tailed distribution.

Among the methods compared by Wilcox, four were bootstrap methods, one was derived by Nanayakkara and Cressie (1991)—which like HC3 uses an estimate of the standard error that stems from MacKinnon and White (1985)—and the sixth was the conventional method. Only one of these six methods gave reasonably accurate probability coverage over the situations considered in simulations: an adjusted percentile bootstrap procedure. Basically, obtain a bootstrap sample by resampling, with replacement, N vectors of observations from $(y_1, x_1), \ldots, (y_N, x_N)$. So heteroscedasticity is addressed based on results in Wu (1986). Let b_1^* be the resulting bootstrap estimate of the slope. Repeat this process 599 times yielding $b_{11}^*, \ldots, b_{1,599}^*$ and let $b_{1(1)}^* \leq \cdots b_{1(599)}^*$ be the bootstrap values written in ascending order. The .95 confidence for the slope was taken to be $(\hat{\beta}_{(a)}^*, \hat{\beta}_{(c)}^*)$, where $N < 40$, $a = 7$, and $c = 593$; for $40 \leq N < 80$, $a = 8$, and $c = 592$; for $80 \leq N < 180$, $a = 11$, and $c = 588$; for $180 \leq N < 250$, $a = 14$, and $c = 585$; while for $N \geq 250$, $a = 15$, and $c = 584$. So for $N \geq 250$ the conventional percentile bootstrap method is used.

Table 1 shows some values of $\hat{\alpha}$, an estimate (based on a simulation with 1,000 replications) of one minus the probability coverage when using the method recommended by Long and Ervin (method LE), the adjusted percentile bootstrap method (method PERAD), and the conventional (homoscedastic) approach (method OLS). Notice that the conventional method is disastrous for VS 1 and 2. Among all the situations considered by Wilcox, generally HC3 gave reasonable results except when x is heavy tailed and the error term has a light-tailed or normal distribution. For VS 1 and the situation considered in Table 1, increasing N to 60, $\hat{\alpha}$ drops to .102 when using method LE.

Table 1. Values of $\hat{\alpha}$, α = .05, N = 20

x		ϵ			Method		
g	h	g	h	VS	LE	OLS	PERAD
0.0	0.5	0.0	0.0	0	0.073	0.055	0.049
				1	0.120	0.327	0.066
				2	0.110	0.523	0.078
				3	0.044	0.005	0.026

In summary, this letter underscores and hopefully supports the main message made by Long and Ervin: Heteroscedasticity is a serious problem that can be addressed. The simple method they recommend appears to perform well except when the predictors have a heavy tailed distribution and simultaneously the error term is normal or light tailed.

Rand R. WILCOX
Department of Psychology
University of Southern California
Los Angeles, CA 90089

REFERENCES

Hoaglin, D. C. (1985), "Summarizing Shape Numerically: The g-and-h Distributions," in *Exploring Data Tables, Trends and Shapes*, eds. D. Hoaglin, F. Mosteller, and J. Tukey, New York: Wiley.

Long, J. S., and Ervin, L. H. (2000), "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, 54, 217–224.

MacKinnon, J. G., and White, H. (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators With Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325.

Nanayakkara, N., and Cressie, N. (1991), "Robustness to Unequal Scale and Other Departures From the Classical Linear Model," in *Directions in Robust Statistics and Diagnostics, Part II*, eds. W. Stahel and S. Weisberg, New York: Springer-Verlag, pp. 65–113.

Wilcox, R. R. (1996), "Confidence Intervals for the Slope of a Regression Line When the Error Terms has Nonconstant Variance," *Computational Statistics and Data Analysis*, 22, 89–98.

Wu, C. J. F. (1986), "Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis," *The Annals of Statistics*, 14, 1261–1295.

WENDELL, J. P., AND SCHMEE, J. (2001), "LIKELIHOOD CONFIDENCE INTERVALS FOR PROPORTIONS IN FINITE POPULATIONS," *THE AMERICAN STATISTICIAN*, 55, 55–61: COMMENT BY VOS AND REPLY

Wendell and Schmee claimed confidence regions obtained from their L method are likelihood based; that is, regions where "values in the confidence region have higher likelihood than those outside." (Cox and Hinkley 1974, p. 218; also quoted in Wendell and Schmee.)

In fact, confidence regions obtained from their L method are not likelihood-based. This is easily seen from their example in Section 3.3 and Table 2. In this example there is a population of size $N = 100$ and a sample of size $n = 30$. The parameter is M, the number in the population with a given characteristic, and the number in the sample having this characteristic is y. If $y = 0$, the L method 90% confidence region for M is $\{0, 1, 2, 3, 4, 5, 6, 8\}$. From the first column in Table 2 we see that the likelihood of $M = 7$ is greater than $M = 8$ and so this region is not likelihood-based.

This L method confidence region is unusual in that it is not an interval. As the authors note, usually the L method produces an interval. However, even if the confidence region is an interval it need not be likelihood-based. Using the same example as before, if $y = 3$ the L method confidence interval for M is $\{5, 6, \ldots, 21\}$. The likelihood evaluated at $M = 4$ is greater than the likelihood evaluated at $M = 20$ and $M = 21$, and so is not likelihood-based.

The fact that the L method does not produce likelihood-based confidence intervals does not affect the authors' claims about their length. It is clear from the construction in Section 3.1 that the L method is designed to closely approximate the nominal level. Naturally, this tends to shorten the length of intervals compared to other conservative methods. Perhaps the L in "L method" should stand for length or level.

Unfortunately, the emphasis on length and level leads to strange properties of the family of tests derived from L method intervals. (The L method intervals determine a test of $H_0 : M = M_0$ by rejecting H_0 if and only if M_0 is not in the confidence region.) Since the confidence intervals are constructed using Q_M, the acceptance region of the above test, L method intervals and L method tests are complementary descriptions of the same inference procedure.

