STATISTICS IN MEDICINE
Statist. Med. 2007; **00**:1–15

Modeling and inference for an ordinal effect size measure

Euijung Ryu *, † and Alan Agresti

Department of Statistics, University of Florida, Gainesville, FL 32611, U.S.A.

SUMMARY

An ordinal measure of effect size is a simple and useful way to describe the difference between two ordered categorical distributions. This measure summarizes the probability that an outcome from one distribution falls above an outcome from the other, adjusted for ties. We develop and compare confidence interval methods for the measure. Simulation studies show that with independent multinomial samples, confidence intervals based on inverting the score test and a pseudo score-type test perform well. This score method also seems to work well with fully-ranked data, but for dependent samples a simple Wald interval on the logit scale can be better with small samples. We also explore how the ordinal effect size measure relates to an effect measure commonly used for normal distributions, and we consider a logit model for describing how it depends on explanatory variables. The methods are illustrated for a study comparing treatments for shoulder tip pain. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: confidence intervals; logit models; Mann-Whitney statistic; matched pairs; multinomial distributions; ordinal data

1. INTRODUCTION

This article considers use of a summary measure to describe the difference between two groups, for observations on an ordered categorical scale. We illustrate using Table I, from a study (Lumley [1]) to compare an active treatment with a control treatment for patients having shoulder tip pain after laparoscopic surgery. The two treatments were randomly assigned to 41 patients. The patients rated their pain level on a scale from 1 (low) to 5 (high) on the fifth day after the surgery.

In practice, when responses are ordered categorical, a common approach is to assign scores to the categories and use methods for comparing means. A modeling approach, such as a cumulative logit model with a proportional odds structure, treats the response as ordinal rather than interval-scale and provides an odds ratio summary using cumulative probabilities.

*Correspondence to: Euijung Ryu, Department of Statistics, University of Florida, Gainesville, FL 32611, U.S.A.

†E-mail: eryu@stat.ufl.edu

Received 3 May 2007
Revised 25 June 2007

In this article we use an alternative measure that treats the response as ordinal but is simpler to interpret for audiences not familiar with odds ratios and that has connections with a commonly used effect size measure for normal distributions. We develop confidence intervals and models for this ordinal effect measure.

The next section introduces the measure and reviews existing confidence interval methods for it. Section 3 proposes other confidence interval methods for the case of independent multinomial samples. Section 4 presents related confidence intervals based on a cumulative logit model. In Section 5, we apply the methods to Table I and some other data sets. Section 6 summarizes simulation studies comparing the methods. Section 7 adapts the confidence interval methods and assesses their performance for fully-ranked data and matched-pairs data, and discusses the connection with the normal effect size measure. Section 8 presents a logit model for the measure with explanatory variables.

2. THE ORDINAL EFFECT SIZE MEASURE

Let Y_1 and Y_2 denote independent random variables that each have the same ordinal scale. A measure that summarizes their relative size without assuming magnitudes for the categories is

$$\theta = P(Y_1 < Y_2) + 0.5P(Y_1 = Y_2).$$

Klotz [2] used θ in testing the hypothesis of equality of the distributions of Y_1 and Y_2 against alternatives for which one is stochastically larger than the other. Vargha and Delaney [3] called θ *a measure of stochastic superiority of* Y_2 *over* Y_1. Bamber [4] showed that θ is the same as the area under a receiver operating characteristic (ROC) curve.

For c outcome categories, we label the categories $1, 2, \cdots, c$, from least to greatest in degree. Let $\pi_i = P(Y_1 = i)$ and $\lambda_i = P(Y_2 = i)$, $i = 1, 2, \cdots, c$, with $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_c)'$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_c)'$. The measure is

$$\theta = \boldsymbol{\lambda}' A \boldsymbol{\pi}, \tag{1}$$

where

$$A = \begin{pmatrix} 0.5 & 0 & & \cdots & 0 \\ 1 & 0.5 & 0 & \cdots & 0 \\ & & \vdots & & \\ 1 & \cdots & 1 & 0.5 & 0 \\ 1 & \cdots & & 1 & 0.5 \end{pmatrix}.$$

When $c = 2$, θ equals $0.5(1 + \pi_1 - \lambda_1)$, a linear function of the difference of proportions, $\pi_1 - \lambda_1$. If categories are reversed in order or if Y_1 and Y_2 are interchanged, then θ changes to $1 - \theta$. If Y_1 and Y_2 are identically distributed or if they both have symmetric distributions (that is, $\pi_1 = \pi_c$, $\pi_2 = \pi_{c-1}$, $\ldots$, and likewise for $\boldsymbol{\lambda}$), then $\theta = 0.50$. If Y_2 is stochastically larger (smaller) than Y_1, then $\theta > 0.50$ (< 0.50). To test that the distributions are identical against $H_a : \theta \neq 0.50$, $H_a : \theta > 0.50$, or $H_a : \theta < 0.50$, one can use the test of Mann and Whitney [5] or the equivalent Wilcoxon test [6]. A likelihood ratio test is available for testing $H_0 : \theta = 0.50$, which contains the null hypothesis that two distributions are identical (Troendle [7]).

Instead of testing a single value for θ, this article focuses on constructing confidence intervals for θ. Hochberg [8] proposed confidence intervals for $P(Y_1 < Y_2) - P(Y_1 > Y_2)$

using U-statistics and the delta method. These can be adapted to apply to θ because $\theta = [P(Y_1 < Y_2) - P(Y_1 > Y_2) + 1]/2$. Halperin, Hamdy, and Thall [9] provided a distribution-free confidence interval for θ using a pivotal quantity. Their simulation study showed that their approach is as good as or better than Hochberg's method, especially for extreme values of θ.

Recently, Newcombe [10] evaluated eight asymptotic confidence intervals for θ. The methods treat the distributions of Y_1 and Y_2 as continuous, but he stated that they also apply with ordinal categorical responses. He recommended a pseudo score-type confidence interval that assumes exponential distributions for Y_1 and Y_2. The following two sections consider other confidence interval methods for θ in the categorical-outcome case and compare them to existing methods.

