Profile Confidence Intervals for Contingency Table Parameters

Joseph B. Lang

July 2005

Technical Report No. 351

Profile Confidence Intervals for Contingency Table Parameters

Joseph B. Lang

Department of Statistics and Actuarial Science
University of Iowa, Iowa City, IA 52245 USA, joseph-lang@uiowa.edu

July 20, 2005

A general investigation of profile score and profile likelihood confidence intervals for contingency table parameters is carried out. The investigation method, which is based on the theory of multinomial-Poisson homogeneous models, lends itself to a general computational algorithm and is applicable for a broad class of parameters and sampling schemes. The method also affords useful theoretical results on estimability and sampling plan invariance. The computational algorithm described herein is straightforward to implement and avoids two main limitations of existing algorithms. The theoretical results make clear how inferences depend on the sampling plan and the characteristics of the estimand function. Among other things, these results lead to practical ways to avoid "estimating" non-estimable parameters, a practice that is all too common with current contingency table software. Examples of profile confidence intervals for a variant of the gamma measure of association, global odds ratios, a difference between marginal means, and a marginal dispersion measure illustrate the method.

KEY WORDS: Categorical data; Contingency tables; Estimability; Multinomial-Poisson homogeneous model; Profile likelihood confidence interval; Profile score confidence interval; Wald confidence interval.

1 Introduction

The contingency table literature suggests that profile score and profile likelihood confidence intervals generally have better coverage properties than their Wald counterpart–see for example, Mee (1984), Miettinen and Nurminen (1985), Bedrick (1987), Gart and Nam (1988a), Meeker and Escobar (1995), Newcombe (1998a,b), Agresti and Coull (1998), Agresti and Caffo (2000), Brown et al. (2001), Agresti (2002:77), Tsimikas et al. (2002). The conclusions in this existing literature serve as the motivation for the current paper.

We carry out a general investigation of profile score and profile likelihood confidence intervals for contingency table parameters. The investigation method, which is based on the theory of multinomial-Poisson homogeneous (MPH) models (Lang 2004), lends itself to a general com-

putational algorithm and is applicable for a broad class of parameters and sampling plans. Rather than adding to existing literature by again comparing coverage properties of different intervals, the current paper has two distinct primary objectives. The first is to describe a general MPH-based algorithm for computing a profile score or likelihood interval for scalar-valued estimand $S(\tau)$, where τ is a vector of contingency table probabilities. The second objective is to give theoretical results on the applicability of the MPH method, parameter estimability, and sampling plan invariance.

Profile intervals have been used on a case-by-case basis for several different contingency table estimands, including a single binomial probability, odds ratio, relative risk, and risk difference. The examples in the literature, several of which are summarized in Sections 2 and 3, use computational approaches that have two main limitations: they are case-specific and they are applicable for a restrictive class of parameters. The computational approach proposed in this paper avoids these limitations. This paper's theoretical results make clear how inferences depend on the sampling plan and the characteristics of the estimand function. Among other things, these results lead to practical ways to avoid "estimating" non-estimable parameters, a practice that is all too common with current contingency table software.

The paper is organized as follows: Section 2 gives several examples from the literature where profile score or likelihood confidence intervals have been considered. Section 3 gives qualitatively different examples that have not been considered previously in the literature. Section 4 gives a very brief overview of the theory underlying profile confidence intervals. Section 5 describes the wide variety of sampling plans under which the method outlined in this paper is applicable. Section 6 describes a new approach, and gives an algorithm, for computing profile intervals that is based on the theory of multinomial-Poisson homogeneous (MPH) models. Section 7 discusses the technical conditions under which the MPH method is applicable. Section 8 gives some useful theoretical results on estimability and sampling plan invariance. Section 9 revisits the examples of Section 3 and computes and compares the profile and Wald intervals. Section 10 gives a brief discussion.

