Journal of Quantitative Economics (2022) 20 (Suppl 1):S31–S48
https://doi.org/10.1007/s40953-022-00309-8

ORIGINAL ARTICLE

A Review of Score-Test-Based Inference for Categorical Data

Alan Agresti[1] · Sabrina Giordano[2] · Anna Gottard[3]

Accepted: 27 April 2022 / Published online: 26 May 2022

Abstract

One of C. R. Rao's many important contributions to statistical science was his introduction of the *score test*, based on the derivative of the log-likelihood function at the null hypothesis value of the parameter of interest. This article reviews methods for constructing score tests and score-test-based confidence intervals for analyzing parameters that arise in analyzing categorical data. A considerable literature indicates that score tests and their inversion for constructing confidence intervals perform well in a variety of settings and sometimes much better than Wald-test and likelihood-ratio test-based methods. We also discuss extensions of score-based inference and potential future research on generalizations for longitudinal data, complex sampling, and high-dimensional data.

Keywords Confidence intervals · Generalized linear models · Likelihood-ratio tests · Pearson chi-squared · C.R. Rao · Wald inference

Introduction

In categorical data analysis, hypothesis tests about parameters of interest are typically based on three classes of methods: The likelihood-ratio (LR) test (Wilks 1938), the Wald test (Wald 1943), and the score test proposed by Rao (1948). These statistics can also be utilized to construct confidence intervals (CI) by inverting test statistics about the parameter.

Prepared for the special issue of Journal of Quantitative Economics in honor of Prof. C. R. Rao.

✉ Alan Agresti
aa@stat.ufl.edu

[1] Department of Statistics, University of Florida, Gainesville, FL 32611, USA

[2] Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria, Cosenza, Italy

[3] Department Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence, Italy

Wald tests and CIs are typically considered the standard approach, because of their computational simplicity and easy availability with software. For inversion of the Wald test, the 95% confidence interval is obtained simply by adding to and subtracting from the parameter estimate 1.96 times the estimated standard error. Thanks to this simplicity, early developments of methods in categorical data analysis were Wald-based, such as weighted least squares methods proposed by Grizzle et al. (1969) and many follow-up articles by Gary Koch and his colleagues. However, the LR test and its inversion for confidence intervals are increasingly available in software. Also, Rao's score test, which is based on the derivative of the log-likelihood function at the null hypothesis value of the parameter of interest, is commonly used in some settings. In honour of Professor C. R. Rao for this special issue, our article focuses on the score test and score-test-based confidence intervals in categorical data analysis. We also present recent extensions of it that utilize the breakthrough impact of Rao's contribution.

We begin by summarizing the three methods. For notational simplicity, we present them for the simple case of a single parameter β for a simple statistical model. Denote by $\ell(\beta)$ the associated log-likelihood function and by $\hat{\beta}$ the maximum likelihood (ML) estimate. The score function is

$$u(\beta) = \partial\ell(\beta)/\partial\beta.$$

Linked to it is the Fisher information, $\imath(\beta) = -\mathbb{E}\left[\partial^2\ell(\beta)/\partial\beta^2\right]$, coinciding with the variance of the score function $u(\beta)$, since $\mathbb{E}[u(\beta)] = 0$.

Consider a two-sided significance test of H_0: $\beta = \beta_0$ against H_a: $\beta \neq \beta_0$. The squared version of the Wald test statistic is

$$\left[\frac{\hat{\beta} - \beta_0}{\mathrm{se}(\hat{\beta})}\right]^2 = \left(\hat{\beta} - \beta_0\right)^2 \imath(\hat{\beta}),$$

where the standard error $\mathrm{se}(\hat{\beta})$ of $\hat{\beta}$ and the Fisher information $\imath(\hat{\beta})$ are evaluated at $\hat{\beta}$. The LR test statistic is

$$-2\left[\ell(\beta_0) - \ell(\hat{\beta})\right],$$

comparing the unconstrained log-likelihood function at its maximum $\ell(\hat{\beta})$ with its value at β_0. Rao's score test statistic is

$$\frac{\left[u(\beta_0)\right]^2}{\imath(\beta_0)} = \frac{\left[\partial\ell(\beta)/\partial\beta_0\right]^2}{-\mathbb{E}\left[\partial^2\ell(\beta)/\partial\beta_0^2\right]},$$

with derivatives evaluated at β_0. Its underlying idea is that when H_0 is true, the score function should be relatively near zero at β_0. In some literature, especially in econometrics, Rao's score test is also known as the Lagrange multiplier test, based on Silvey (1959).

Under H_0, all three test statistics have asymptotic chi-squared distributions and are asymptotically equivalent (Cox and Hinkley 1974). When H_0 is false, the three statistics have approximate non-central chi-squared distributions, with different non-centrality parameters. A $100(1-\alpha)\%$ CI can be derived by inverting the tests, constructing the set of β_0 values such that the two-sided significance test has p-value $>$ α. An advantage of Rao's score test is that it applies even when the other two tests cannot be used, for instance when β falls on the boundary of the parametric space under H_0. A disadvantage of the Wald test is its lack of invariance, with results depending on the scale of measurement for β. Likewise, the Wald CI for a nonlinear function $g(\beta)$ of β is not $g(\cdot)$ applied to the Wald CI for β. Thus, in using the Wald method, a wise choice of scale is needed.

In this paper, we introduce frameworks in which score test-based inference is useful for categorical data analysis. "Tests for Categorical Data as Score Tests" summarizes score tests, "Score-Test-Based Confidence Intervals" summarizes score-test based confidence intervals, and "Small-Sample Score-Test-Based Inference" considers small-sample methods. "Extensions of Score-Test-Based Inference for Categorical Data" discusses further extensions such as a "pseudo-score" method that applies when ordinary score methods are not readily available, generalizations for high dimensional cases and for complex and longitudinal settings, and potential future research.