3. CONFIDENCE INTERVALS FOR INDEPENDENT MULTINOMIAL DISTRIBUTIONS

Suppose there are n_1 *i.i.d* observations of Y_1 and n_2 *i.i.d* observations of Y_2, with results summarized by frequencies $\{n_{ij},\ i = 1, 2,\ j = 1, \cdots, c\}$ in a $2 \times c$ contingency table. The frequencies in each row have a multinomial distribution.

Except for a constant term, the product multinomial log-likelihood is

$$l(\boldsymbol{\pi}, \boldsymbol{\lambda}) = \boldsymbol{y}_1' \log(\boldsymbol{\pi}) + \boldsymbol{y}_2' \log(\boldsymbol{\lambda}), \tag{2}$$

where $\boldsymbol{y}_1 = (n_{11}, \cdots, n_{1c})'$, $\boldsymbol{y}_2 = (n_{21}, \cdots, n_{2c})'$, $\log(\boldsymbol{\pi}) = (\log(\pi_1), \cdots, \log(\pi_c))'$, and $\log(\boldsymbol{\lambda}) = (\log(\lambda_1), \cdots, \log(\lambda_c))'$. The maximum likelihood (ML) estimates for $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ are $\hat{\pi}_j = n_{1j}/n_1$, $\hat{\lambda}_j = n_{2j}/n_2$, $j = 1, \cdots, c$, and the ML estimate of θ is

$$\hat{\theta} = \hat{\boldsymbol{\lambda}}' A \hat{\boldsymbol{\pi}} = \frac{1}{n_1 n_2} \left[\sum_{i=1}^{c-1} \sum_{j>i}^{c} n_{1i} n_{2j} + 0.5 \sum_{i=1}^{c} n_{1i} n_{2i} \right].$$

This is the Mann-Whitney U-statistic, allowing ties, divided by the product of the sample sizes (Klotz [2]). From Halperin et al. [9], the variance of $\hat{\theta}$ is

$$V_{\hat{\theta}} = \frac{1}{n_1 n_2} \left[\theta - (n_1 + n_2 - 1)\theta^2 + (n_2 - 1)C + (n_1 - 1)D - \frac{1}{4} \sum_{i=1}^{c} \pi_i \lambda_i \right], \tag{3}$$

where $C = \sum_{i=1}^{c-1} \pi_i (\sum_{j=i+1}^{c} \lambda_j + \lambda_i/2)^2 + \pi_c \lambda_c^2/4$ and $D = \sum_{j=2}^{c} \lambda_j (\sum_{i=1}^{j-1} \pi_i + \pi_j/2)^2 + \pi_1^2 \lambda_1/4$. From properties of U-statistics, provided $0 < \theta < 1$,

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{V}_{\hat{\theta}}}} \xrightarrow{d} N(0, 1), \tag{4}$$

where $\hat{V}_{\hat{\theta}}$ is the estimated variance of $\hat{\theta}$ obtained by substituting the ML estimates of θ, $\boldsymbol{\pi}$, and $\boldsymbol{\lambda}$ into (3).

We now consider five asymptotic confidence intervals for θ: the Wald interval, the Wald interval applied to logit($\hat{\theta}$), the likelihood-ratio test (LRT)-based interval, the score-test based interval, and a pseudo score-type interval. We obtain them by inverting the corresponding tests of $H_0 : \theta = \theta_0$.

3.1. Wald-type Confidence Intervals

From the asymptotic normality of $\hat{\theta}$ in (4), the $100(1-\alpha)\%$ Wald confidence interval for θ is

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\hat{V}_{\hat{\theta}}},$$

where $z_{\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution. A degenerate confidence interval results if $\hat{\theta}$ equals either 0 or 1, or if $\hat{\theta} = 0.50$ with all observations falling in a single column. In near-extreme cases, the distribution of $\hat{\theta}$ is usually highly skewed, and the lower bound or the upper bound of this interval may fall outside of $[0, 1]$. Along with these problems, Wald intervals generally perform poorly for parameters based on proportions. For example, Brown, Cai and DasGupta [11] showed the Wald confidence interval for a binomial proportion has chaotic coverage probabilities even with large sample sizes.

A more promising Wald approach constructs the interval for a transformation of θ, such as $\text{logit}(\theta)$, and then inverts it to the θ scale. From the delta method, the Wald confidence interval for $\text{logit}(\theta)$ is

$$\text{logit}(\hat{\theta}) \pm z_{\alpha/2}\frac{\sqrt{\hat{V}_{\hat{\theta}}}}{\hat{\theta}(1-\hat{\theta})}.$$

Its bounds (LB, UB) induce the interval $(\exp(LB)/1+\exp(LB),\ \exp(UB)/1+\exp(UB))$ for θ. If $\hat{\theta}$ is either 0 or 1, we take the interval to be $[0, 1]$, which is unappealing compared to the intervals obtained with the following methods.

3.2. LRT-based Confidence Interval

To find the LRT confidence interval, we need the restricted ML estimates of the cell probabilities under $H_0 : \theta = \theta_0$ for all possible θ_0. We regard each such null hypothesis as a constraint function of the cell probabilities. That is, the cell probabilities under the null hypothesis satisfy $\boldsymbol{\lambda}' A\boldsymbol{\pi} - \theta_0 = 0$. We denote the restricted ML estimates satisfying $\theta = \theta_0$ by $\tilde{\boldsymbol{\pi}}(\theta_0)$ and $\tilde{\boldsymbol{\lambda}}(\theta_0)$.

The LRT statistic $G^2(\theta_0)$ for $H_0 : \theta = \theta_0$ is

$$G^2(\theta_0) = 2\left(\boldsymbol{y}_1'[\log(\hat{\boldsymbol{\pi}}) - \log(\tilde{\boldsymbol{\pi}}(\theta_0))] + \boldsymbol{y}_2'[\log(\hat{\boldsymbol{\lambda}}) - \log(\tilde{\boldsymbol{\lambda}}(\theta_0))]\right).$$

It has an asymptotic chi-square null distribution with $df = 1$ (Lang [14]). The $100(1-\alpha)\%$ LRT-based confidence interval for θ is the set of θ_0 that satisfies $G^2(\theta_0) < \chi^2_{(1-\alpha),1}$, where $\chi^2_{(1-\alpha),1}$ denotes the $(1-\alpha)$ quantile of the chi-square distribution with $df = 1$.

