Score and Pseudo-Score Confidence Intervals for Categorical Data Analysis

Alan AGRESTI

This article reviews methods for constructing confidence intervals for analyzing categorical data. A considerable literature indicates that the method of inverting score tests performs well for a variety of cases. When the sample size is small or the parameter is near the parameter space boundary, this method usually performs much better than inverting Wald tests and sometimes better than inverting likelihood-ratio tests. For small samples, *exact* methods are also available. Although these methods can be quite conservative, inverting a score test using the *mid P-value* provides a sensible compromise that uses the small-sample distribution while reducing the conservatism and only slightly sacrificing the lower bound for the desired confidence level. For some models, score-test-based inferences are impractical, such as when the likelihood function is not an explicit function of the model parameters. For such cases, pseudo-score inference can be based on a Pearson-type chi-squared statistic that compares fitted values for a working model with fitted values when the parameter of interest takes a fixed value. For some simple cases involving proportions and their differences, a different pseudo-score approach that adds artificial observations before forming Wald confidence intervals provides a simple way of approximating score confidence intervals.

Key Words: Discrete data; Multinomial models; Pearson chi-squared; Odds ratio; Proportions; Score test.

1. Introduction

Confidence intervals for a parameter can be constructed by inverting significance tests about the value of that parameter. The usual approach is based on inverting one of three large-sample chi-squared tests—the likelihood-ratio test proposed by Sam Wilks (1938), the Wald test proposed by Abraham Wald (1943), or the score test proposed by C. R. Rao (1948).

For inversion of the Wald test, the 95% confidence interval has the generic form, estimate $\pm$ 1.96 estimated standard errors. Historically, Wald test-based inference was the standard approach in the early development of methods for categorical data analysis, because of its computational simplicity. A good example is the article by Grizzle, Starmer, and Koch (1969) applying weighted least squares methods to analyze categorical data, and many follow-up articles by Gary Koch and his colleagues and students that considered various types of applications (such as repeated measures analyses, in the influential article by Koch et al. 1977). The Wald method is still the most commonly used method, because of its ease of use with statistical software that outputs parameter estimates and standard errors. Increasingly used also in the present era is the inversion of the likelihood-ratio test, which compares the maximized log-likelihood function under null and alternative hypotheses for the possible null parameter values to generate the profile likelihood confidence interval. Less commonly used is the inversion of the score test, which is based on the derivative of the log-likelihood function at the null hypothesis.

Statistics in Biopharmaceutical Research
2011, Vol. 3, No. 2
DOI: 10.1198/sbr.2010.09053

For a generic parameter β, consider a confidence interval (CI) based on inverting a two-sided significance test of H_0: $\beta = \beta_0$ for the set of possible values β_0. The $100(1-\alpha)\%$ CI is the set of β_0 values for which the test has P-value $> \alpha$. For a log-likelihood function $L(\beta)$ (using a single parameter here for notational simplicity), denote the maximum likelihood (ML) estimate by $\hat{\beta}$, the score function by $u(\beta) = \partial L(\beta)/\partial\beta$, and the information by $\iota(\beta) = -E[\partial^2 L(\beta)/\partial\beta^2]$. The Wald test uses

$$[(\hat{\beta} - \beta_0)/SE]^2 = (\hat{\beta} - \beta_0)^2 \iota(\hat{\beta}),$$

where $\iota(\hat{\beta})$ denotes $\iota(\beta)$ evaluated at $\hat{\beta}$. The 95% Wald CI for β is $\hat{\beta} \pm 1.96(\text{SE})$. The likelihood-ratio (LR) test statistic is

$$-2[L(\beta_0) - L(\hat{\beta})].$$

The score test statistic is

$$\frac{[u(\beta_0)]^2}{\iota(\beta_0)} = \frac{[\partial L(\beta)/\partial\beta_0]^2}{-E[\partial^2 L(\beta)/\partial\beta_0^2]},$$

where the partial derivatives are evaluated at β_0. The three chi-squared tests are asymptotically equivalent under H_0 (Cox and Hinkley 1974). The Wald method, unlike the other two methods, has the disadvantage that it depends on the scale of measurement. For example, a Wald CI for e^{β} does not consist of the exponentiated values of the Wald CI for β. Thus, in using the Wald method, a sensible choice of scale is imperative, such as using the log scale for the odds ratio and relative risk.

Section 2 summarizes research that has found that for parameters in some basic categorical data analyses, inverting the score test is an effective, well-performing method. Even for small samples, the large-sample score-test-based interval often performs surprisingly well. By good performance, we mean that actual error probabilities (i.e., the size of the test and the confidence level of the CI) are usually close to their nominal levels. [More detailed evaluations can compare other criteria as well, such as length and types of bias; e.g., see Vos and Hudson (2005).] Section 3 discusses how a slightly adjusted method based on inverting small-sample score tests using the mid P-value also performs well. Score confidence intervals are sometimes difficult to construct, however, such as when the likelihood function is not an explicit function of the model parameters. For interval estimation of a parameter in a multinomial model, Agresti and Ryu (2010) proposed a "pseudo-score" method that inverts a test using a modified Pearson statistic comparing the fitted values for the model to the fitted values assuming a particular value of that parameter. Section 4 introduces this method and Section 5 outlines generalizations and potential related research.

Although the Wald-test-based confidence interval is simple and is commonly taught in introductory statistics courses and used in practice, it often performs poorly when the sample size is small or the parameter is near the boundary of the parameter space. A type of application in which this poor behavior is often relevant is meta-analysis using information from many studies or centers in comparing two treatments on a binary response, when some studies have no or very few outcomes of a particular type but results are combined using weights based on estimated variances. Section 6 summarizes simple adjustments of Wald intervals for proportions and their differences such that the intervals resemble and perform similarly to score intervals.