We return once again to the example of Section 3.3. If $y = 3$ and $\alpha = .10$, the L method rejects $H_0 : M = 4$ but accepts $H_0 : M = 21$ even though the probability of observing $Y = 3$ is much greater if $M = 4$ than if $M = 21$. In particular, $P(Y = 3|M = 4) = .0725$ while $P(Y = 3|M = 21) = .0461$. Of course, this is a direct result of the fact that the L method is not likelihood-based. Another means of comparing these hypotheses is in terms of p values. Specifically, we compare these hypotheses in terms of the probability of observing a value as extreme or more extreme than the observation $y = 3$. That is, for $H_0 : M = 4$ the one-sided p value is $P(Y \geq 3|M = 4) = .0795$ while for $H_0 : M = 21$ the one-sided p value is $P(Y \leq 3|M = 21) = .0622$. The L method rejects the hypothesis with the larger p value while accepting the hypothesis with the smaller p value.

The reason the L method behaves in this manner is that it produces two-sided tests. The difficulties are seen even more clearly in the situation where $y = 0 (N = 100, n = 30$, and M is the parameter as before). Since $N = 100$ is given, an equivalent parameterization is the population proportion $p = M/N$. For $\alpha = .10$ and $\hat{p} = y/n = 0$, the L method accepts $p = .06$ and $p = .08$ but rejects $p = .07$. This odd behavior does not occur with a one-sided test because the probability of $\hat{p} = 0$, that is, the p value, decreases as p increases.

The authors acknowledge that this behavior, which they term gapping, is somewhat problematic, but it is unclear whether they recognize it as related to the two-sided nature of their tests. Using the L method, the p value for testing $H_0 : p = .07$ is the probability that $Y \in \{0, 5, 6, 7\}$ while the p value for testing $H_0 : p = .08$ is the probability that $Y \in \{0, 5, 6, 7, 8\}$. The p values are .0991 and .1012, respectively. The second p value is larger because there is more probability in the tail where the data did not occur; that is, $\{5, 6, 7, 8\}$. The problem with the L method test (and confidence interval) is that it ignores the direction of the departure of the data from the null value. According to Cox and Hinkley (1974, p. 106),

> [I]t is essential to consider the direction of departure in interpreting the result of the significance test; that is, to conclude that there is evidence that $\theta > \theta_0$ when in fact $\theta < \theta_0$ would normally be very misleading . . .. This argument shows that we are really involved with two tests, one to examine the possibility $\theta > \theta_0$, the other $\theta < \theta_0$.

A simple and natural way to make the L method tests reect the direction of the departure is to use one-sided tests. However, it is the exibility of having two-tail errors that allows the L method intervals to have coverage probabilities closer to the nominal level. If the L method was restricted to probabilities in a single tail, I suspect its length would differ very little from the T method intervals.

Paul VOS

Biostatistics Department
East Carolina University
Greenville, NC 27858

REFERENCES

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

REPLY

We thank Professor Vos for his very careful reading of our article. Vos states that the "L method confidence region is not an interval" and objects to describing L method confidence intervals for proportions in finite populations as likelihood-based. To prove his objection he uses the case of $N = 100, n = 30$, and $y_{\text{obs}} = 0$ as shown in Table 2 of Section 3.3 of our article. Table 2 was intended to point out a peculiarity in the derivation of the L method. It shows a gap at $M = 7$ in the confidence set $C_{.1} = \{0, 1, 2, 3, 4, 5, 6, 8\}$. In this example $C_{.1}$ is not the L method confidence interval. Section 3.1 defines the L method confidence interval for the number M of items with a specified characteristic in a population of size N as $M_{L,L}(y_{\text{obs}}) \leq M \leq M_{U,L}(y_{\text{obs}})$. The explanatory text of the example clearly states that, consistent with this definition of the L method, "The confidence interval is $[0, 8]$. This is the same interval as the one obtained by the T method (p. 58)." The L method confidence interval includes the value of $M = 7$ missing in $C_{.1}$. L method intervals are always intervals.

Vos correctly points out that the L method algorithm as stated in our article does not always produce likelihood confidence intervals in the strict sense of the definition. The intention of the L method was to produce confidence intervals that are short and match nominal confidence levels. In order to match nominal confidence coefficients closely and to achieve short interval lengths, on occasion a confidence interval endpoint with a lower likelihood may be included in preference to one with a higher likelihood value outside the opposite limit of the interval. The example for $y_{\text{obs}} = 3, N = 100$, and $n = 30$ shows this. It seems reasonable to change this feature by checking for the likelihood value at the other end of the interval and exchanging a value of lower likelihood with one of higher likelihood. In the cited example this would change the confidence interval to $4 \leq M \leq 20$ from $5 \leq M \leq 21$. It would raise the coverage probability for $M = 4$ and reduce it for $M = 21$. This simple change would maintain both the spirit and the rule of likelihood confidence intervals.

L method is a two-sided procedure. If one were to apply the L method to one-sided confidence intervals, it would reduce to the T method. In the last paragraph of our article we unambiguously recommend the T method for one-sided confidence bounds. For two-sided applications, however, the L method provides shorter confidence intervals than the T method most of the time, and with closer to nominal (conservative) coverage probabilities. We cannot concur with Vos that the use of one-sided tests is always "simple and natural." Experience shows a myriad of inherently two-sided applications requiring two-sided procedures such as confidence intervals. When these involve proportions in finite populations or related parameters, the L method is a good choice when interval length is a consideration.

Josef SCHMEE

Graduate Management Institute
Union College
Schenectady, NY 12308

John WENDELL

College of Business of Business Administration
University of Hawai`i at Manoa
Honolulu, HI 96822