Tests for Categorical Data as Score Tests

Many standard significance tests for categorical data can be derived as score tests that a parameter or a set of parameters equal 0. Methods that construct their estimates of variability under a null hypothesis are often score tests or are closely related to score tests. A landmark example is the Pearson chi-squared test of independence for a two-way contingency table. With cell counts $\{y_{ij}\}$ for a sample of size n and with expected frequency estimates $\{\hat{\mu}_{ij} = y_{i+}y_{+j}/n\}$ based solely on the row and column marginal counts, the Pearson statistic is

$$X^2 = \sum_i \sum_j \frac{\left(y_{ij} - \hat{\mu}_{ij}\right)^2}{\hat{\mu}_{ij}}.$$

Some details follow about other statistics for categorical data. See Agresti (2013) for a summary of the methods mentioned.

Smyth (2003) showed that the Pearson chi-squared statistic X^2 for testing independence in a two-way contingency table is a score statistic, under the assumption that the cell counts are independent and Poisson distributed. For multiway contingency tables, Smyth proved that the score test of the hypothesis that any chosen subset of the pairs of faces in the table are independent yields a Pearson-type statistic. For I independent Binom(n_i, π_i) random variables $\{Y_i\}$ and a binary linear trend model $\pi_i = H(\alpha + \beta x_i)$ with a twice differentiable monotone function H, Tarone and Gart (1980) proved that the score statistic for testing $H_0 : \beta = 0$ does not depend on H. It follows that the Cochran-Armitage test, which is the score test of $H_0 : \beta = 0$ in a linear probability model $\pi_i = \alpha + \beta x$, is equivalent to the score statistic for testing $H_0 : \beta = 0$ in the logistic regression model. The Cochran-Mantel-Haenszel test of conditional independence in $2 \times 2 \times K$ tables that compare two groups on a binary response while adjusting for a categorical covariate is a score test for the logistic model assuming no interaction between the group variable and the categorical covariate (Birch 1964, 1965; Darroch 1981). Day and Byar (1979) demonstrated the equivalence of Cochran-Mantel-Haenszel statistics and score tests for testing independence in case-control studies, investigating the risk associated with a dichotomous exposure and with individuals stratified in groups. Another special case of the Cochran-Mantel-Haenszel test is a score test applied to binary responses of n matched pairs displayed in n partial 2×2 tables, commonly known as McNemar's test. A generalized Cochran-Mantel-Haenszel test for two multicategory variables is the score test for the null hypothesis of conditional independence in a generalized logistic model (Day and Byar 1979). For testing conditional independence in three-way contingency tables that relate a nominal explanatory variable to an ordinal response variable while adjusting for a categorical variable, Iannario and Lang (2016) presented a generalization of the Cochran-Mantel-Haenszel test by proposing score tests based on first moments and constrained correlation.

For generalized linear models with the canonical link function, such as binomial logistic regression models and Poisson loglinear models, the likelihood function simplifies with the data reducing to sufficient statistics. For subject i, letting y_i denote the observed response and x_{ij} the value of the explanatory variable j for which β_j is the coefficient, the sufficient statistic for β_j is $\sum_i x_{ij} y_i$. The score test statistic for $H_0 : \beta_j = 0$ can be expressed as a standardization of its sufficient statistic. In this case, Lovison (2005) gave a formula for the score statistic that resembles the Pearson statistic, being a quadratic form comparing fitted values for two models. Let $\mathbf{X}$ be the model matrix for the full model and let $\widehat{\mathbf{V}}_0$ be the diagonal matrix of estimated variances under the null model (e.g., with $\beta_j = 0$), with fitted values $\widehat{\boldsymbol{\mu}}$ for the full model and $\widehat{\boldsymbol{\mu}}_0$ for the reduced model. Then, the score statistic is

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

Lang et al. (1999) gave this formula for the loglinear case.

Another setting for the Pearson chi-squared statistic occurs in testing model goodness-of-fit. Let $\{y_i\}$ denote multinomial cell counts for a contingency table of arbitrary dimensions. Let $\{\widehat{\mu}_i\}$ be the ML fitted values for a particular model. For testing goodness-of-fit, the score test statistic is the Pearson-type statistic,

$$X^2 = \sum_i \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}.$$

Cox and Hinkley (1974, p. 326) noted this, and Smyth (2003) extended it to a corresponding statistic for generalized linear models.

Score-Test-Based Confidence Intervals

Although the score test is well established for categorical data, CIs based on the score function are less utilized. The best known and most utilized score CI is Wilson's CI for a binomial parameter π (Wilson 1927). This CI uses the score test statistic

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

which has an asymptotic standard normal distribution under H_0: $\pi = \pi_0$. The $100(1 - \alpha)\%$ CI consists of those π_0 values for which the p-value $> \alpha$. For instance, the endpoints of the 95% CI are the roots of the quadratic equation $z_W^2 = (1.96)^2$. By contrast, the Wald CI is based on the z statistic with $\hat{\pi}$ in the denominator instead of π_0. It has midpoint $\hat{\pi}$ and thus zero length whenever $\hat{\pi} = 0$ or 1.

In other settings, score CIs are less commonly known and used. Studies often aim to compare two groups (e.g. different treatments) on a binary response (success, failure), and we focus on that case. A 2 × 2 contingency table, with observed frequencies $\{y_{11}, y_{12}, y_{21}, y_{22}\}$, shows the results, with rows for the groups and columns for response categories. Let $n_1 = y_{11} + y_{12}$, $n_2 = y_{21} + y_{22}$, denote the sample sizes of the two groups. For a subject in row i, $i = 1, 2$, let π_i denote the binomial probability that the response is category 1 (success). For relevant parameters such as the difference of probabilities, the ratio of probabilities, and the odds ratio, Agresti (2011) summarized score-test based CIs. We next summarize them.