2. Score-Test-Based Confidence Intervals

For generalized linear models with the canonical link function, such as binomial logistic regression models and Poisson loglinear models, the score test statistic for a parameter can be expressed as a standardization of its sufficient statistic. For subject i, letting y_i denote the response outcome and x_{ij} denote the value of the explanatory variable j for which β_j is the coefficient, the sufficient statistic for β_j has the simple form $\sum_i x_{ij} y_i$. Many popular tests in categorical data analysis can be derived as score tests that a parameter or a set of parameters equal 0. Examples are the Pearson chi-squared test of independence in two-way contingency tables, the McNemar test for comparing proportions with binary matched pairs, the Cochran–Mantel–Haenszel test of conditional independence for stratified 2×2 tables, and the Cochran–Armitage trend test for several ordered binomial samples. Methods that construct their estimates of variability under a null hypothesis condition are often score tests or are closely related to score tests.

Although score tests are well established for practical applications, score CIs are much less well known and used. The only score CI that seems to receive considerable use is Wilson's (1927) CI for a binomial parameter π. It can be expressed as an inversion of the asymptotic standard normal test using test statistic

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}},$$

for which the endpoints are roots of a quadratic equation. Score CIs are less well known for other cases, even for basic parameters for 2×2 contingency tables $\{n_{ij}\}$.

Three particularly important parameters for 2×2 tables are the difference of proportions, the odds ratio, and the relative risk. For the difference of proportions $\pi_1 - \pi_2$ for two independent binomial samples, Mee (1984) and Miettinen and Nurminen (1985) showed that the score CI inverts the test of H_0: $\pi_1 - \pi_2 = \beta_0$ using test statistic that

is the square of

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \beta_0}{\sqrt{[\hat{\pi}_1(\beta_0)(1-\hat{\pi}_1(\beta_0))/n_1] + [\hat{\pi}_2(\beta_0)(1-\hat{\pi}_2(\beta_0))/n_2]}},$$

where $\hat{\pi}_1$ and $\hat{\pi}_2$ are the sample proportions (i.e., unrestricted ML estimates) and $\hat{\pi}_1(\beta_0)$ and $\hat{\pi}_2(\beta_0)$ are the ML estimates subject to the constraint $\pi_1 - \pi_2 = \beta_0$. (When $\beta_0 = 0$, z^2 is the Pearson chi-squared statistic for testing independence, applied to a 2×2 table.) For a fixed β_0 the restricted ML estimates have closed form, but the set of such β_0 having sufficiently small $|z|$ to fall in the CI must be determined iteratively. For interval estimation of an odds ratio, for a given β_0 let $\{\hat{\mu}_{ij}(\beta_0)\}$ be the unique values that have the same row and column margins as $\{n_{ij}\}$ and satisfy

$$\frac{\hat{\mu}_{11}(\beta_0)\hat{\mu}_{22}(\beta_0)}{\hat{\mu}_{12}(\beta_0)\hat{\mu}_{21}(\beta_0)} = \beta_0.$$

Let $\chi^2_{1,a}$ denote the $100(1-a)$ percentile of a chi-squared distribution with $df = 1$. The set of β_0 satisfying

$$X^2 = \sum (n_{ij} - \hat{\mu}_{ij}(\beta_0))^2/\hat{\mu}_{ij}(\beta_0) \le \chi^2_{1,a}$$

form a $100(1-a)\%$ conditional score CI for the odds ratio (Cornfield 1956). Likewise, score CIs exist for the relative risk (Koopman 1984), the difference of proportions with matched samples (Tango 1998), logistic regression parameters, and generic measures of association (Lang 2008).

More generally, let $\{n_i\}$ denote multinomial cell counts for a contingency table of arbitrary dimensions. Let $\{\hat{\mu}_i\}$ be the ML fitted values for a particular model. For testing goodness of fit, the score test statistic is the Pearson statistic,

$$X^2 = \sum \frac{(n_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

This fact was observed for multinomial models by Cox and Hinkley (1974, p. 326) and then extended to a corresponding statistic for generalized linear models by Smyth (2003). When the model corresponds to taking the saturated model and imposing a particular constraint for a parameter β, then inverting the Pearson chi-squared test of H_0: $\beta = \beta_0$ yields the score CI.

For parameters in categorical data analysis, the literature so far indicates that large-sample two-sided score tests and corresponding confidence intervals perform well, often much better than Wald inference. Even with small samples for which asymptotics would be expected to fail, this method performs surprisingly well for simple contingency table parameters and often outperforms likelihood-ratio-test-based inference. This may reflect the fact that for canonical models, the score statistic (a) is a standardization of a sufficient statistic that is a linear combination of the observations and (b) uses a null rather than estimated non-null standard error. For details and evidence such as simulation studies, see Koehler and Larntz (1980) for testing independence in two-way contingency tables; Newcombe (1998a) and Agresti and Coull (1998) for CIs for binomial proportions; Miettinen and Nurminen (1985), Newcombe (1998b), and Agresti and Min (2005a) for CIs for the difference of proportions and relative risk; Tango (1998) and Agresti and Min (2005b) for inference about the difference of proportions for dependent samples; Miettinen and Nurminen (1985) and Agresti and Min (2005a) for CIs for the odds ratio; Agresti and Klingenberg (2005) for multivariate comparisons of proportions; Agresti et al. (2008) for simultaneous CIs comparing several binomial proportions; and Ryu and Agresti (2008) for effect measures comparing two groups on an ordinal scale.

