VOS

The existence of the confidence interval described in the title is of theoretical interest. However, there are difficulties with this interval that should be considered before using this interval in practice.

Wall, Boen, and Tweedie (WBT) considered confidence intervals $C_{\overline{x},\alpha}$ of the form

$$C_{\overline{x},\alpha} = \{\mu : \overline{x} - \zeta_\alpha|\overline{x}| \le \mu \le \overline{x} + \zeta_\alpha|\overline{x}|\}, \quad (1)$$

where $\zeta_\alpha > 1$ is chosen so that $C_{\overline{x},\alpha}$ covers the true value with probability at least $1 - \alpha$. The test derived from this interval rejects $H_0 : \mu = \mu_0$ at level α if and only if $C_{\overline{x},\alpha}$ does not contain μ_0, and is defined by the rejection regions

$$\begin{aligned} R_{\mu_0,\alpha} &= \{\overline{x} : \mu_0 \notin C_{\overline{x},\alpha}\} \\ &= \left\{\overline{x} : \frac{-|\mu_0|}{\zeta_\alpha - \text{sign}(\mu_0)} < \overline{x} < \frac{|\mu_0|}{\zeta_\alpha + \text{sign}(\mu_0)}\right\}, \end{aligned}$$

where $\text{sign}(0) = 0$ and $\text{sign}(\mu_0) = \mu_0/|\mu_0|$ for $\mu_0 \neq 0$.

These are very unusual rejection regions. Consider the two-sided test of $H_0 : \mu = 10$ at $\alpha = .10$. Using $\zeta_{.10} = 4.84$ gives rejection region $\{\overline{x} : -2.6 < \overline{x} < 1.7\}$. If $\overline{x} = -2$, $H_0 : \mu = 10$ is rejected, but if $\overline{x} = -3$, H_0 is accepted. The confidence intervals (1) order the sample space so that observations near zero are evidence against H_0 for any $\mu_0 \neq 0$. In fact, if $\mu_0 \neq 0$ and $\overline{x} = 0$, then H_0 is rejected at any level α. On the other hand, $H_0 : \mu = 0$ is accepted at any level α no matter what the value $\overline{x}$ is observed. Thus, this inferential procedure tells us nothing about the hypothesis that the mean is zero.

In contrast, the t interval (available for sample sizes $n \ge 2$) provides a sensible ordering of the sample space for any μ_0 : values of $\overline{x}$ that are far from μ_0 (on the estimated standard error scale) are evidence against H_0. I disagree with the authors' suggestion that (1) be used in practice because it may have shorter average length than the t interval. I agree with WBT that "there is a good use for this example in the classroom." However, I would use this example to show that in defining an effective confidence interval, one needs to consider more than interval length.

Paul VOS
Biostatistics Department
East Carolina University
Greenville, NC 27858

WITTKOWSKI

Wall, Boen, and Tweedie suggested $x \pm c|x|$ as a confidence interval (CI) for μ with $c = 4.84$ for the 90% level. This gives rise to two "counter examples":

First: If $x = 0$ has been observed, this yields a CI of width 0, regardless of the level.

Second: The 90% CI for 10 °C is (−38 °C ... 48 °C) or (−37 °F ... 119 °F), while the 90% CI for 50 °F (= 10 °C) is (−192 °F ... 292 °F).

Clearly, some explanation is warranted.

Knut M. WITTKOWSKI
The Rockefeller University Hospital
1230 York Ave 121B, Box 322
New York, NY 10021

REPLY

I am very glad to see that this article sparked much interest (as it did for me). Just before his untimely death, Richard Tweedie joked with me that the 12 e-mail responses and inquiries we received within a month of this little four-page article appearing were more than he had ever seen for any of his papers in his distinguished career. This article started as a lunchroom conversation (or sparring) between Jim Boen and myself and I am grateful to Richard Tweedie for encouraging and helping us to turn it into an article.

The letter by Hodge and Huang points out that the intuition we provided in the conclusion is not satisfactory (perhaps even incorrect) for explaining why this implausible approach works. We struggled with giving an intuitive explanation for the results. Unlike most problems I have worked on where intuition comes first followed by difficult mathematical theorems and proof, in this case, the theorem/proof came relatively easy yet the intuition is still somewhat foggy. Hodge and Huang add to the discussion by arguing that our confidence interval demonstrates that a confidence interval does not necessarily give information about the variability. Although true, it is not clear that this discussion lends to the intuition of why this confidence interval works in the first place. Perhaps intuition can better be gained by emphasizing that the confidence interval we present is *not equivariant* (i.e., its coverage probability depends on the ratio of $|\mu|/\sigma$ as seen in Equation 2), and all we are guaranteeing with the interval is that the coverage probability will be *at least* $100(1-\alpha)\%$ for all $(|\mu|, \sigma)$ pairs.