3.3. Score Confidence Interval

The score confidence interval for θ under multinomial sampling is obtained by inverting the score test statistic for $H_0 : \theta = \theta_0$. With the restricted ML estimates of cell probabilities under H_0, several authors including Aitchison and Silvey ([12], [13]), Bera and Bilias [15], and Lang [14] showed that for any multinomial model the score statistic is equivalent to the Pearson-type statistic,

$$S^2(\theta_0) = \sum_{j=1}^{c}\left[\frac{(y_{1j} - n_1\tilde{\pi}_j(\theta_0))^2}{n_1\tilde{\pi}_j(\theta_0)} + \frac{(y_{2j} - n_2\tilde{\lambda}_j(\theta_0))^2}{n_2\tilde{\lambda}_j(\theta_0)}\right].$$

This is also the Lagrange multiplier test statistic (Silvey [16]). It has an asymptotic chi-square null distribution with *df* $= 1$ (Lang [14]). Thus, the $100(1-\alpha)\%$ score confidence interval is the set of θ_0 satisfying $S^2(\theta_0) < \chi^2_{(1-\alpha),1}$. This approach has been used to obtain confidence intervals for many categorical measures, such as difference of proportions and Kappa for matched pairs (Agresti and Min [17], and Donner and Eliasziw [18]).

3.4. Pseudo Score-type Confidence Interval

An alternative to the Wald method directly uses the asymptotic normality of $\hat{\theta}$ with the null rather than non-null variance, following Wilson's ([19]) approach for proportions. To obtain the null variance of $\hat{\theta}$, we substitute the restricted ML estimates of cell probabilities under $H_0 : \theta = \theta_0$ in $V_{\hat{\theta}}$ in (3). Denote the corresponding estimated variance by $\tilde{V}_{\hat{\theta}}(\theta_0)$. Then, a $100(1-\alpha)\%$ pseudo score-type confidence interval for θ is the set of θ_0 that satisfies

$$PS^2(\theta_0) = \frac{(\hat{\theta}-\theta_0)^2}{\tilde{V}_{\hat{\theta}}(\theta_0)} < \chi^2_{(1-\alpha),1}.$$

Generally, this method differs from the score confidence interval method in Section 3.3.

3.5. Algorithms for Finding the Intervals

Finding the restricted ML estimates used in the LRT-based, score, and pseudo score confidence intervals entails finding ML estimates satisfying $\theta = \theta_0$ for various θ_0 constrained values. One can do this by finding saddlepoints of a Lagrange multiplier function expressed in terms of this constraint. Methods for doing this include a Newton-Raphson algorithm for constraint functions (Aitchison and Silvey [12], [13]) or a modified Newton-Raphson algorithm (Lang [14]) that avoids some difficulties related to the inversion of matrices in the Aitchison-Silvey algorithm.

Lang's approach is available with the "mph.fit" function in the *R* software, available from Dr. Joseph B. Lang ("joseph-lang@uiowa.edu"). We found it useful in applying this algorithm to use as initial values the endpoints of a more easily computable interval such as Newcombe's pseudo score interval. With such algorithms, there are occasional problematic cases caused by the algorithm not converging for certain θ_0 values. In future research, it would be useful to develop a special-purpose algorithm that works well for doing this. A function for *R* software to find these confidence intervals using Lang's function is available from the first author of this paper (Euijung Ryu).

4. CONFIDENCE INTERVALS FOR θ UNDER A PARAMETRIC MODEL

The confidence interval methods in the previous sections used $2(c-1)$ cell probability parameters. To reduce the number of parameters, substantially for large c, we can apply θ to a model for the table. A simple cumulative logit model for a $2 \times c$ table is

$$\text{logit}[P(Y_k \leq j)] = \alpha_j - (k-1)\beta, \quad j = 1, \cdots, c-1, \; k = 1, 2,$$

which has c parameters. One way that this model arises is by assuming that counts in the first row have an underlying latent variable with a standard logistic distribution (location parameter 0 and scale parameter 1), and counts in the second row have the same distribution except for a location shift β. Below, we make inference about θ without assuming any latent structure. An alternative approach, not explored here, makes inference about θ for assumed underlying parametric distributions.

Let $\gamma_{1j} = P(Y_1 \leq j)$ and $\gamma_{2j} = P(Y_2 \leq j)$ with $\boldsymbol{\gamma}_1 = (\gamma_{11}, \cdots, \gamma_{1(c-1)})'$ and $\boldsymbol{\gamma}_2 = (\gamma_{21}, \cdots, \gamma_{2(c-1)})'$. We can express cell probabilities using the cumulative probabilities. For example, $\pi_1 = \gamma_{11}$, $\pi_j = \gamma_{1j} - \gamma_{1(j-1)}$ for $j = 2, \cdots, c-1$, and $\pi_c = 1 - \gamma_{1(c-1)}$. Substituting the cumulative probabilities into (1), we get

$$\theta = \boldsymbol{\gamma}_1' D \boldsymbol{\gamma}_2 + 0.5(1 + \gamma_{1(c-1)} - \gamma_{2(c-1)}), \tag{5}$$

where

$$D = \begin{pmatrix} 0 & 0.5 & 0 & & \cdots & 0 \\ -0.5 & 0 & 0.5 & 0 & \cdots & 0 \\ & & \vdots & & & \\ & & \vdots & & & \\ 0 & \cdots & 0 & -0.5 & 0 & 0.5 \\ 0 & \cdots & & 0 & -0.5 & 0 \end{pmatrix}.$$

From the form of θ, the sign of $(\theta - 0.5)$ is the same as the sign of β.

Under the cumulative logit model, the log-likelihood function except for a constant term is

$$l(\boldsymbol{\alpha}, \beta) = \sum_{k=1}^{2} \sum_{j=1}^{c} n_{kj} \log \left[\frac{\exp(\alpha_j - (k-1)\beta)}{1 + \exp(\alpha_j - (k-1)\beta)} - \frac{\exp(\alpha_{j-1} - (k-1)\beta)}{1 + \exp(\alpha_{j-1} - (k-1)\beta)} \right], \tag{6}$$

where $\alpha_0 = 0$ and $\alpha_c = \infty$. The ML estimates of $\alpha_1, \cdots, \alpha_{c-1}$, and β can be obtained by standard statistical software such as SAS or R. Let $\hat{\theta}$ denote the ML estimate of θ for this model, that is, substituting the model-based ML estimates for $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ into (5).