In practice, Wald CIs and likelihood-ratio-test-based profile likelihood CIs are easily accessible with statistical software. For example, profile likelihood CIs are available with PROC GENMOD in SAS (with the LRCI option) and in R by applying the `confint()` function to an object corresponding to a generalized linear model fit. By contrast, score CIs are not as well known as they deserve to be, given how well they perform, and they have relatively little availability in the primary statistical software packages, even for simple settings such as 2×2 tables.

3. Small-Sample Score Confidence Intervals

Using the score (or other) test statistic, it is possible to apply small-sample distributions, rather than their large-sample normal and chi-squared approximations, to obtain P-values and CIs. To illustrate, consider inference for a parameter of a logistic regression model. For subject i with binary outcome y_i and values for k explanatory variables $x_{i0} = 1, x_{i1}, x_{i2}, \ldots, x_{ik}$, the model is

$$\text{logit}[P(y_i = 1)] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

The score statistic for β_j is based on $T_j = \sum_i y_i x_{ij}$. Starting with the binomial likelihood, one can base a test on the conditional distribution of T_j after eliminating the other parameters by conditioning on their sufficient statistics. For example, with the equal-tail method, bounds (β_{1L}, β_{1U}) of a $100(1-a)\%$ CI for β_1 satisfy

$$P(T_1 \ge t_{1,\text{obs}} | t_0, t_2, \ldots, t_k; \beta_{1L}) = a/2,$$

$$P(T_1 \le t_{1,\text{obs}} | t_0, t_2, \ldots, t_k; \beta_{1U}) = a/2.$$

See Mehta and Patel (1995) for details. Software is available for doing this, such as LogXact (Cytel Software 2005).

Because of discreteness, it is not possible for a significance test to have a fixed size such as $a = 0.05$ at all possible null values for a parameter. In rejecting the null hypothesis whenever the P-value $\leq a$, the actual size has a as an upper bound. Hence, actual confidence levels for small-sample interval estimation inverting such tests do not exactly equal the nominal values, and the inferences are *conservative*. Inverting a test with actual size $\leq a$ for all β_0 guarantees that the CI has actual coverage probability $\geq (1-a)$. In practice, the actual coverage probability varies for different β values and is unknown.

When the conservatism is problematic, there are ways of alleviating it (Agresti 2003). One way, which also results in a narrower interval, is to invert a single two-sided test instead of two equal-tail one-sided tests. Another approach that is feasible when the parameter space is small (such as for 2×2 tables) uses an unconditional approach to eliminate nuisance parameters, because the conditional approach exacerbates the discreteness. For H_0: $\beta = \beta_0$ with nuisance parameter ψ, let $p(\beta_0; \psi)$ be the P-value for a given value of ψ. The unconditional P-value is $\sup_\psi p(\beta_0; \psi)$ and the $100(1-a)\%$ CI consists of β_0 for which $\sup_\psi p(\beta_0; \psi) > a$. Agresti and Min (2002) found that this approach works well for the odds ratio, using a two-sided score statistic as the criterion. Agresti and Min (2001) inverted two-sided score tests for the difference and ratio of proportions, and this method is available in the StatXact software (Cytel 2005). Coe and Tamhane (1993) proposed an alternative unconditional approach for the difference and ratio of proportions that is more complex but performs well. Santner et al. (2007) reviewed several such methods.

For single-parameter problems and for the conditional approach that eliminates nuisance parameters, the discreteness and consequent conservatism can be eliminated completely by using a randomized type of inference. With a discrete test statistic T such as a score statistic, let $\mathcal{U}$ be a uniform(0,1) random variable. For testing $H_0 : \beta = \beta_0$ against $H_a : \beta > \beta_0$ using T, the randomized test has P-value

$$P_{\beta_0}(T > t_{\text{obs}}) + \mathcal{U} \times P_{\beta_0}(T = t_{\text{obs}}).$$

This has a uniform(0,1) null distribution, which is always the case for ordinary P-values with test statistics having a continuous distribution. A CI with actual coverage probability exactly equal to $(1-a)$ has endpoints (β_L, β_U) satisfying

$$P_{\beta_U}(T < t_{\text{obs}}) + \mathcal{U} \times P_{\beta_U}(T = t_{\text{obs}}) = a/2$$

$$P_{\beta_L}(T > t_{\text{obs}}) + (1-\mathcal{U}) \times P_{\beta_L}(T = t_{\text{obs}}) = a/2.$$

Stevens (1950) suggested this approach for the binomial parameter. Although in this modern era it is viewed as unacceptable to use a method for which the results depend on a random number, it is a historical curiosity that Stevens and other statisticians apparently believed that this approach would be adopted for applied work. For example, Pearson (1950) argued that statisticians may well come to accept randomization *after* performing an experiment just as they had come to accept Fisher's ideas about randomization *before* performing the experiment. Stevens (1950) argued that an advantage of eliminating the uncertainty about the actual coverage probability is that a narrower CI results than with standard small-sample methods.

These days, randomized inference of this type is not used. However, some authors advocate a *fuzzy inference* approach that portrays graphically all such possible randomized CIs (Geyer and Meeden 2005). We find that approach useful for motivating an alternative method based on inverting tests using the *mid P-value* (Lancaster 1961). For $H_a : \beta > \beta_0$, the mid P-value is

$$P_{\beta_0}(T > t_{\text{obs}}) + (1/2)P_{\beta_0}(T = t_{\text{obs}}).$$

Unlike the randomized P-value, it depends only on the data. Under H_0, the ordinary P-value is stochastically larger than uniform in distribution, but the mid P-value is not and has $E(\text{mid } P\text{-value}) = 1/2$. The sum of right-tail and left-tail P-values equals $1 + P_{\beta_0}(T = t_{\text{obs}})$ for the ordinary P-value but equals 1 for the mid P-value. Using the small-sample distribution, a $100(1-a)\%$ CI for β based on the mid P-value is determined by

$$P_{\beta_U}(T < t_{\text{obs}}) + (1/2) \times P_{\beta_U}(T = t_{\text{obs}}) = a/2,$$

$$P_{\beta_L}(T > t_{\text{obs}}) + (1/2) \times P_{\beta_L}(T = t_{\text{obs}}) = a/2.$$

Although the coverage probability of this interval is not guaranteed to be $\geq (1-a)$, in practice it is usually close to that value. Numerical evaluations suggest that it tends to be a bit conservative in an average sense.