Vos and Wittkowski both point out similar "unusual" behavior that this confidence interval exhibits given an observed x value of 0. Although this may seem strange, it is not incorrect. Recall that the coverage probability of a confidence interval is based on the random variable X, not the given data x. Once we are given a dataset, in this case one data point, the confidence interval either covers the true value of the parameter or it does not cover the parameter. That is, given the data, the coverage probability is 1 or 0. It is the method itself (based on the random sampling distribution of the statistic) that yields coverage probability $(1 - \alpha)$.

Wittkowski's second point is meant to expose the bizarre behavior of the confidence interval when the scale is changed. He shows that the interval obtained given x is not the same as the interval obtained given $ax + b$. Although this may seem strange, (again) it is not incorrect. As we mentioned in the article, $X \pm \zeta|X|$ is not in the equivariant class of confidence intervals (it is not based on a pivotal quantity) and as such it will exhibit the behavior Wittkowski has shown.

When we chose to use the word "effective" in our title to describe our confidence interval, we were thinking in terms of efficiency (i.e., expected margin of error). That is, the interval for a sample of size one has expected margin of error less than infinity and with a sample of size two it has margin of error less than Student's t (for some $(|\mu|, \sigma)$ pairs). For practical purposes though, it is easily argued that this improvement in efficiency is not worthy since it means we have to give up equivariance (a property we are used to having our practical statistics exhibit).

Melanie M. WALL
Division of Biostatistics
School of Public Health
University of Minnesota

CHEN, Z., AND KUO, L. (2001), "A NOTE ON THE ESTIMATION OF THE MULTINOMIAL LOGIT MODEL WITH RANDOM EFFECTS," *THE AMERICAN STATISTICIAN*, 55, 89–95: COMMENT BY ZASLAVSKY AND VAN DYK

As Chen and Kuo noted, a multinomial logit model can be fit as a set of Poisson (log-linear) regression models, where the responses are the number of cases in each category. These models are parameterized with regression coefficients β and an intercept parameter θ_k for each covariate class (defined by common values of the covariates). When this model is fitted via maximum likelihood, the maximizing value of each θ_k is the log-normalizing constant of the multinomial distribution. Hence, the profile of the Poisson likelihood (maximized over θ_k) is the multinomial likelihood; maximizing this multinomial profile likelihood is equivalent to maximizing the Poisson likelihood. This is simpler than explicitly including the normalizing constant in the likelihood, and permits use of ordinary Poisson regression software after suitable manipulation of the data, a neat application of a technique sometimes called "Poissonization" in the categorical data literature.

Care must be taken, however, when extending these methods to a multinomial model with random effects. Suppose that we add multivariate normal random effects (together, a vector $\mathbf{u}_k$) to the linear predictors for the various outcomes. Chen and Kuo proposed to include a θ_k parameter for each cluster, and to maximize the marginal likelihood of the Poisson random-effects model over $\beta, \boldsymbol{\theta} = \{\theta_k\}$, and the random effects variances Σ. Thus, they calculated $\arg\max_{\beta\Sigma} \max_{\boldsymbol{\theta}} \int \mathrm{P}(\mathrm{data}|\beta, \boldsymbol{\theta}, \mathbf{u})\ \mathrm{P}(\mathbf{u}|\Sigma) d\mathbf{u}$, where the integral marginalizes the likelihood over the distribution of the entire set of random effects $\mathbf{u} = \{\mathbf{u}_k\}$. The integrand, however, is *not* the multinomial likelihood, but rather a Poisson likelihood *conditional* on $\boldsymbol{\theta}$. The maximum likelihood estimate for the desired marginalized multinomial likelihood, rather, is $\arg\max_{\beta\Sigma} \int [\max_{\boldsymbol{\theta}} \mathrm{P}(\mathrm{data}|\beta, \boldsymbol{\theta}, \mathbf{u})] \mathrm{P}(\mathbf{u}|\Sigma) d\mathbf{u}$, with the order of integration and maximization reversed. In general these operations do not commute and the expressions are not equal.

If the random effects variances are sufficiently small, integrating and then maximizing may not be a bad approximation to maximizing and then integrating. In effect, maximization over θ_k restricts the linear predictor (including the random effects terms) for the Poisson mean for the corresponding class of observations to a (curved) manifold in which the sum of the predicted counts equals the observed count for each k. With a suitably chosen random effects covariance matrix, the predictor can be restricted to the approximating (tangent) plane for a particular class, but in general not for all classes at once.

This example illustrates the computational complexities introduced by the implicit integration in random effects models, and the pitfalls introduced when maximization and integration appear in the same calculation. An alternative approach would be to omit $\boldsymbol{\theta}$ from the model altogether; this would induce a different random-effects model that might be more or less plausible in any given application. We note that Chen and Kuo's method using nonlinear modeling is not affected by the problem we describe, since the desired likelihood is expressed explicitly.