The asymptotic variance of $\hat{\theta}$ is obtained using the delta method. Denoting $\boldsymbol{d} = (\partial\theta/\partial\alpha_1, \cdots, \partial\theta/\partial\alpha_{c-1}, \partial\theta/\partial\beta)'$, and denoting the expected Fisher information matrix by B, the asymptotic variance of $\hat{\theta}$ is $V_{\hat{\theta}} = \boldsymbol{d}' B^{-1} \boldsymbol{d}$.

The methods considered in the previous section can be applied to this measure using similar methods. For example, the $100(1-\alpha)\%$ logit Wald confidence interval is based on inverting $\text{logit}(\hat{\theta}) \pm z_{\alpha/2} \sqrt{\hat{V}_{\hat{\theta}}}/\hat{\theta}(1-\hat{\theta})$. Under $H_0 : \theta = \theta_0$, the restricted ML estimates $(\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_{c-1}\tilde{\beta})$ of the model parameters maximize the log-likelihood in (6). We obtain them using Lagrange multipliers with a constraint function

$$\boldsymbol{\gamma}_1' D \boldsymbol{\gamma}_2 + 0.5(1 + \gamma_{1(c-1)} - \gamma_{2(c-1)}) - \theta_0 = 0$$

by employing either the Newton-Raphson algorithm or Lang's algorithm. These are the basis of LRT-based and score confidence intervals. A pseudo score-type confidence interval substitutes the restricted ML estimates into $V_{\hat{\theta}}$ for each θ_0. Denoting the corresponding variance form by $\tilde{V}_{\hat{\theta}}(\theta_0)$, a $100(1-\alpha)\%$ pseudo score-type confidence interval for θ is the set of θ_0 satisfying

$$\frac{(\hat{\theta} - \theta_0)^2}{\tilde{V}_{\hat{\theta}}(\theta_0)} < \chi^2_{(1-\alpha),1}.$$

5. EXAMPLES

We now apply $\hat{\theta}$ and the confidence interval methods to Table I on shoulder tip pain scores. Assuming independent multinomial sampling, the unrestricted ML estimates are $\hat{\boldsymbol{\pi}} = (0.864,$ $0.091,\ 0.045,\ 0.000,\ 0.000)'$, $\hat{\boldsymbol{\lambda}} = (0.368, 0.158, 0.211, 0.158, 0.105)'$, and $\hat{\theta} = 0.771$. For the cumulative logit model, the Pearson test statistic for testing fit is 0.94 with $df = 3$, so the model fits adequately. Under this model, $\hat{\theta} = 0.773$. Table II shows the unrestricted and model-based confidence intervals for θ. By comparison, the Halperin et al. interval is (0.619, 0.875) and the Newcombe's pseudo score-type interval is (0.596, 0.880). Except for the ordinary Wald interval and the lower bound of the Newcombe interval, different methods give similar results.

Since all confidence intervals do not contain 0.5, we infer that the active treatment works better than the control treatment to reduce shoulder pain. The chance that a patient with the active treatment feels less pain than one receiving the control treatment could be only slightly greater than 0.5, or much greater. The imprecision reflects the relatively small sample sizes.

Table III illustrates how the methods perform with tables that can cause problems with simple Wald methods. When all observations fall in one column (case 1), the Wald intervals are degenerate (since the estimated standard error then equals 0). When $\hat{\theta}$ is near the boundary (case 2), the ordinary Wald interval can overshoot the boundary. When $\hat{\theta}$ is at the boundary (case 3), the Wald interval is degenerate and the logit Wald interval provides no information. For these problematic cases in Table III, the LRT-based intervals are shorter than the score and pseudo score-type intervals. We study next, with a simulation, how these methods and the Wald methods tend to perform.

6. SIMULATION STUDY

We used a simulation study to evaluate the confidence interval methods presented in Sections 3 and 4, the Halperin et al. interval, and the Newcombe's pseudo score-type interval. This study varied the number of columns c, the true θ value, whether the cumulative logit model holds, and group sample sizes.

We obtained the cell probabilities by categorizing two logistic distributions. For the first row, we used equal cell probabilities. The second row probabilities were obtained by possibly changing the location and scale parameters. Using the standard logistic distribution with location = 0 and scale = 1 for the first row, we determined $c - 1$ cutoff points of the $2 \times c$ table. For the second row we set the scale parameter to equal either 1 (cumulative logit model holds) or 2 (model does not hold). The location parameter for the second row was determined so that either $\theta = 0.5$ (no effect) or 0.8 (strong effect). We took the group sample sizes (n_1, n_2) = (10, 10), (50, 50), (100, 100), (10, 50), (50, 100) and (10, 100).

For each condition, we ran 10,000 simulations and estimated the actual coverage probability of nominal 95% confidence intervals. An estimate of the true coverage probability then has standard error that is about 0.002. Our evaluations found that the unrestricted score and pseudo score-type methods perform much better than the other methods, especially with small sample sizes. We found that the coverage probabilities for the unrestricted LRT confidence interval tend to be too low when sample sizes are small, especially when the sample sizes are highly unbalanced, regardless of whether or not the assumed model holds. We also found that Newcombe's pseudo score-type method tends to have coverage probabilities that are too high, but, recall that this method is designed for fully-ranked data. In addition, that method has the simplifying property, not shared by confidence intervals proposed in this paper, by which the $\hat{\theta}$ value and the sample sizes are sufficient to determine the interval; that is, the confidence limits for Newcombe's method depend only on $\hat{\theta}$, n_1, and n_2 (Newcombe [10]). The logit Wald confidence interval, which is simple to use, performs well with equal sample sizes, but sometimes has poor coverage probabilities with unequal sample sizes. All the parametric-based intervals often had poor coverage performance when the assumed model does not hold, so we cannot recommend them for general use.