To illustrate, suppose T is a binomial random variable. Using this construction with binomial probabilities and (1/2) replaced by 1.0 yields the standard Clopper and Pearson (1934) exact (conservative) CI. Considering this method and the mid P-based CI over all the possible parameter values between 0 and 1, Figure 1 from Agresti and Gottard (2007) shows the quartiles of nominal 95% coverage probabilities as a function of n. The median coverage probability (represented in each case by the middle of the three curves) is much closer to the nominal level for the mid-P-based CI. It would be useful if statistical software could provide mid-P-based small-sample CIs for cases with a single parameter, such as the binomial and Poisson parameters, and for cases in which nuisance parameters are eliminated, such as odds ratios and logistic regression parameters.

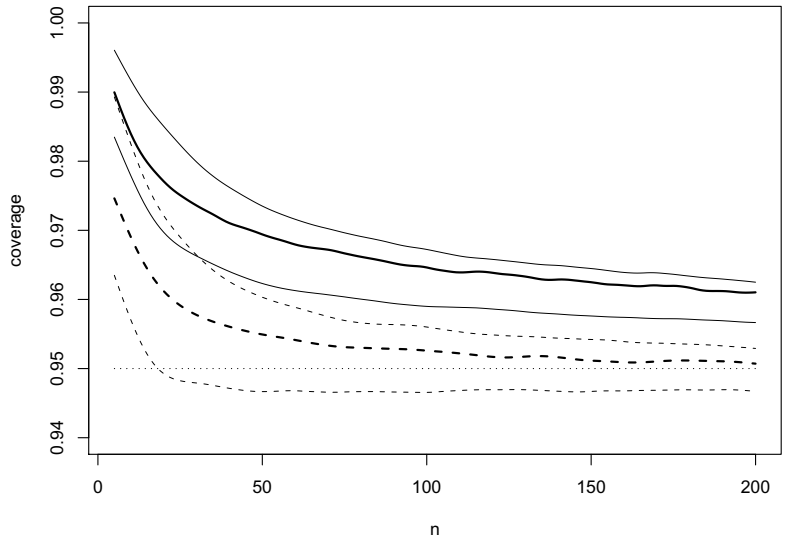

Figure 1. Quartiles of coverage probabilities for Clopper–Pearson (—) and mid-P (- - -) small-sample confidence intervals for binomial parameter, from Agresti and Gottard (2007). For example, at a particular n, Clopper-Pearson coverage probabilities fall above top curve for 25% of π values in (0, 1).

4. Pseudo-Score Inference Using the Pearson Chi-Squared Statistic

Although the method of inverting score tests to obtain CIs works well for parameters in simple models, it is often difficult or even infeasible to implement. A prime example is the set of models for which the likelihood function is not an explicit function of the model parameters.

To illustrate, consider Table 1 showing data from Kenward and Jones (1991) on results from a crossover study designed to compare two dosages of a treatment for relief of severe uterine pain during a woman's menstrual cycle. (The study also used a placebo treatment, not shown here.) To detect whether responses tend to be more positive for one dose than the other, we could compare the marginal distributions using the cumulative logit marginal model for the responses (y_1, y_2),

$$\begin{aligned} \text{logit}[P(y_1 \le j)] &= \alpha_j, \\ \text{logit}[P(y_2 \le j)] &= \alpha_j + \beta, \quad j = 1, 2. \end{aligned}$$

The multinomial log-likelihood function, in terms of cell probabilities $\{\pi_{ij}\}$ and cell counts $\{n_{ij}\}$, is

$$L(\boldsymbol{\pi}) \propto \log[\pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \ldots \pi_{33}^{n_{33}}],$$

but the model parameters refer to marginal probabilities so cannot be substituted into the likelihood function. Other models for this table for which the score method would be difficult to implement are the random effects analog of this marginal model, the association model that specifies a common value for the four global odds ratios, and a model by which the mean for one response changes linearly across categories of the other response, for a particular choice of category scores.

For a multinomial model for cell counts $\{n_i\}$ with ML fitted values $\{\hat{\mu}_i\}$, let $\{\hat{\mu}_{i0}\}$ denote fitted values for a simpler "null" model (e.g., with a certain parameter $\beta = \beta_0$). For testing the simpler model against the full model, the LR statistic is

$$G^2 = 2 \sum_i \hat{\mu}_i \log(\hat{\mu}_i / \hat{\mu}_{i0}).$$

The profile likelihood $100(1-a)\%$ CI for β is

$$\{\beta_0\} \text{ such that } G^2 \le \chi^2_{1,a}.$$

Agresti and Ryu (2010) proposed a method that parallels this one, but using the Pearson statistic, with the purpose of making score-type CIs available when the score CI itself is not easily obtainable. Rao (1961) suggested the Pearson-type statistic for comparing models,

$$X^2 = \sum_i \frac{(\hat{\mu}_i - \hat{\mu}_{i0})^2}{\hat{\mu}_{i0}}.$$

This is a quadratic approximation for G^2. From a Taylor series expansion, X^2 has the same limiting null distribution as G^2 even under sparse asymptotics in which the number of cells in the contingency table grows with the

Table 1. Contingency table used to illustrate pseudo-score inference

	y_2 = High Dose			
y_1 = Low Dose	No relief	Moderate relief	Complete relief	Total
No relief	9	7	9	25
Moderate relief	4	16	11	31
Complete relief	4	10	16	30
Total	17	33	36	86

sample size, as is the case when at least one explanatory variable is continuous (Haberman 1977). For general categorical data modeling, the alternative $100(1-a)\%$ CI for β is

$$\{\beta_0\} \text{ such that } X^2 \leq \chi^2_{1,a}.$$

When the full model is saturated, this method yields the score CI. When the model is unsaturated, X^2 is not the score test statistic. The test using X^2 to compare models in that case is a *pseudo-score test* and the CI is a *pseudo-score confidence interval*.