Consider a score CI for the difference, $\delta = \pi_1 - \pi_2$. For H_0: $\pi_1 - \pi_2 = \delta_0$, let $\hat{\pi}_1$ and $\hat{\pi}_2$ be the unrestricted ML estimates, which are the sample proportions $\hat{\pi}_i = y_{i1}/n_i$, $i = 1, 2$, and let $\hat{\pi}_1(\delta_0)$ and $\hat{\pi}_2(\delta_0)$ be the ML estimates subject to the constraint $\pi_1 - \pi_2 = \delta_0$. Mee (1984) obtained an asymptotic score CI by inverting the test statistic that is the square of

$$z_{\text{diff}} = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \delta_0}{\sqrt{[\hat{\pi}_1(\delta_0)(1 - \hat{\pi}_1(\delta_0))/n_1] + [\hat{\pi}_2(\delta_0)(1 - \hat{\pi}_2(\delta_0))/n_2]}}.$$

The restricted ML estimates $\hat{\pi}_i(\delta_0)$, $i = 1, 2$, have closed form, but the computation of the set of δ_0 that fall in the CI requires an iterative algorithm. When $\delta_0 = 0$, z_{diff}^2 is the Pearson chi-squared statistic for testing independence. Therefore, when zero is included in this $100(1 - \alpha)\%$ score CI, the Pearson test has p-value $> \alpha$. Miettinen and Nurminen (1985) proposed to multiply z_{diff} by $\left(1 - (n_1 + n_2)^{-1}\right)^{1/2}$ to improve performance with small samples. Newcombe (1998a) proposed another score-test based CI for $\pi_1 - \pi_2$. This interval combines Wilson's individual score CIs for the two proportions. See also Fagerland et al. (2017, Sec. 4.5.4) for details. In large samples, Newcombe's interval tends to be close to Mee's asymptotic score interval, and both have higher actual coverage probabilities than the Wald interval (which inverts the z statistic with unrestricted ML estimates in the standard error), particularly in unbalanced samples ($n_1 \neq n_2$). Fagerland et al. (2015) recommended Newcombe's hybrid score intervals as the best when sample sizes are moderate or large. The Newcombe CI and the Miettinen-Nurminen CI perform similarly, with coverage probabilities close to the nominal level, but the Newcombe CI performs better when proportions are close to the boundaries.

Sometimes it is more informative to consider the ratio, $\phi = \pi_1/\pi_2$, instead of the difference $\delta = \pi_1 - \pi_2$, especially when both π_1 and π_2 are near 0. This ratio, sometimes referred to as the *relative risk*, is estimated by

$$\hat{\phi} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{y_{11}/n_1}{y_{21}/n_2}.$$

Koopman (1984) proposed an asymptotic score CI for the relative risk. Under $H_0 : \pi_1/\pi_2 = \phi_0$, that is $H_0 : \pi_1 = \phi_0\pi_2$, the chi-squared statistic is

$$u_{\mathrm{RR}} = \frac{\left(y_{11} - n_1\hat{\pi}_1(\phi_0)\right)^2}{n_1\hat{\pi}_1(\phi_0)\left(1-\hat{\pi}_1(\phi_0)\right)} + \frac{\left(y_{21} - n_2\hat{\pi}_2(\phi_0)\right)^2}{n_2\hat{\pi}_2(\phi_0)\left(1-\hat{\pi}_2(\phi_0)\right)},$$

where $\hat{\pi}_i(\phi_0)$, $i = 1, 2$, denote the ML estimates of π_1 and π_2 under H_0, which are

$$\begin{aligned}\hat{\pi}_1(\phi_0) &= \frac{\phi_0\left(n_1 + y_{21}\right) + y_{11} + n_2}{2(n_1+n_2)} + \\ &\quad - \frac{\sqrt{\left(\phi_0\left(n_1 + y_{21}\right) + y_{11} + n_2\right)^2 - 4\phi_0\left(n_1+n_2\right)\left(y_{11}+y_{21}\right)}}{2(n_1+n_2)},\end{aligned}$$

and $\hat{\pi}_2(\phi_0) = \hat{\pi}_1(\phi_0)/\phi_0$. The endpoints are the two solutions to $u_{\mathrm{RR}} = \chi^2_{1,1-\alpha}$, equating u_{RR} to the $1-\alpha$ quantile of the chi-squared distribution with one degree of freedom. The CI limits are zero or infinity when a cell count is 0. Miettinen and Nurminen (1985) proposed another asymptotic score CI for ϕ, by inverting under H_0,

$$z_{\mathrm{MN}} = \frac{1}{s}\left(\hat{\pi}_1 - \phi_0\hat{\pi}_2\right)\sqrt{1 - \frac{1}{n_1+n_2}},$$

where

$$s = \sqrt{\frac{\hat{\pi}_1(\phi_0)\left(1-\hat{\pi}_1(\phi_0)\right)}{n_1} + \frac{\phi_0^2\,\hat{\pi}_2(\phi_0)\left(1-\hat{\pi}_2(\phi_0)\right)}{n_2}}.$$

Also, Zou and Donner (2008) proposed a hybrid score CI combining the Wilson's score CIs of each parameter π_1 and π_2 and exploiting the logarithm of a ratio equaling the difference of the logarithms. See also Sec. 4.7.5 of Fagerland et al. (2017). Fagerland et al. (2015) recommended Koopman's asymptotic CI for small, moderate and large sample size. According to a comparative study by Price and Bonett (2008), this CI performs very well, with coverage probabilities always close to the nominal level, and from this point of view, it is clearly superior to other non-score-based intervals.

The *odds ratio* is a parameter of special interest for categorical data, because it is linked to the coefficient of an explanatory variable in logistic regression via exponentiation. In a 2 × 2 table with two independent binomials, the odds ratio is

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)},$$

estimated by $(y_{11}y_{22})/(y_{12}y_{21})$. To construct a score-test-based CI for θ, for a given θ_0, let $\{\hat{\mu}_{ij}(\theta_0)\}$ be the unique values that have the same row and column margins as $\{y_{ij}\}$ and such that

$$\frac{\hat{\mu}_{11}(\theta_0)\hat{\mu}_{22}(\theta_0)}{\hat{\mu}_{12}(\theta_0)\hat{\mu}_{21}(\theta_0)} = \theta_0.$$

The set of θ_0 satisfying

$$X^2 = \sum_{ij}(y_{ij} - \hat{\mu}_{ij}(\theta_0))^2/\hat{\mu}_{ij}(\theta_0) \leq \chi^2_{1,1-\alpha},$$

forms a $100(1-\alpha)\%$ conditional score CI for the odds ratio (Cornfield 1956) that also applies for a multinomial sample over the four cells.