Table IV summarizes the simulation results, averaged over all the sample size cases, over $c = 3$ and 6, over $\theta = 0.5$ and 0.8, and over whether the model holds. The table reports three overall summaries of performance: the mean coverage probability, the mean of the absolute differences between coverage probabilities and the nominal value, and the proportion of coverage probabilities with absolute distance more than 0.02 from the nominal value of 0.95. For the cases for which the absolute distance exceeded 0.02, coverage probabilities tended to be too large for the Newcombe pseudo score-type method and too small for the Halperin et al., Wald, logit Wald, and LRT intervals.

In developing a rank-based test about θ, mainly for the purpose of testing that $\theta = 0.5$ without assuming identical distributions, Brunner and Munzel [20] mentioned in passing that their test could be inverted to obtain a confidence interval for θ. However, this interval shares the disadvantage of the Wald interval of being centered at $\hat{\theta}$, and hence not working when $\hat{\theta}$ equals either 0 or 1. We also considered this method in our simulation. Although it performed better than the Wald method, it did not perform as well as the score and pseudo score-type methods.

In summary, for relatively small sample sizes the score and pseudo score-type methods seem best among the various methods, although this conclusion is tentative because of the limited cases this simulation study considered.

7. OTHER DATA STRUCTURES

Next we consider two other cases in which θ can provide a useful summary: (1) matched-pairs data, and (2) fully-ranked data, in which independent samples come from two continuous distributions, rather than multinomial distributions.

7.1. Matched-Pairs Data

Suppose each observation in one sample pairs with an observation in the other sample, for an ordered categorical response with c categories. The data are summarized by a $c \times c$ contingency table. Let $\boldsymbol{\pi} = \{\pi_{ij},\ i, j = 1, \ldots, c\}$ denote cell probabilities for this table. We assume the cell counts have a multinomial distribution with sample size n and cell probabilities $\boldsymbol{\pi}$.

Applying θ to the marginal row probabilities $(\pi_{1+}, \cdots, \pi_{c+})$ and marginal column probabilities $(\pi_{+1}, \cdots, \pi_{+c})$, the matched-pairs (MP) version of θ is

$$\theta_{MP} = (\pi_{+1}, \cdots, \pi_{+c})A(\pi_{1+}, \cdots, \pi_{c+})'.$$

Marginal homogeneity implies $\theta_{MP} = 0.5$, and θ_{MP} provides a comparison of the two marginal distributions that is sensitive to stochastic orderings.

The ML estimate of θ_{MP} is $\hat{\theta}_{MP} = (\hat{\pi}_{+1}, \cdots, \hat{\pi}_{+c})A(\hat{\pi}_{1+}, \cdots, \hat{\pi}_{c+})'$, which is obtained by substituting ML estimates of marginal row and column probabilities into θ_{MP}. It is straightforward to show that $\hat{\theta}_{MP}$ equals $\hat{\boldsymbol{\pi}}' R \hat{\boldsymbol{\pi}}$, where $\hat{\boldsymbol{\pi}}$ is a vector of sample proportions and

$$R = \left(\begin{array}{cccc|c|cccc} 0.5\mathbf{1}_c & \mathbf{1}_c & \cdots & \mathbf{1}_c & \cdots & 0.5\mathbf{1}_c & \mathbf{1}_c & \cdots & \mathbf{1}_c \\ \mathbf{0}_c & 0.5\mathbf{1}_c & \cdots & \mathbf{1}_c & \cdots & \mathbf{0}_c & 0.5\mathbf{1}_c & \cdots & \mathbf{1}_c \\ \vdots & \vdots & & \vdots & \cdots & \vdots & \vdots & & \vdots \\ \mathbf{0}_c & \mathbf{0}_c & \cdots & 0.5\mathbf{1}_c & \cdots & \mathbf{0}_c & \mathbf{0}_c & \cdots & 0.5\mathbf{1}_c \end{array} \right)'.$$

It is known that $\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \xrightarrow{d} N(\mathbf{0}, \Sigma)$, where $\Sigma = Diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$. Since $\hat{\theta}_{MP} = \hat{\boldsymbol{\pi}}'[0.5(R + R')]\hat{\boldsymbol{\pi}}$, where $0.5(R + R')$ is symmetric, $\hat{\theta}_{MP}$ has variance (Schott [21], p. 395)

$$var(\hat{\theta}_{MP}) = \frac{1}{n^2}\{0.5tr([(R + R')\Sigma]^2)\} + \frac{1}{n}\{\boldsymbol{\pi}'(R + R')\Sigma(R + R')\boldsymbol{\pi}\}. \quad (7)$$

The methods discussed in Sections 3 can be applied, using a multinomial log-likelihood and the variance form in (7) instead of (3).

In a limited simulation study, we used the cell probabilities from Section 6 as marginal row and column probabilities. For joint cell probabilities, we used an underlying bivariate normal distribution with correlations 0.4 and 0.8 satisfying the marginal row and column probabilities. For large samples, all methods performed reasonably well. With small sample sizes ($n = 25$, 50, and 75), however, we found that the logit Wald method performs better than the LRT method and better than the score and pseudo score-type methods when the effect is large, in which cases these other methods tend to be quite conservative. For example, see Table V with $n = 50$. This result was surprising, and the tentative superiority for small samples of the logit Wald method must be qualified by its inappropriateness for some cases (e.g., when $\hat{\theta} = 0$ or 1). Newcombe [22] has recently proposed a somewhat different approach for matched-pairs using a summary measure motivated by the Wilcoxon signed-rank statistic.

7.2. Fully-Ranked Data

Next, suppose independent samples of size n_1 and n_2, say $(X_{11}, \cdots, X_{n_1})$ and $(X_{21}, \cdots, X_{2n_2})$, come from unknown continuous distributions F_1 and F_2. To apply $\hat{\theta}$, we rank the data from the smallest to the largest and construct a $2 \times c$ table with $c = n_1 + n_2$, assuming there are no ties. We refer to this case as "fully-ranked data." By treating the data as if they come from two independent multinomial distributions, with frequencies n_{ij}, $i = 1, 2$, $j = 1, \cdots, c$ (each being a 0 or 1), we have $\hat{\pi}_j = n_{1j}/n_1$ and $\hat{\lambda}_j = n_{2j}/n_2$. The estimate of θ is

$$\hat{\theta} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I(X_{1i} < X_{2j})/(n_1 n_2),$$

where I is an indicator function. The numerator is the Mann-Whitney U statistic with no ties. An asymptotic normality of $\hat{\theta}$ in this case is well known, and so Wald confidence intervals apply directly. The estimated asymptotic variance of $\hat{\theta}$, say $v\hat{a}r(\hat{\theta})$, is obtained by substituting $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\lambda}}$ into the form in (2) with $\hat{\pi}_i\hat{\lambda}_i = 0$ for each i. The resulting confidence interval is the same as the Hanley-McNeil Wald method discussed by Newcombe [10].