As an aside, we mention that in the case of the canonical link function for a generalized linear model, Lovison (2005) gave a formula for the score statistic that resembles the Pearson statistic, being a quadratic form comparing fitted values for the two models. Let $\mathbf{X}$ be the model matrix for the full model and let $\hat{\mathbf{V}}_0$ be the diagonal matrix of estimated variances under the null model, with fitted values $\hat{\boldsymbol{\mu}}$ for the full model and $\hat{\boldsymbol{\mu}}_0$ for the reduced model. Then, the score statistic is

$$(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0)'\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}_0\mathbf{X})^{-1}\mathbf{X}'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0).$$

Lang, McDonald, and Smith (1999) gave this formula for the log-linear case. Moreover, for the canonical-link case, Lovison showed that the score statistic bounds below the Pearson statistic comparing the models and is a first-order approximation for it. For such cases, which include binomial logistic regression and Poisson loglinear models, it follows that asymptotic P-values for the ordinary score test are at least as large as those for the pseudo-score test, and CIs based on inverting the score test contain CIs based on inverting the pseudo-score test. However, we stated this is an "aside" because the pseudo-score method is intended for more complex, noncanonical link cases, in which the score statistic is often not available.

For Table 1, the ML estimate of β for the marginal cumulative logit model is $\hat{\beta} = -0.389$, with SE $= 0.251$. The model fits well, with a Pearson goodness-of-fit statistic value of 0.45 ($df = 1$). One can fit the marginal model for various fixed β_0 (taking that value times the margin indicator as an offset) by using the R function `mph.fit` available from Joseph Lang at the University of Iowa. The 95% pseudo-score CI for β is $(-0.891, 0.110)$. Here, results are similar to those for the profile likelihood CI of $(-0.891, 0.104)$.

Pseudo-score methods are useful for three reasons: First, for models that are not generalized linear models with canonical link, ordinary score methods often are not practical but the pseudo-score methods can be implemented with the same level of difficulty as profile likelihood confidence intervals. Second, as the next section discusses, extensions apply to settings in which profile likelihood methods are not available. Third, as Section 2 mentioned, research has shown that ordinary score inferences (when available) perform well in terms of having actual coverage probability near the nominal level for a variety of measures for discrete data. Through simulations, Agresti and Ryu (2010) found that the pseudo score method has similar behavior, performing similarly to the profile likelihood interval and sometimes even a bit better when sample sizes are small.

5. Generalizations of Pseudo-Score Inference

Agresti and Ryu (2010) also proposed generalizations of pseudo-score inference. We briefly mention two such generalizations here, the second of which has potential for future research.

First, the method generalizes to discrete distributions other than the multinomial and to sampling schemes more complex than simple multinomial sampling. Suppose $\{y_i, i = 1, \ldots, n\}$ are independent observations assumed to have a specified discrete distribution. A Pearson-type statistic for comparing models has the form

$$X^2 = \sum_i \frac{(\hat{\mu}_i - \hat{\mu}_{i0})^2}{v(\hat{\mu}_{i0})} = (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0)'\hat{\mathbf{V}}_0^{-1}(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0),$$

where $v(\hat{\mu}_{i0})$ denotes the estimated variance of y_i assuming the null distribution for y_i and $\hat{\mathbf{V}}_0$ is the diagonal matrix containing such values (Lovison 2005). The pseudo-score methods for multinomial responses extend to parameters of models for discrete data using this general-

ized statistic, such as parameters of Poisson regression models.

Many data are obtained with a complex sampling scheme. For example, most surveys do not use simple random sampling but instead a multi-stage sample that employs stratification and clustering. One can then replace $\hat{V}_0$ by an appropriately inflated or nondiagonal estimated covariance matrix. Suppose, for example, that the sampling variances of estimates are approximately 50% larger than obtained with simple random sampling (as is the case according to the codebook of the General Social Survey). We can then obtain more relevant confidence intervals from the set of β_0 with $\sum_i(\hat{\mu}_i - \hat{\mu}_{i0})^2/(1.50\hat{\mu}_{i0}) \leq \chi^2_{1,a}$. For such complex sampling designs, profile likelihood confidence intervals are not available and need to be replaced by quasi-likelihood adaptations.