The research literature suggests that asymptotic score tests and corresponding CIs perform well, usually much better than Wald CIs. Even with small samples, score CIs perform surprisingly well and often out-perform likelihood-ratio-test-based inference. This behavior may be a consequence of the score statistic in canonical models being the standardization of a sufficient statistic that uses standard errors computed under H_0. For evaluations based on simulations, see Fagerland et al. (2015) and Fagerland et al. (2017). In addition, for comparisons specific to CIs for binomial proportions, see Newcombe (1998b) and Agresti and Coull (1998). See Miettinen and Nurminen (1985), Newcombe (1998a), and Agresti and Min (2005a) for comparison of CIs for the difference of proportions and the 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, Ryu and Agresti (2008) for effect measures comparing two groups on an ordinal scale, Lang (2008) for logistic regression parameters and generic measures of association, and Tang (2020) for score CIs for stratified comparisons of binomial proportions.

Statistical software provides functions for computing score CIs. The Appendix lists some useful R (R Core Team 2022) functions.

Small-Sample Score-Test-Based Inference

Asymptotic tests based on large-sample approximations may perform poorly for small n, although research suggests that score tests often perform well even in quite small samples. One can instead perform tests and construct CIs by applying relevant small-sample distributions, such as the binomial. For instance, consider inference for a coefficient β_j in a logistic regression model

$$\text{logit}\left[P(Y_i = 1)\right] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

for binary response Y_i. With $x_{i0} = 1$ for the intercept, the score statistic for β_j is based on the sufficient statistic $T_j = \sum_i x_{ij} y_i$. Starting with the binomial likelihood for independent observations, one can base a test on the conditional distribution of T_j after eliminating the other nuisance parameters by conditioning on their sufficient statistics. For example, with the equal-tail method, bounds (β_{1L}, β_{1U}) of a $100(1-\alpha)\%$ CI for β_1 are obtained by solving the two equations

$$P(T_1 \geq t_{1,obs} | t_0, t_2, \dots, t_k) = \alpha/2,$$
$$P(T_1 \leq t_{1,obs} | t_0, t_2, \dots, t_k) = \alpha/2.$$

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

Discreteness implies that a significance test cannot have a fixed size α, at all possible null values for a parameter. In rejecting the null hypothesis whenever the p-value $\leq \alpha$, the actual size has α as an upper bound. Hence, actual confidence levels for small-sample interval estimation inverting such tests do not exactly equal the nominal values. Inferences are *conservative*, in the sense that coverage probabilities are bounded below by the nominal level and CIs are wider than ideal. The actual coverage probability varies for different parameter values and is unknown.

Agresti (2003) shows some remedies to alleviate the conservatism. One approach that is feasible when the parameter space is small 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 $100(1-\alpha)\%$ CI consists of β_0 for which $\sup_\psi p(\beta_0;\psi) > \alpha$. Chan and Zhang (1999) proposed an exact unconditional interval for $\pi_1 - \pi_2$ by inverting two one-sided exact score tests of size at most $\alpha/2$ each. Agresti and Min (2001) inverted a single two-sided exact unconditional score test, which results in a narrower interval, available in StatXact software (Cytel 2005). Agresti and Min (2002) found that the unconditional exact approach with two-sided score statistic also works well for the odds ratio. See Fagerland et al. (2017) for Chan-Zhang and Agresti-Min forms of CIs for $\pi_1 - \pi_2$ (pp. 118–119), the relative risk (pp. 139–141), and the odds ratio (pp. 159–160). Coe and Tamhane (1993) proposed an alternative unconditional approach for $\pi_1 - \pi_2$ and π_1/π_2 that is more complex but performs well. Santner et al. (2007) reviewed several such methods.

Agresti and Gottard (2007) showed that an alternative way to reduce conservativeness with discrete data is to base tests and CIs on the *mid-P-value* (Lancaster 1961). For testing $H_0 : \beta = \beta_0$ versus $H_a : \beta > \beta_0$ based on a discrete test statistic T such as a score statistic, the mid-P-value is

$$P(T > t_{obs} \mid H_0) + \frac{1}{2} P(T = t_{obs} \mid H_0).$$

Under H_0, the ordinary p-value is stochastically larger than uniform(0,1) in distribution (which is the exact distribution in the continuous case), but the mid-P-value is not and it has the same mean and a slightly smaller variance than a uniform random variable. The sum of right-tail and left-tail p-values equals $1 + P(T = t_{obs} \mid H_0)$ for the ordinary p-value but equals 1 for the mid-P-value. Using the small-sample distribution, a $100(1-\alpha)\%$ mid-P-value CI (β_L, β_U) for β is determined by

$$P_{\beta_U}(T < t_{obs}) + (1/2) \times P_{\beta_U}(T = t_{obs}) = \alpha/2,$$
$$P_{\beta_L}(T > t_{obs}) + (1/2) \times P_{\beta_L}(T = t_{obs}) = \alpha/2.$$

Its coverage probability is not guaranteed to be $\geq (1-\alpha)$, but it is usually close to that value. Numerical evaluations, such as in Agresti and Gottard (2007), suggest that it tends to be a bit conservative in an average sense.

Several examples in Fagerland et al. (2017) show the good performance of mid-P-based inference, compared with commonly used methods. An example is the Cochran-Armitage mid-P score test and related CIs for trend in logit models for $I \times 2$ contingency tables with ordered rows and possibly small samples (Fagerland et al. 2017, Table 5.12, p. 221). Other mid-P versions of score-type tests include a mid-P version of the Pearson chi-square test statistic for independence in unordered $I \times J$ tables, and the McNemar mid-P test. This latter test does not violate the nominal level in all of the 10000 scenarios evaluated by Fagerland et al. (2013).

An alternative small-sample approach uses asymptotic statistics but employs a *continuity correction*, to better align standard normal tail probabilities with the binomial tail probabilities used to construct exact intervals and exact tests. However, because such exact tests are conservative, doing this provides some protection from the actual coverage probability being too low but sacrifices performance in terms of length, with the average coverage probability being too high.

Extensions of Score-Test-Based Inference for Categorical Data

This section describes some extensions of score-test-based inference and potential future research about such methods.