We conducted simulation evaluations, generating data from normal distributions with identical variances but possibly different means. Based on Table VI, for large samples, the

Halperin et al. method, the Newcombe's pseudo score-type method, the unrestricted logit Wald method, the LRT method, and the score methods all seemed to perform well. For small samples, the score method and Newcombe's pseudo score-type method performed better than the other methods.

7.3. Connections with an Effect Size Measure for Normal Distributions

The previous section treated the response as having an unspecified continuous distribution. It is natural to inquire how much efficiency loss could occur from using the strictly ordinal approach when a particular parametric distribution truly holds.

For example, suppose the parametric model $X_1 \sim N(\mu_1, \sigma^2)$ and $X_2 \sim N(\mu_2, \sigma^2)$ truly holds. Then,

$$\theta = \Phi(\frac{\mu_2 - \mu_1}{\sqrt{2}\sigma}),$$

where Φ is the cdf of the standard normal distribution. So, in this case θ relates to the effect size measure $\Delta = (\mu_2 - \mu_1)/\sigma$ often used for approximately normal distributions with common variance. With equal sample sizes n, a natural parametric estimate of θ is

$$\hat{\theta}^* = \Phi(\frac{\bar{X}_2 - \bar{X}_1}{\sqrt{2}s}),$$

where $\bar{X}_1 = \sum_{i=1}^n X_{1i}/n$, $\bar{X}_2 = \sum_{i=1}^n X_{2i}/n$, and $s^2 = [\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^n (X_{2i} - \bar{X}_2)^2]/2(n-1)$. Based on Reiser and Guttman [23] and the delta method, the asymptotic variance of $\sqrt{n}\hat{\theta}'$ is

$$var(\sqrt{n}\hat{\theta}^*) = \phi^2(\frac{\Delta}{\sqrt{2}}) \left(1 + \frac{\Delta^2}{8}\right),$$

where $\phi(x)$ is the standard normal density. By contrast, for the ordinal (Mann-Whitney-type) estimate $\hat{\theta}$ of θ, under the normality assumptions the asymptotic variance is

$$var(\sqrt{n}\hat{\theta}) = 2\left[P(Z_1 \leq \frac{\Delta}{\sqrt{2}} \text{ and } Z_2 \leq \frac{\Delta}{\sqrt{2}}) - \theta^2\right],$$

where Z_1 and Z_2 are jointly bivariate normal with zero means and unit variances and correlation 0.5.

The asymptotic efficiency of the ordinal estimate relative to the parametric estimate of θ is the limit as n increases of

$$eff = \frac{var(\hat{\theta}^*)}{var(\hat{\theta})}.$$

For $\Delta = 0.0$, 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0, the asymptotic relative efficiencies are 0.955, 0.961, 0.974, 0.979, 0.957, 0.892, and 0.782. Therefore, under normality, the parametric estimate can be much better than the ordinal estimate when the effect is very large. Otherwise, the ordinal estimate holds up well, much as the corresponding Mann-Whitney test does in terms of the classic result about its local efficiency compared to the t test for normal distributions.

When a parametric model is plausible in the fully-ranked case and the effect is very large, it might be preferable to estimate θ using that model. In practice, even then this must be weighed against the possibility of actual coverage probabilities for corresponding confidence intervals possibly being far from nominal levels when there is model misspecification.

8. LOGIT MODELING OF θ WITH EXPLANATORY VARIABLES

The confidence interval methods for θ discussed in Sections 2 through 6 are designed for a single $2 \times c$ table. In practice, it can be useful to describe how θ depends on certain explanatory variables. We illustrate with Table VII, which shows the shoulder tip pain scores after laparoscopic surgery of Table I, now stratified by gender and age (Lumley [1]).

Let K denote the number of $2 \times c$ tables. This is the product of the number of levels of all covariates considered. For each k, $k = 1, \cdots, K$, let Y_{k1} and Y_{k2} denote the ordinal responses for the first and second rows of table k. Let $\boldsymbol{y}_{k1} = (n_{k11}, n_{k12}, \ldots, n_{k1c})'$ and $\boldsymbol{y}_{k2} = (n_{k21}, n_{k22}, \ldots, n_{k2c})'$ denote multinomial counts in those rows, for $n_{k1} = \sum_j^c n_{k1j}$ trials with cell probabilities $\boldsymbol{\pi}_{k1} = (\pi_{k11}, \pi_{k12}, \ldots, \pi_{k1c})'$ and $n_{k2} = \sum_j^c n_{k2j}$ trials with cell probabilities $\boldsymbol{\pi}_{k2} = (\pi_{k21}, \pi_{k22}, \ldots, \pi_{k2c})'$. For table k, let

$$\theta_k = P(Y_{k1} < Y_{k2}) + \frac{1}{2}P(Y_{k1} = Y_{k2}) = \boldsymbol{\pi}'_{k2} A \boldsymbol{\pi}_{k1},$$

and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$. Here, we consider modeling logit$(\boldsymbol{\theta})$ using covariates, where logit$(\boldsymbol{\theta})$ denotes $(\text{logit}(\theta_1), \cdots, \text{logit}(\theta_K))'$.