Second, the pseudo-score inference presented above may extend to other types of quasi-likelihood analyses. A possible application is marginal modeling of clustered categorical responses. A popular approach for marginal modeling uses the method of generalized estimating equations (GEE). Because of the lack of a likelihood function with this method, Wald methods are commonly employed, together with a sandwich estimator of the covariance matrix of model parameter estimates. For binary data, let y_{it} denote observation t in cluster i, for $t = 1,\ldots,T_i$ and $i = 1,\ldots,n$. Let $\mathbf{y}_i = (y_{i1},\ldots,y_{iT_i})'$ and let $\boldsymbol{\mu}_i = E(\mathbf{y}_i) = (\mu_{i1},\ldots,\mu_{iT_i})'$. Let $\mathbf{V}_i$ denote the $T_i \times T_i$ covariance matrix of $\mathbf{y}_i$. For a particular marginal model, let $\hat{\boldsymbol{\mu}}_i$ denote an estimate of $\boldsymbol{\mu}_i$, such as the ML estimate under the naive assumption that the $\sum_i T_i$ observations as independent. Let $\hat{\boldsymbol{\mu}}_{i0}$ denote the corresponding estimate under the constraint that a particular parameter β takes value β_0. Let $\hat{\mathbf{V}}_{i0}$ denote an estimate of the covariance matrix of $\mathbf{y}_i$ under this null model. The main diagonal elements of $\hat{\mathbf{V}}_{i0}$ are $\hat{\mu}_{it0}(1-\hat{\mu}_{it0})$, $t = 1,\ldots,T_i$. Separate estimation is needed for the null covariances, which are not part of the marginal model.

Now, consider

$$X^2 = \sum_i (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_{i0})'\hat{\mathbf{V}}_{i0}^{-1}(\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_{i0}).$$

With categorical explanatory variables, X^2 applies to two sets of fitted marginal proportions for the contingency table obtained by cross-classifying the multivariate binary response with the various combinations of explanatory variable values. The set of β_0 values for which $X^2 \leq \chi^2_{1,a}$ is a CI for β. Unlike the GEE approach, this method does not require using the sandwich estimator, which can be unreliable unless the number of clusters is quite large (Firth 1993; Kauermann and Carroll 2001). Even with consistent estimation of $\mathbf{V}_{i0}$, however, the limiting null distribution of X^2 need not be exactly chi-squared because the fitted values result from inefficient estimates. However, based on preliminary simulations, it seems that the chi-squared distribution often provides a good approximation. Boos (1992) and Rotnitzky and Jewell (1990) presented score-type tests for the clustered-data setting.

6. Pseudo-Score CIs That Adjust Wald CIs

Of the three types of tests inverted to construct CIs, the Wald test tends to have the poorest performance. This is unfortunate, as it is the informal way that most methodologists inspecting software output will form a CI. The Wald method is usually fine for producing "rough-and-ready" results, especially when n is large, but for many purposes it is better to use safer methods such as score and profile likelihod CIs. However, for the simple problem of estimating a binomial parameter or comparing two such parameters, simple rough-and-ready adjustments of Wald CIs approximate score CIs and have similar good performance, even with small samples.

Suppose y has a binomial distribution with parameter π, and let $\hat{\pi} = y/n$. Agresti and Coull (1998) noted that in the 95% case, finding all π_0 such that $|\hat{\pi} - \pi_0|/\sqrt{\pi_0(1-\pi_0)/n} < 2$ provides the score CI of form $M \pm 2s$ with

$$M = \left(\frac{n}{n+4}\right)\hat{\pi} + \left(\frac{4}{n+4}\right)\frac{1}{2} = \frac{y+2}{n+4},$$

and

$$s^2 = \frac{1}{n+4}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+4}\right) + \frac{1}{2}\frac{1}{2}\left(\frac{4}{n+4}\right)\right].$$

They used this to motivate the 95% CI, now referred to in some elementary texts as the *plus four CI*,

$$\tilde{\pi} \pm 1.96\sqrt{\tilde{\pi}(1-\tilde{\pi})/\tilde{n}}$$

with $\tilde{\pi} = (y+2)(n+4)$ and $\tilde{n} = n+4$. This has the same midpoint as the 95% score CI, when we round the normal percentile 1.96 to 2. It is slightly wider by Jensen's inequality, because the variance is found at the weighted average of $\tilde{\pi}$ and $1/2$ instead of using a weighted average of variances.

In fact, this adjustment of the Wald CI has much better performance than the ordinary Wald CI, and when π is close to 0 or 1 it also performs better than the ordinary score CI. Figure 2, from Agresti and Caffo (2000), shows the coverage probabilities of the ordinary and adjusted Wald methods, for various sample sizes and for 95% and 99% CIs. For estimating the difference between two proportions with independent samples, Agresti and Caffo (2000) showed that constructing the Wald CI after adding 4 outcomes, one "success" (S) and one "failure" (F) to each sample, also yields a much better CI. Agresti and Min (2005) showed a similar result for comparing