Pseudo-Score Inference with the Pearson Chi-Squared Statistic

Consider a multinomial model for cell counts $\{y_i\}$ and its ML fitted values $\{\hat{\mu}_i\}$. Consider a simpler, *null* model obtained from the full model by imposing a constraint on a model parameter, say $\beta = \beta_0$, with ML fitted values $\{\hat{\mu}_{i0}\}$. The LR statistic

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

can be used to compare the models and to construct the profile likelihood $100(1-\alpha)\%$ CI for β, which is the set of $\{\beta_0\}$ such that $G^2 \leq \chi^2_{1,1-\alpha}$. To provide a score-type CI, Agresti and Ryu (2010) proposed instead inverting a Pearson-type statistic proposed by Rao (1961),

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

This statistic is a quadratic approximation for G^2 and is equivalent to the Pearson statistic for goodness-of-fit testing when the full model is the saturated one. Haberman (1977) showed that under $H_0 : \beta = \beta_0$, X^2 has the same limiting distribution as G^2 for large, sparse tables. This includes the case in which the number of cells in the table grows with the sample size, such as occurs with a continuous explanatory variable.

Agresti and Ryu (2010) proposed the asymptotic $100(1-\alpha)\%$ CI for a generic parameter β as the set of β_0 values such that $X^2 \leq \chi^2_{1,1-\alpha}$. This is a *pseudo-score* CI, as X^2 is the score test statistic only if the full model is saturated. Agresti and Ryu (2010) noted that the pseudo-score CI is available even when the score CI itself is not easily obtainable. In addition, the pseudo-score method generalizes to sampling schemes more complex than simple multinomial sampling and to discrete distributions other than the multinomial, such as Poisson regression models. For generalized linear models with canonical link function and independent observations $\{y_i\}$ from a specified discrete distribution, Lovison (2005) showed that the bounds obtained by inverting the generalized Pearson-type statistic (see "Tests for Categorical Data as Score Tests") are bounded above by the Pearson statistic. Consequently, the 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. Nonetheless, the pseudo-score method is useful when ordinary score methods are not practical, such as in more complex cases or when the link function is not canonical. In these situations, the pseudo-score CIs can be implemented with the same difficulty level as profile likelihood confidence intervals. Through simulations, Agresti and Ryu (2010) found that the pseudo score method has similar behavior as the profile likelihood interval and sometimes with even a bit better performance with small samples. Also, as discussed in the next subsection, extensions of the pseudo-score method may apply to settings in which profile likelihood methods are not available.

The pseudo-score method generalizes to parameters of generalized linear models for discrete data, for instance in Poisson and negative binomial regression. Suppose $\{Y_i, i = 1, \dots, n\}$ are independent observations assumed to have a specified discrete distribution. Let $v(\hat{\mu}_{i0})$ denote the estimated variance of Y_i assuming the null distribution for Y_i and let $\hat{\mathbf{V}}_0$ be the diagonal matrix containing such values. A Pearson-type statistic for comparing models in the generalized linear model setting (Lovison 2005) has the form

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

This statistic also applies to a quasi-likelihood setting, in which one needs only to specify the expected values under the assumed models and the variance function (or more generally a matrix of covariance functions), without specifying a particular distribution (Lovison 2005).

Other Extensions of Score-Test-Based Inference

For longitudinal data and many other forms of clustered data, score tests are not readily available for popular models. A prime example is the set of models for which the likelihood function is not an explicit function of the model parameters, such as marginal models for longitudinal data. 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 estimators. Boos (1992) and Rotnitzky and Jewell (1990) presented score-type tests for this setting.

In future research, the pseudo-score inference presented in "Pseudo-Score Inference with the Pearson Chi-Squared Statistic" may also extend to marginal modeling of clustered categorical responses. 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 $\boldsymbol{\mu}_i = \mathbb{E}(\mathbf{Y}_i) = (\mu_{i1}, \ldots, \mu_{iT_i})'$. Let $\mathbf{V}_i$ denote the $T_i \times T_i$ covariance matrix of $\mathbf{Y}_i$. For a particular marginal model, let $\widehat{\boldsymbol{\mu}}_i$ denote an estimate of $\boldsymbol{\mu}_i$, such as the ML estimate under the naive assumption that the $\sum_i T_i$ observations are independent. Let $\widehat{\boldsymbol{\mu}}_{i0}$ denote the corresponding estimate under the constraint that a particular parameter β takes value β_0. Let $\widehat{\mathbf{V}}_{i0}$ denote an estimate of the covariance matrix of $\mathbf{Y}_i$ under this null model. The main diagonal elements of $\widehat{\mathbf{V}}_{i0}$ are $\widehat{\mu}_{it0}(1 - \widehat{\mu}_{it0})$, $t = 1, \ldots, T_i$. Separate estimation is needed for the null covariances, which are not part of the marginal model. Now, consider the statistic

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

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 $X^2 \leq \chi^2_{1,1-\alpha}$ 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 large. 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. It is of interest to analyze whether the chi-squared distribution tends to provide a good approximation. Extensions are possible for correlated discrete cases other than correlated categorical responses. As pointed out by Lovison (2005), unlike likelihood ratio test-type statistics, a Pearson-type statistic can be defined for any quasi-likelihood model, needing only to specify expected values under the model and variance-covariance functions.

Many research studies, especially those using surveys, obtain data 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 $\widehat{V}_0$ in the Pearson-type statistic just mentioned by an appropriately inflated or non-diagonal estimate of the covariance matrix. For such complex sampling designs, profile likelihood CIs are not available and need to be replaced by quasi-likelihood adaptations.

In one approach of this type, Rao and Scott (1981) proposed an extension of the Pearson chi-squared statistic for testing independence in a two-way contingency table when the data result from a complex survey design and the observations cannot be treated as realizations of *iid* random variables. In particular, they provide a correction of $\widehat{V}_0$ for stratified random sampling and two-stage sampling. However, their test statistic requires that none of the observed cell counts equals zero. To solve this limitation, Lipsitz et al. (2015) proposed Wald and score statistics for testing independence based on weighted least squares estimating equations.