8.1. Logit Models for θ

Let X be a model matrix for the explanatory variables, and let $\boldsymbol{\beta}$ denote parameters for their effects. We consider the model with logit link,

$$\text{logit}(\boldsymbol{\theta}) = X\boldsymbol{\beta}. \tag{8}$$

Our interest here is to obtain ML estimates and confidence intervals for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

Under the assumption of independent multinomial distributions, the log-likelihood function except for a constant term is

$$l(\boldsymbol{\pi}) = \sum_{k=1}^{K} [\boldsymbol{y}'_{k1} \log(\boldsymbol{\pi}_{k1}) + \boldsymbol{y}'_{k2} \log(\boldsymbol{\pi}_{k2})],$$

where $\boldsymbol{\pi} = (\boldsymbol{\pi}'_{11}, \boldsymbol{\pi}'_{12}, \cdots, \boldsymbol{\pi}'_{K1}, \boldsymbol{\pi}'_{K2})'$. The cell probabilities cannot be expressed in terms of the model parameters, which complicates ML estimation. The ML estimates of $\boldsymbol{\pi}$ are values that maximize the log-likelihood function under the model constraint (8). One can use Lang's algorithm [14] to obtain the ML estimates of $\hat{\boldsymbol{\pi}}$ and hence of

$$\hat{\theta}_k = \hat{\boldsymbol{\pi}}'_{k2} A \hat{\boldsymbol{\pi}}_{k1}, \quad \text{and} \quad \hat{\boldsymbol{\beta}} = (X'X)^{-1} X' \ \text{logit}(\hat{\boldsymbol{\theta}}).$$

For confidence intervals, here we consider only the score-test-based approach, which was found to perform well in Sections 6 and 7. To obtain the model-based interval for θ_k, we need to find restricted ML estimates of the cell probabilities, say $\tilde{\boldsymbol{\pi}}(k)$, under the model (8) with the constraint $\theta_k - \theta_0 = 0$. This can be done again using Lang's algorithm. Then the Pearson-type statistic for testing $H_0 : \theta_k = \theta_0$ assuming (8) compares the two sets of fitted values using

$$S_k^2(\theta_0) = \sum_{i=1}^{K} \sum_{j=1}^{2} \sum_{r=1}^{c} \frac{(n_{ij}\hat{\pi}_{ijr} - n_{ij}\tilde{\pi}_{ijr}(k))^2}{n_{ij}\tilde{\pi}_{ijr}(k)},$$

with $df = 1$ (Lang [24]). The $100(1-\alpha)\%$ score confidence interval for θ_k is a set of θ_0 satisfying $S_k^2(\theta_0) < \chi^2_{(1-\alpha),1}$. The score confidence interval for β_j results from a similar argument.

8.2. Modeling θ for Shoulder-Tip Pain

Now we return to Table VII to describe the treatment effect on shoulder tip pain. We use a main-effects model

$$\text{logit}(\theta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where $x_1 = 0$ for a female patient, $x_1 = 1$ for a male, $x_2 = 0$ if a patient's age is between 20 and 70, and $x_2 = 1$ for age above 70. The Pearson statistic for testing the model fit is 0.36 with $df = 1$. The model seems to fit adequately, but this test provides only a rough indication because of the data sparseness.

Table VIII shows ML estimates and their 95% score confidence intervals. The confidence intervals for the effects of gender and age include 0, but are very wide because of the small sample size (with only six observations at the higher age level). Because of the data sparseness, the standard errors for ML estimates based on the sample proportions (treating each 2×5 table separately) are unreliable, and the estimate and standard error for the females of age 71+ are degenerate. Table VIII also shows score confidence intervals for the effect size measure. The first two confidence intervals indicate that the active treatment is significantly better than the control treatment for patients whose ages are between 20 and 70.

9. CONCLUSIONS

The simulation study in Section 6 was limited in scope but suggested that the pseudo score and score confidence intervals under an unrestricted model perform best among the methods discussed for independent multinomial samples. The score method also seems to perform well for fully-ranked data. For matched-pairs data, the logit Wald method seems to perform better than the score method for small n. Newcombe's pseudo score-type method for fully-ranked data performs well for such data, but its coverage probabilities tend to be too high when it is applied to ordered categorical data. Section 8 discussed how to estimate parameters in logit models for the measure, using the score method when all covariates are categorical. A function for *R* software to find the confidence intervals in the unrestricted case for a $2 \times c$ table is available from E. Ryu.

REFERENCES

1. Lumley T. Generalized estimating equations for ordinal data: A note on working correlation structures. *Biometrics* 1996; **52**: 354–361.
2. Klotz JH. The Wilcoxon, ties, and the computer. *Journal of the American Statistical Association* 1966; **61**: 772–787.
3. Vargha A, Delaney HD. The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics* 1998; **59**: 137–142.
4. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic Graph. *Journal of Mathematical Psychology* 1975; **12**: 387–145.
5. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947; **18**: 50–60.
6. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bulletin* 1945; *1*: 80–83.
7. Troendle JF. A likelihood ratio test for the nonparametric Beherens-Fisher problem. *Biometrical Journal* 2002; *44*: 813–824.
8. Hochberg Y. On the variance estimate of a Wilcoxon-Mann Whitney statistic for group ordered data. *Communications in Statistics - Theory and Methods* 1981; **A10**: 1719–1732.
9. Halperin M, Hamdy MI, Thall PF. Distribution-free confidence intervals for a parameter of Wilcoxon-Mann-Whitney type for ordered categories and progressive censoring. *Biometrics* 1989: **45**: 509–521.
10. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: Asymptotic methods and evaluation. *Statistics in Medicine* 2006; **25**: 559–573.
11. Brown LD, Cai T, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001; **16**: 101–133.
12. Aitchison J, Silvey SD. Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics* 1958; **29**: 813–828.
13. Aitchison J, Silvey SD. Maximum-likelihood estimation procedures and associated tests of significance. *Journal of the Royal Statistical Society, Ser. B.* 1960; **1**: 154–171.
14. Lang JB. Multinomial-Poisson homogeneous models for contingency tables. *The Annals of Statistics* 2004; **32**: 340–383.
15. Bera AK, Bilias Y. Rao's score, Neyman's $C(\alpha)$ and Silvey's LM tests: an essay on historical developments and some new results. *Journal of Statistical Planning and Inference* 2001; **97**: 9–44.
16. Silvey SD. The Lagrangian multiplier test. *Annals of Mathematical Statistics* 1959; **30**: 389–407.
17. Agresti A, Min Y. Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine* 2005; **24**: 729–740.
18. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the Kappa statistis: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* 1991; **11**: 1511–1519.
19. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927; **22**: 209–212.
20. Brunner E, Munzel, U. The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal* 2000; **42**: 17-25.
21. Schott JR. *Matrix analysis for statistics.* Wiley: New York, 1997.
22. Newcombe RG. A relative measure of effect size for paired data generalizing the wilcoxon matched-pairs signed-ranks test statistic. *Unpublished manuscript* 2007.
23. Reiser B, Guttman I. Statistical inference for $Pr(Y < X)$: The normal case. *Technometrics* 1986; **28**: 253–257.
24. Lang JB. Homogeneous linear predictor models for contingency tables. *Journal of the American Statistical Association* 2005; **100**: 121–134.