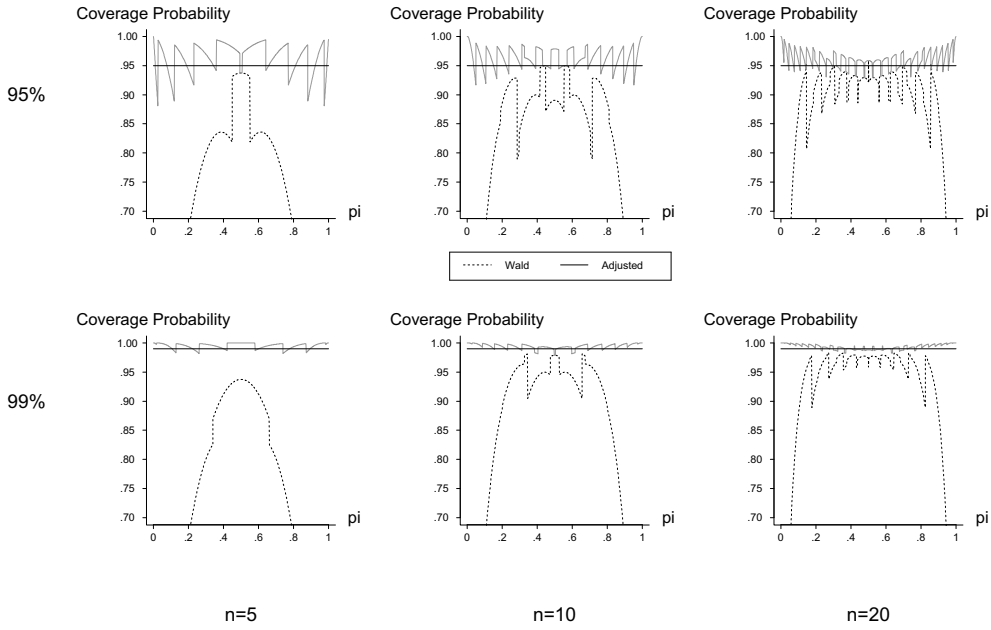

Figure 2. Coverage probabilities for Wald (···) and adjusted Wald pseudo-score (—) confidence intervals for a binomial parameter, from Agresti and Caffo (2000)

proportions with dependent samples, an improved CI resulting from adding one S and one F to each sample such that there is half an additional observation for each possible sequence {(S,S), (S,F), (F,S), (F,F)} for the matched pairs.

Brown, Cai, and Das Gupta (2001) showed further evidence of the poor performance of the Wald method. For example, when a single proportion $\pi = 0.01$ or 0.99, the value of n_0 needed in order for the coverage probability of a nominal 95% Wald CI to exceed 0.94 uniformly in $n \geq n_0$ is about 8000, compared to 1 for the adjusted CI. The poor performance of the Wald CI is due to centering at the point estimate when the parameter space is bounded, not because the CI is too short. In fact, the Wald CI has greater length than an adjusted CI unless the parameters are relatively near the boundary of the parameter space.

The shrinkage form of the adjusted Wald CIs also suggests that CIs resulting from the Bayesian approach can also perform well in a frequentist sense. This was shown with relatively diffuse prior distributions for a single proportion by Brown et al. (2001) and for the difference of proportions, relative risk, and odds ratio by Agresti and Min (2005a). These articles found that the Bayesian probability interval from the $a/2$ to $(1 - a/2)$ quantiles of the posterior distribution perform well in a frequentist sense when the prior distributions for binomial parameters are the Jeffreys prior, which is the U-shaped beta distribution with parameters 0.5 and 0.5.

7. Summary

For basic categorical data analyses, inverting the large-sample score test provides CIs having coverage probabilities near the nominal level. For small-sample distributions with a single parameter, inverting score tests using the mid *P*-value also provides good CIs. Score CIs should be added to the major statistical software packages, now being available mainly in specialized software such as StatXact. Specially written functions for the free software R are available for many such CIs at *www.stat.ufl.edu/~aa/cda/software.html*. The ordinary R function `prop.test` provides it for a single binomial parameter, with the option `CORRECT=FALSE` to delete the Yates continuity correction.

Ordinary score tests and CIs are often infeasible, and Agresti and Ryu (2010) proposed a pseudo-score CI for a multinomial model parameter based on inverting a test using the Pearson statistic. This is a simple unified method that performs well in a wide variety of settings

and can be implemented with ordinary model-fitting software.

Acknowledgments

Thanks to Dr. Euijung Ryu for the pseudo-score results for the example in Section 4.

[Received December 2009. Revised April 2010.]

References

Agresti, A. (2003), "Dealing with Discreteness: Making 'Exact' Confidence Intervals for Proportions, Differences of Proportions, and Odds Ratios More Exact," *Statistical Methods in Medical Research*, 12, 3–21. 166

Agresti, A., Bini, M., Bertaccini, B., and Ryu, E. (2008), "Simultaneous Confidence Intervals for Comparing Binomial Parameters," *Biometrics*, 64, 1270–1275. 165

Agresti, A., and Caffo, B. (2000), "Simple and Effective Confidence Intervals for Proportions and Difference of Proportions Result from Adding Two Successes and Two Failures, *The American Statistician*, 54, 280–288. 169, 170

Agresti, A., and Coull, B. A. (1998), "Approximate is Better Than Exact for Interval Estimation of Binomial Parameters," *The American Statistician*, 52, 119–126. 165, 169

Agresti, A., and Gottard, A. (2007), "Nonconservative Exact Small-Sample Inference for Discrete Data," *Computational Statistics and Data Analysis*, 51, 6447–6458. 166, 167

Agresti, A., and Klingenberg, B. (2005), "Multivariate Tests Comparing Binomial Probabilities, With Application to Safety Studies for Drugs," *Applied Statistics*, 54, 691–706. 165, 169

Agresti, A., and Min, Y. (2001), "On Small-Sample Confidence Intervals for Parameters in Discrete Distributions," *Biometrics*, 57, 963–971. 166

——— (2002), "Unconditional Small-Sample Confidence Intervals for the Odds Ratio," *Biostatistics*, 3, 379–386. 166

——— (2005a), "Frequentist Performance of Bayesian Confidence Intervals for Comparing Proportions in 2 × 2 Contingency Tables," *Biometrics*, 61, 515–523. 165, 170

——— (2005b), "Simple Improved Confidence Intervals for Comparing Matched Proportions," *Statistics in Medicine*, 24, 729–740. 165

Agresti, A., and Ryu, E. (2010), "Pseudo-Score Confidence Intervals for Parameters in Discrete Statistical Models," *Biometrika*, 97, 215–222. 164, 167, 168, 170

Boos, D. D. (1992), "On Generalized Score Tests," *The American Statistician*, 46, 327–333. 169

Brown, L. D., Cai, T. T., and Das Gupta, A. (2001), "Interval Estimation for a Binomial Proportion," *Statistical Science*, 16, 101–133. 170