Another possible extension of score-based inference concerns constrained statistical inference. Constrained statistical inference problems arise in categorical data analysis when there are inequality constraints on parameters, such as functions of conditional probabilities in a $I \times J$ table. They are used to specify hypotheses of stochastic dominance, monotone dependence and positive association in contingency tables (Agresti and Coull 1998; Dardanoni and Forcina 1998; Bartolucci et al. 2007, among others). To test them, the literature on constrained inference for categorical data (see Colombi and Forcina 2016, and the references therein quoted) concentrated on the LR statistic and its asymptotic chi-bar-squared distribution, a weighted sum of chi-squared variables whose weights can be calculated exactly or sufficiently precisely via simulation (see R package `ic-infer` by Grömping 2010, and `hmmm` by Colombi et al. 2014).

Silvapulle and Sen (2005) presented an extensive review on testing under inequality restrictions, and described two possible ways (global and local) to extend score statistics for inequality constrained testing problems, giving proofs of the asymptotic equivalence, under some conditions, of these score-type and LR statistics. However, the LR seems more used in constrained inference, possibly because of analytical and computational advantages (e.g. Molenberghs and Verbeke 2007). A research challenge could be in the direction of investigating, also through simulations, where score-type testing is convenient.

Inference in High-Dimensional Settings

In high-dimensional settings, the number of parameters can be very large, sometimes even exceeding the sample size. Then, a fundamental issue is to derive the theoretical properties of regularized estimators such as those using a lasso-type (Tibshirani 1996) penalty term. While several properties on regularized point estimators have been assessed, methods to adequately quantify estimate uncertainty and derive confidence intervals is an important topic under investigation, usually referred as *post selection inference* or *selective inference*. Classical inferential theory is not valid. Even if interest focuses only on few parameters with the others considered a nuisance, the score function is seriously affected by the dimension of the nuisance parameter. Recent developments explore how extensions of Rao's score test function can be utilized both for hypothesis testing and confidence intervals in high-dimensional generalized linear models.

A key contribution is due to Ning et al. (2017). For a subset of parameters of interest, they proposed a new device, called a *decorrelated score function*, that can be used with high dimensional logistic and Poisson regression models, among others. To illustrate, suppose the assumed model is characterized by a set of parameters $\boldsymbol{\theta}$ that can be partitioned as $\boldsymbol{\theta} = (\beta, \boldsymbol{\gamma})$, where β is a finite-dimensional parameter of interest and $\boldsymbol{\gamma}$ is a high-dimensional nuisance parameter. Ning et al. (2017) applied a decorrelation operation to the score function, obtaining a score function for β that is uncorrelated with the nuisance score function. The decorrelated score test can be viewed as an extension of Rao's score test, and it is equivalent to this in a low-dimensional setting. For instance, consider a logistic regression model with covariates $\boldsymbol{Q} = (z, \boldsymbol{x})' \in \mathbb{R}^p$, where z is the variable of interest with coefficient β and $\boldsymbol{x}$ are other covariates, with coefficients $\boldsymbol{\gamma}$, assumed as sparse. Then, the log-likelihood function is

$$\ell(\theta, \boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i\left(\theta z_i + \boldsymbol{\gamma}'\boldsymbol{x}_i\right) + \log\left[1 + \exp\left(\theta z_i + \boldsymbol{\gamma}'\boldsymbol{x}_i\right)\right] \right\}.$$

Ning et al. (2017) showed that when $\boldsymbol{\gamma}$ has high dimension, Rao's score test statistic with maximum likelihood and regularized estimator fails its asymptotic optimal properties. In particular, the score functions no longer have a simple limiting distribution. The decorrelated score function of β is defined as

$$S(\beta, \boldsymbol{\gamma}) = u_\beta(\beta, \boldsymbol{\gamma}) - \boldsymbol{w}'\, u_{\boldsymbol{\gamma}}(\beta, \boldsymbol{\gamma}),$$

where $u_\beta(\beta, \boldsymbol{\gamma}) = \partial\ell(\beta, \boldsymbol{\gamma})/\partial\beta$ is the score function with respect to β, and $\boldsymbol{w} = \imath(\beta, \boldsymbol{\gamma})\, \imath(\boldsymbol{\gamma})^{-1}$. The resulting function $S(\beta, \boldsymbol{\gamma})$ is uncorrelated with the score function for the nuisance parameters $u_{\boldsymbol{\gamma}}(\beta, \boldsymbol{\gamma})$.

The score test for β requires an estimate for both $\boldsymbol{\gamma}$ and $\boldsymbol{w}$ to compute the test statistic $\widehat{S}(\beta, \widehat{\boldsymbol{\gamma}})$ to be evaluated under $H_0 : \beta = \beta_0$. Ning et al. (2017) proposed an algorithm for such computation and showed that it applies to several models, to several regularized estimators, and also for a multi-dimensional parameter of interest. Under H_0, the test statistic

$$z_{DS} = \sqrt{n}\, \widehat{S}(\beta_0, \widehat{\boldsymbol{\gamma}}) \,/\, \sqrt{\widehat{\sigma}_S},$$

with $\widehat{\sigma}_S$ a consistent estimator of the variance of the decorrelated score function has, asymptotically, a standard normal distribution, under mild assumptions. In comparison with Wald-type tests for high dimensional settings, such as the desparsifying method (Van de Geer et al. 2014), the decorrelated score test was shown through simulation to be slightly more powerful. The decorrelated score function can also generate valid confidence intervals for the parameters of interest (Shi et al. 2020).

High-dimensional data typically are sparse data, which can cause problems such as infinite estimates in models for categorical data because of complete separation or quasi-complete separation. Generally, with sparse data or infinite maximum likelihood estimates, it is popular to use Firth's penalized-likelihood approach (Firth 1993). Siino et al. (2018) have developed the penalized score statistic test for logistic regression in the presence of sparse data by modifying the classical score function to partly remove the bias of the ML estimates due to sparseness. In particular, for logistic regression parameters, the authors showed through simulations that the score-based CIs with Firth's penalization perform better than some competitors such as Wald and likelihood ratio statistics in terms of coverage level and average width, even with small samples, strong sparsity, and sampling zeros, and also for any number of covariates in the model.