Table I. Shoulder tip pain scores after laparoscopic surgery

	Pain scores				
Treatments	1	2	3	4	5
Active	19	2	1	0	0
Control	7	3	4	3	2

Table II. Confidence intervals for θ in Table I, with and without assuming a cumulative logit model

Assume model		Wald	Logit Wald	LRT	Score	Pseudo score
No	Lower endpoint	0.644	0.621	0.635	0.633	0.628
	Upper endpoint	0.900	0.874	0.882	0.875	0.874
Yes	Lower endpoint	0.645	0.621	0.632	0.629	0.627
	Upper endpoint	0.901	0.876	0.885	0.876	0.876

Table III. Confidence intervals for θ with extreme cases

Counts	$\hat{\theta}$		Wald	Logit Wald	LRT	Score	Pseudo score
First row: $(10, 0, 0, 0, 0)$	0.500	Lower endpoint	0.500	0.500	0.412	0.361	0.361
Second row: $(20, 0, 0, 0, 0)$		Upper endpoint	0.500	0.500	0.546	0.581	0.581
First row: $(4, 5, 1, 0, 0)$	0.975	Lower endpoint	0.926	0.840	0.810	0.718	0.715
Second row: $(0, 0, 10, 8, 2)$		Upper endpoint	1.024	0.997	0.999	0.996	0.996
First row: $(4, 6, 0, 0, 0)$	1.000	Lower endpoint	1.000	0.000	0.834	0.736	0.734
Second row: $(0, 0, 10, 8, 2)$		Upper endpoint	1.000	1.000	1.000	1.000	1.000

Table IV. Overall performance summaries of coverage probability (CP) from simulation study for seven methods, averaged over several sample sizes, $c = 3$ and 6, $\theta = 0.5$ and 0.8, and whether or not a cumulative logit model holds

Methods	Mean of CP	Mean of $\|CP - 0.95\|$	Proportion of $(\|CP - 0.95\| > 0.02)$
Halperin et al.	0.943	0.010	0.083
Newcombe's pseudo score ‡	0.962	0.013	0.250
Wald	0.922	0.028	0.500
logit Wald	0.943	0.010	0.125
LRT	0.942	0.008	0.125
Score	0.952	0.003	0.000
Pseudo score	0.951	0.003	0.000

Table V. Coverage probabilities of five methods from simulation study for matched-pairs data, with sample sizes = 50 and c =6

	$\theta = 0.5$		$\theta = 0.65$		$\theta = 0.8$		$\theta = 0.95$	
	$\rho = 0.4$	$\rho = 0.8$	$\rho = 0.4$	$\rho = 0.8$	$\rho = 0.4$	$\rho = 0.8$	$\rho = 0.4$	$\rho = 0.8$
Wald	0.942	0.952	0.945	0.951	0.934	0.941	0.922	0.917
logit Wald	0.947	0.954	0.951	0.954	0.945	0.950	0.941	0.945
LRT	0.945	0.953	0.949	0.958	0.947	0.956	0.963	0.970
Score	0.948	0.976	0.954	0.973	0.969	0.975	0.967	0.973
Pseudo score	0.951	0.976	0.956	0.976	0.968	0.977	0.967	0.976

Table VI. Coverage probabilities of seven methods from simulation study for fully-ranked data, with sample sizes (10, 10) and (20, 30)

	$(n_1, n_2) = (10, 10)$		$(n_1, n_2) = (20, 30)$	
	$\theta = 0.5$	$\theta = 0.8$	$\theta = 0.5$	$\theta = 0.8$
Halperin et al.	0.946	0.922	0.946	0.945
Newcombe's pseudo score	0.948	0.952	0.948	0.960
Wald	0.920	0.886	0.938	0.925
logit Wald	0.966	0.973	0.953	0.955
LRT	0.947	0.916	0.946	0.946
Score	0.950	0.945	0.949	0.956
Pseudo score	0.963	0.943	0.952	0.972

Table VII. Shoulder tip pain scores after laparoscopic surgery, stratified by age and gender

Age	Gender	Treatment	1	2	3	4	5
20-70	Female	Active	12	1	0	0	0
		Control	3	2	3	0	2
	Male	Active	5	1	1	0	0
		Control	1	0	1	3	0
71+	Female	Active	1	0	0	0	0
		Control	1	0	0	0	0
	Male	Active	2	1	0	0	0
		Control	1	0	0	0	0

Table VIII. ML estimates of θ_k and β_j parameters, with their 95% score confidence intervals

		ML estimates		Score intervals	
		Model-based (s.e.)	Sample prop.-based (s.e.)	Lower endpoints	Upper endpoints
$\boldsymbol{\theta}$	$x_1 = 0$, $x_2 = 0$	0.850 (0.068)	0.831 (0.079)	0.685	0.942
	$x_1 = 1$, $x_2 = 0$	0.810 (0.104)	0.857 (0.117)	0.522	0.944
	$x_1 = 0$, $x_2 = 1$	0.500 (0.011)	0.500 (0.000)	0.372	0.877
	$x_1 = 1$, $x_2 = 1$	0.429 (0.171)	0.333 (0.136)	0.349	0.822
$\boldsymbol{\beta}$	intercept (β_0)	1.735 (0.537)	-	0.780	2.123
	gender (β_1)	-0.283 (0.696)	-	-1.938	1.322
	age (β_2)	-1.736 (0.538)	-	-3.444	0.314

[‡] Method designed for fully-ranked data but applied here for categorical data