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. 166

Coe, P. R., and Tamhane, A. C. (1993), "Small Sample Confidence Intervals for the Difference, Ratio, and Odds Ratio of Two Success Probabilities," *Communications in Statistics, Simulation and Computation*, 22, 925–938. 166

Cornfield, J. (1956), "A Statistical Problem Arising from Retrospective Studies," *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, 4, Berkeley, CA: University of California Press, pp. 135–148. 165

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall. 164, 165

Cytel (2005), *StatXact 7 User Manual*, volumes 1 and 2, and *LogXact 7 User Manual*. Cambridge, MA: Cytel Inc. 165, 166

Firth, D. (1993), "Recent Developments in Quasi-Likelihood Methods," *Proceedings of the International Statistical Institute, 49th Session*, 341–358. 169

Geyer, C. J., and Meeden, G. D. (2005), "Fuzzy and Randomized Confidence Intervals and *P*-values," *Statistical Science*, 20, 358–366. 166

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969), "Analysis of Categorical Data by Linear Models," *Biometrics*, 25, 489–504. 163

Haberman, S. (1977), "Log-Linear Models and Frequency Tables with Small Expected Cell Counts," *The Annals of Statistics*, 5, 1148–1169. 168

Kauermann, G., and Carroll, R. J. (2001), "A Note on the Efficiency of Sandwich Covariance Matrix Estimation," *Journal of the American Statistical Association*, 96, 1387–1396. 169

Kenward, M. G., and Jones, B. (1991), "The Analysis of Categorical Data from Cross-over Trials using a Latent Variable Model," *Statistics in Medicine*, 10, 1607–1619. 167

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehnen, R. G. (1977), "A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data," *Biometrics*, 33, 133–158. 163

Koehler, K., and Larntz, K. (1980), "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials," *Journal of the American Statistical Association*, 75, 336–344. 165

Koopman, P. A. R. (1984), "Confidence Intervals for the Ratio of Two Binomial Proportions," *Biometrics*, 40, 513–517. 165

Lancaster, H. O. (1961), "Significance Tests in Discrete Distributions," *Journal of the American Statistical Association*, 56, 223–234. 166

Lang, J. B. (2008), "Score and Profile Likelihood Confidence Intervals for Contingency Table Parameters," *Statistics in Medicine*, 27, 5975–5990. 165

Lang, J. B., McDonald, J. W., and Smith, P. W. F. (1999), "Association-Marginal Modeling of Multivariate Categorical Responses: A Maximum Likelihood Approach," *Journal of the American Statistical Association*, 94, 1161–1171. 168

Lovison, G. (2005), "On Rao Score and Pearson X^2 Statistics in Generalized Linear Models," *Statistical Papers*, 46, 555–574. 168

Mee, R. W. (1984), "Confidence Bounds for the Difference Between two Probabilities" (letter), *Biometrics*, 40, 1175–1176. 164

Mehta, C. R., and Patel, N. R. (1995), "Exact Logistic Regression: Theory and Examples," *Statistics in Medicine*, 14, 2143–2160. 165

Miettinen, O., and Nurminen, M. (1985), "Comparative Analysis of Two Rates," *Statistics in Medicine*, 4, 213–226. 164, 165

Newcombe, R. (1998a), "Two-Sided Confidence Intervals for the Single Proportion," *Statistics in Medicine*, 17, 857–872. 165

——— (1998b), "Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods," *Statistics in Medicine*, 17, 873–890. 165

Pearson, E. S. (1950), "On Questions Raised by the Combination of Tests Based on Discontinuous Distributions," *Biometrika*, 37, 383–398. 166

Rao, C. R. (1948), "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Application to Problems of Estimation," *Proceedings of the Cambridge Philosophical Society*, 44, 50–57. 163

——— (1961), "A Study of Large Sample Test Criteria Through Properties of Efficient Estimates," *Sankhya*, A23, 25–40. 167

Rotnitzky, A., and Jewell, N. P. (1990), "Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data," *Biometrika*, 77, 485–497. 169

Ryu, E., and Agresti, A. (2008), "Modeling and Inference for an Or-

dinal Effect Size Measure," *Statistics in Medicine*, 27, 1703–1717. 165

Santner, T. J., Pradhan, V., Senchaudhuri, P., Mehta, C. R., and Tamhane, A. (2007), "Small-Sample Comparisons of Confidence Intervals for the Difference of Two Independent Binomial Proportions," *Computational Statistics & Data Analysis*, 51, 5791–5799. 166

Smyth, G. K. (2003), "Pearson's Goodness of Fit Statistic as a Score Test Statistic," in *Science and Statistics: A Festschrift for Terry Speed*, ed. D. R. Goldstein, IMS Lecture Notes–Monograph Series, Vol. 40, Hayward, CA: Institute of Mathematical Statistics, pp. 115–126. 165

Stevens, W. L. (1950), "Fiducial Limits of the Parameter of a Discontinuous Distribution," *Biometrika*, 37, 117–129. 166

Tango, T. (1998), "Equivalence Test and Confidence Interval for the Difference in Proportions for the Paired-Sample Design," *Statistics in Medicine*, 17, 891–908. 165

Vos, P. W., and Hudson, S. (2005), "Evaluating Criteria for Discrete Confidence Intervals: Beyond Coverage and Length," *The American Statistician*, 59, 137–142. 164

Wald, A. (1943), "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Transactions of the American Mathematical Society*, 54, 426–482. 163

Wilks, S. S. (1938), "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *Annals of Mathematical Statistics*, 9, 60–62. 163

Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–212. 164

About the Author

Alan Agresti is Distinguished Professor Emeritus, Department of Statistics, University of Florida, Gainesville, Florida 32611 (E-mail: *aa@stat.ufl.edu*).