Appendix: Score-Based Tests and Confidence Intervals in R

The Wald test and CIs and likelihood-ratio-test-based profile likelihood CIs are easily accessible with statistical software, while score test and CIs are less commonly used and not generally available as default in the statistical software packages. However, functions implementing score-type inference for parameters in basic settings can be easily written by users, and such functions are increasingly available also in some computationally demanding contexts. In the following, we list some R functions {packages} to calculate score-type tests and CIs, for the parameters of interest in the categorical data analysis, in basic and more advanced methodological approaches. The list is not exhaustive.

Functions in the R Package `propCIs`

- `diffscoreci`
  Score CI for difference of proportions with independent samples
- `riskscoreci`
  Score CI for the relative risk in a 2 × 2 table
- `orscoreci`
  Score CI for an odds ratio in a 2 × 2 table
- `scoreci.mp`
  Score confidence interval (Tango 1998) for a difference of proportions with matched-pairs data
- `scoreci`
  Wilson's confidence interval (Wilson 1927) for a single proportion
- `midPci`
  mid-P confidence interval adaptation of the Clopper-Pearson exact interval

Functions in Other R Packages

- `binom.conf.int` {epitools}
  Calculates Wilson confidence intervals for binomial parameters
- `binconf` {Hmisc}
  CI for proportion with “wilson" option for score CI
- `score.stat` {VGAM}

Generic function that computes Rao's score test statistics evaluated at the null values.

- `scoretest` {StepReg}
  This function can compute score test statistic and *p*-value for a linear model when one adds an explanatory variable
- `score_test` {tram}
  p-values and confidence intervals for parameters in linear transformation models (Hothorn et al. 2018) obtained by the score test principle
- `summarylr` {glmglrt}
  summarylr is an improved summary function for standard glm (stats package) adding LRT or Rao score *p*-values
- `HypoTest` {CompRandFld}
  The function performs statistical hypothesis tests for nested models based on composite likelihood versions of: Wald-type, score-type and Wilks-type (LR) statistics
- `sig` {LogisticDx}
  Significance tests (LR, Score, Wald) for a binary regression model fit with glm
- `confint2` {glmtoolbox}
  CIs based on Wald, likelihood-ratio, Rao's score tests for a generalized linear model
- `anova2` {glmtoolbox}
  Can compare nested generalized linear models using Wald, score, and LR tests
- `glm.scoretest` {statmod}
  Computes score test statistic for adding covariate to a generalized linear model
- `confint_contrast` {glmglrt}
  Can compute contrasts of fixed-effects in many models. The default implementation computes Wald's confidence intervals with any model. It has specialized use for GLMs with Wald's, LRT and score CIs and may be used with other models.
- `binom.midp` {binomSamSize}
  Calculate mid-p confidence interval for binomial proportion
- `exact.test` {exact}
  Unconditional exact tests for 2 × 2 tables with independent samples
- `binomDiffCI` {MKinfer}
  Confidence intervals for difference of two binomial proportions
- `scoreci` {ratesci}
  Score confidence intervals for comparisons of independent binomial rates
- `pairbinci` {ratesci}
  Confidence intervals for comparisons of paired binomial rates
- `bgtCI` {binGroup}
  Confidence intervals for a proportion in binomial group testing
- `binomTest` {conf}
  Confidence intervals for binomial proportions

Acknowledgements Thanks to a referee for helpful comments.

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 (1): 3–21.

Agresti, A. 2011. Score and pseudo-score confidence intervals for categorical data analysis. *Statistics in Biopharmaceutical Research* 3 (2): 163–172.

Agresti, A. 2013. *Categorical Data Analysis*. Oxford: Wiley.

Agresti, A., M. Bini, B. Bertaccini, and E. Ryu. 2008. Simultaneous confidence intervals for comparing binomial parameters. *Biometrics* 64 (4): 1270–1275.

Agresti, A., and B. Coull. 1998. Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician* 52: 119–126.

Agresti, A., and A. Gottard. 2007. Nonconservative exact small-sample inference for discrete data. *Computational Statistics and Data Analysis* 51 (12): 6447–6458.

Agresti, A., and B. Klingenberg. 2005. Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54 (4): 691–706.

Agresti, A., and Y. Min. 2001. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* 57 (3): 963–971.

Agresti, A., and Y. Min. 2005. Frequentist performance of Bayesian confidence intervals for comparing proportions in 2× 2 contingency tables. *Biometrics* 61 (2): 515–523.

Agresti, A., and Y. Min. 2005. Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine* 24 (5): 729–740.

Agresti, A., and E. Ryu. 2010. Pseudo-score confidence intervals for parameters in discrete statistical models. *Biometrika* 97 (1): 215–222.

Bartolucci, F., R. Colombi, and A. Forcina. 2007. An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica* 17 (2): 691–711.

Birch, M. 1964. The detection of partial association, I: the 2×2 case. *Journal of the Royal Statistical Society: Series B (Methodological)* 26 (2): 313–324.

Birch, M. 1965. The detection of partial association, II: the general case. *Journal of the Royal Statistical Society: Series B (Methodological)* 27 (1): 111–124.

Boos, D. 1992. On generalized score tests. *The American Statistician* 46 (4): 327–333.

Chan, I.S., and Z. Zhang. 1999. Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* 55 (4): 1202–1209.

Coe, P.R., and A.C. Tamhane. 1993. Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Communications in Statistics - Simulation and Computation* 22 (4): 925–938. https://doi.org/10.1080/03610919308813135

Colombi, R., and A. Forcina. 2016. Testing order restrictions in contingency tables. *Metrika* 79 (1): 73–90.

Colombi, R., S. Giordano, and M. Cazzaro. 2014. hmmm: an R package for hierarchical multinomial marginal models. *Journal of Statistical Software* 59 (1): 1–25.

Cornfield, J. 1956. A statistical problem arising from retrospective studies. In Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, pp. 135–148. ed. J. Neyman.

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

Cytel,. 2005. *StatXact 7 User Manual and LogXact 7 User Manual*. Cambridge, Massachusetts: Cytel Inc.

Dardanoni, V., and A. Forcina. 1998. A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *Journal of the American Statistical Association* 93 (443): 1112–1123.