Alan M. ZASLAVSKY
Department of Health Care Policy
Harvard Medical School
Boston, MA 02115

David A. VAN DYK
Department of Statistics
Harvard University
Cambridge, MA 02138

SCHENKER, N., AND GENTLEMAN, J. F. (2001), "ON JUDGING THE SIGNIFICANCE OF DIFFERENCES BY EXAMINING THE OVERLAP BETWEEN CONFIDENCE INTERVALS," *THE AMERICAN STATISTICIAN*, 55, 182–186: COMMENT BY BARTKO

An earlier statistics article (Browne 1979) discussed the visual assessment of the significance of a mean difference when graphically displayed as the mean plus or minus the standard deviation, the standard error or by the 95% confidence interval on the mean. Browne presented some rules of thumb for decision making and discussed unequal variances, the ratio of the interval lengths, and the sample size of the longer interval and the role each plays in devising rules for visual interpretation.

REFERENCES

Browne, R. H. (1979), "On Visual Assessment of the Significance of a Mean Difference," *Biometrics*, 35, 657–667.

John J. BARTKO
Bethesda, MD 20817

HANLEY, J. A., JOSEPH, L., PLATT, R. W., CHUNG, M. K., AND BÉLISLE, P. (2001), "VISUALIZING THE MEDIAN AS THE MINIMUM-DEVIATION LOCATION," *THE AMERICAN STATISTICIAN*, 55, 150–152: COMMENT BY LEVINE

The enjoyable description of the Java applet in Hanley et al. (2001) stimulated me to write these remarks. I have no doubt that students, roused by this infusion of Java, are likely to connect to their proof of the median as the minimum-deviation location. For years I have presented to students, including those at the most elementary level, a simple geometric proof written in informal language. Students are encouraged to draw the proof and try different placements of a point that would minimize the sum of the distances. Their drawings usually lead to that "aha!" experience. Though undoubtedly a Java-enhanced proof attracts more student attention, there are times when the old technologies hold up. So whether you are technologically deprived or your Net is down and the Java won't brew, this proof will see you through!

A Geometric Proof that the Sum of the Distances is Minimized by the Median

The following shows that, for an ordered set of n numbers $y_1 \ldots y_n$, a point that minimizes the sum of the distances from all the n numbers to that point is the median.

Let $y_1 \ldots y_n$ be an ordered set of n numbers and p be a point in $[y_1, y_n]$. Now assume, for the moment, $n = 2$. From Figure 1 we can see that the distance between y_1 and p (i.e., $|y_1 - p|$) *plus* the distance between y_2 and p (i.e., $|y_2 - p|$) is the whole length of the interval or line, which equals the distance between y_1 and y_2. So, the sum of the distances of y_1 and y_2 from p equals the distance between y_2 and y_1. Thus, $S = |y_1 - p| + |y_2 - p| = L = |y_2 - y_1|$.

Figure 1. The sum of the distances of y_1 and y_2 from p when (1) p is within $[y_1, y_2]$ and (2) when $p > y_2$.

If $p > y_2$, then we have a situation also depicted in Figure 1, where p is shown in a lighter shade. In this case the distance from y_2 to p is the short distance represented by the dashed line. The distance from y_1 to p is $|y_2 - y_1|$ plus the distance from y_2 to p. Specifically, the sum of the distances, S, is: $S = |y_2 - p| + |y_1 - p| = L + 2|y_2 - p|$. Thus, S contains an extra term and so is larger than L by $2|y_2 - p|$.

In sum, when p is outside of the interval $[y_1, y_2]$, S will be larger than L. To minimize S, we must have p be in the interval $[y_1, y_2]$ so that $S = L$. Since the median is in the interval it will minimize S, as will any other number in the interval.

The same logic applies even when there are more than two points. To avoid extra terms in S, p must be within the interval $[y_1, y_n]$. When p is within an interval, the sum of the distances to p from the ends of the interval is equal to the distance between the ends of the interval, as shown in Figure 1 (i.e., $|y_2 - p| + |y_1 - p| = |y_1 - y_2|$). Moving inward from the outermost interval $[y_1, y_n]$, examine whether p is within each interval (Figure 2 shows an example where p is within all intervals). If p is within an interval $[y_i, y_{n+1-i}]$, the logic above applies and the sum of the distances of y_i and y_{n+1-i} to p will be the distance between the ends of the interval (i.e., $|y_i - y_{n+1-i}|$). If p is not within an interval, the knowledge gained from Figure 1 is still pertinent; the sum of the distances is not minimized because extra term(s) are needed. The examination of the intervals continues until p is either not in an interval (in which case S has not been minimized) or p has been found to be within all intervals. If n is even, and p falls within each of the $n/2$ intervals then, using the logic from above, we know that $\sum |y_i - p|$ has been minimized.

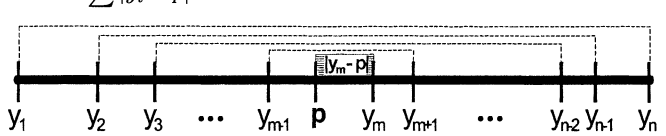

Figure 2. The distance between the median (y_m) and p showing that the sum of the distances is minimized when $p = y_m$.