Darroch, J.N. 1981. The Mantel-Haenszel test and tests of marginal symmetry; fixed-effects and mixed models for a categorical response, correspondent paper. *International Statistical Review/Revue Internationale de Statistique* 49 (3): 285–307.

Day, N., and D. Byar. 1979. Testing hypotheses in case-control studies-equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics* 35 (3): 623–630.

Fagerland, M., S. Lydersen, and P. Laake. 2017. *Statistical Analysis of Contingency Tables*. Boca Raton: CRC Press.

Fagerland, M.W., S. Lydersen, and P. Laake. 2013. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology* 13 (1): 1–8.

Fagerland, M.W., S. Lydersen, and P. Laake. 2015. Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research* 24 (2): 224–254.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80 (1): 27–38.

Grizzle, J.E., C.F. Starmer, and G.G. Koch. 1969. Analysis of categorical data by linear models. *Biometrics* 25: 489–504.

Grömping, U. 2010. Inference with linear equality and inequality constraints using R: The package ic-infer. *Journal of Statistical Software* 33 (1): 1–31.

Haberman, S.J. 1977. Log-linear models and frequency tables with small expected cell counts. *The Annals of Statistics* 5: 1148–1169.

Hothorn, T., L. Moest, and P. Buehlmann. 2018. Most likely transformations. *Scandinavian Journal of Statistics* 45 (1): 110–134.

Iannario, M., and J.B. Lang. 2016. Testing conditional independence in sets of I×J tables by means of moment and correlation score tests with application to hpv vaccine. *Statistics in Medicine* 35 (25): 4573–4587.

Koopman, P. 1984. Confidence intervals for the ratio of two binomial proportions. *Biometrics* 40: 513–517.

Lancaster, H.O. 1961. Significance tests in discrete distributions. *Journal of the American Statistical Association* 56 (294): 223–234.

Lang, J.B. 2008. Score and profile likelihood confidence intervals for contingency table parameters. *Statistics in Medicine* 27 (28): 5975–5990.

Lang, J.B., J.W. McDonald, and P.W. Smith. 1999. Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach. *Journal of the American Statistical Association* 94 (448): 1161–1171.

Lipsitz, S.R., G.M. Fitzmaurice, D. Sinha, N. Hevelone, E. Giovannucci, and J.C. Hu. 2015. Testing for independence in contingency tables with complex sample survey data. *Biometrics* 71 (3): 832–840.

Lovison, G. 2005. On Rao score and Pearson χ^2 statistics in generalized linear models. *Statistical Papers* 46 (4): 555–574.

Mee, R.W. 1984. Confidence bounds for the difference between two probabilities (letter). *Biometrics* 40: 1175–1176.

Mehta, C.R., and N.R. Patel. 1995. Exact logistic regression: Theory and examples. *Statistics in Medicine* 14 (19): 2143–2160.

Miettinen, O., and M. Nurminen. 1985. Comparative analysis of two rates. *Statistics in Medicine* 4 (2): 213–226.

Molenberghs, G., and G. Verbeke. 2007. Likelihood ratio, score, and wald tests in a constrained parameter space. *The American Statistician* 61 (1): 22–27.

Newcombe, R.G. 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 17 (8): 873–890.

Newcombe, R.G. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 17 (8): 857–872.

Ning, Y., H. Liu, et al. 2017. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics* 45 (1): 158–195.

Price, R.M., and D.G. Bonett. 2008. Confidence intervals for a ratio of two independent binomial proportions. *Statistics in Medicine* 27 (26): 5497–5508.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. R. 1948. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In Mathematical Proceedings of the Cambridge Philosophical Society, Volume 44: 50–57. Cambridge University Press.

Rao, C. R. 1961. A study of large sample test criteria through properties of efficient estimates: Part I: Tests for goodness of fit and contingency tables. *Sankhyā*: *The Indian Journal of Statistics, Series A* 23(1): 25–40.

Rao, J.N., and A.J. Scott. 1981. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association* 76 (374): 221–230.

Rotnitzky, A., and N.P. Jewell. 1990. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 77 (3): 485–497.

Ryu, E., and A. Agresti. 2008. Modeling and inference for an ordinal effect size measure. *Statistics in Medicine* 27 (10): 1703–1717.

Santner, T.J., V. Pradhan, P. Senchaudhuri, C.R. Mehta, and A. Tamhane. 2007. Small-sample comparisons of confidence intervals for the difference of two independent binomial proportions. *Computational Statistics and Data Analysis* 51 (12): 5791–5799.

Shi, C., R. Song, W. Lu, and R. Li. 2020. Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association* 116 (535): 1–12.

Siino, M., S. Fasola, and V.M. Muggeo. 2018. Inferential tools in penalized logistic regression for small and sparse data: A comparative study. *Statistical Methods in Medical Research* 27 (5): 1365–1375.

Silvapulle, M.J., and P.K. Sen. 2005. *Constrained Statistical Inference: Inequality, Order and Shape Restrictions*. Oxford: Wiley.

Silvey, S.D. 1959. The Lagrangian multiplier test. *The Annals of Mathematical Statistics* 30 (2): 389–407.

Smyth, G. K. 2003. Pearson's goodness of fit statistic as a score test statistic. In *Statistics and Science: a Festschrift for Terry Speed*. Lecture notes-monograph series 40: 115–126, Institute of Mathematical Statistics, Hayward, CA.

Tang, Y. 2020. Score confidence intervals and sample sizes for stratified comparisons of binomial proportions. *Statistics in Medicine* 39 (24): 3427–3457.

Tango, T. 1998. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine* 17 (8): 891–908.

Tarone, R.E., and J.J. Gart. 1980. On the robustness of combined tests for trends in proportions. *Journal of the American Statistical Association* 75 (369): 110–116.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.

Van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42 (3): 1166–1202.

Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* 54 (3): 426–482.

Wilks, S.S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 9 (1): 60–62.

Wilson, E.B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22 (158): 209–212.

Zou, G., and A. Donner. 2008. Construction of confidence limits about effect measures: A general approach. *Statistics in Medicine* 27 (10): 1693–1702.