2 Examples of Likelihood-Explicit Estimands

Practically speaking, there are two main types of estimands: (i) Likelihood-explicit estimands are those that afford a simple explicit reparameterization of the data likelihood in terms of

the estimand and nuisance parameters. (ii) Likelihood-implicit estimands are those that do not afford such a simple explicit reparameterization. The majority of the examples in the contingency table literature restrict attention to likelihood-explicit estimands. This section gives several, mostly context-free, examples.

Example 2.1 (Difference between Marginal Probabilities). Define the 2×2 table probabilities as $\tau_{ij} \equiv P(A = i, B = j), i, j = 1, 2$. Several researchers have discussed profile score or likelihood intervals, or closely-related variants, for estimand $S(\boldsymbol{\tau}) = \tau_{1+} - \tau_{+1}$ (Quesenberry and Hurst 1964; Lloyd 1990; May and Johnson 1997; Tango 1998, Agresti 2002:411). That this estimand is likelihood-explicit follows because $S(\boldsymbol{\tau}) = \Delta$ if and only if $\boldsymbol{\tau} = (\eta_1, \eta_2 + \Delta, \eta_2, 1 - \eta_1 - 2\eta_2 - \Delta)$, where $\boldsymbol{\eta} \equiv (\tau_{11}, \tau_{21})$ is a nuisance parameter that lies in a space that contains a 2-dimensional rectangle. ■

Example 2.2 (Relative Difference). Define the 2×2 table probabilities as $\tau_{ij} \equiv P(B = j | A = i)$. Gart and Nam (1988b) discuss profile score intervals for estimand $S(\boldsymbol{\tau}) = (\tau_{21} - \tau_{11})/\tau_{11}$. This is a likelihood-explicit estimand because $S(\boldsymbol{\tau}) = \Delta$ if and only if $\boldsymbol{\tau} = (\eta, 1-\eta, (1+\Delta)\eta, 1 - (1 + \Delta)\eta)$, where $\eta = \tau_{11}$. ■

Example 2.3 (Risk Rate Difference for Incomplete Table). Tang and Tang (2003) consider data on two-step tuberculosis skin testing. Define the 2×2 table probabilities as $\tau_{ij} \equiv P(\text{TB1} = i, \text{TB2} = j)$, where TB$k = 1$ or 2 as tuberculosis test k is negative or positive. If a patient tests positive the first time, it is assumed they would test positive the second time, so they are not tested again. This leads to the structural zero $\tau_{21} = 0$. Tang and Tang (2003) show how to compute the profile score interval for the estimand $S(\boldsymbol{\tau}) = \tau_{1+} - \tau_{11}/\tau_{1+}$; here, $\boldsymbol{\tau} = (\tau_{11}, \tau_{12}, \tau_{22})$. Notice that $S(\boldsymbol{\tau}) = \Delta$ if and only if $\boldsymbol{\tau} = (\eta^2 - \eta\Delta, \eta(1 + \Delta) - \eta^2, 1 - \eta)$, where $\eta = \tau_{1+}$. ■

Example 2.4 (Mean of $\{0, 1, 2\}$ Variable). Define the 1×3 table probabilities as $\tau_i \equiv P(R = i), i = 0, 1, 2$. Newcombe (2003) considers profile intervals for estimand $S(\boldsymbol{\tau}) = 0\tau_0 + 1\tau_1 + 2\tau_2 = E(R)$. Notice that $S(\boldsymbol{\tau}) = \Delta$ if and only if $\boldsymbol{\tau} = (1 - \Delta + \eta, \Delta - 2\eta, \eta)$, where $\eta = \tau_2$. ■

Other examples in the literature consider estimands such as a single binomial probability (e.g. Agresti and Coull 1998, Newcombe 1998a, Brown et al. 2001), the relative risk (e.g. Bedrick 1987), and the difference between two probabilities when estimates are from independent samples (e.g. Newcombe 1998b, Agresti and Caffo 2000). These examples, along with Examples 2.1 through 2.4, all share the important characteristic that there exists an explicitly invertible one-to-one function of the form $\boldsymbol{\tau} \mapsto (S(\boldsymbol{\tau}), \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a vector of nuisance pa-

rameters. This means that the data log likelihood can be explicitly reparameterized in terms of the estimand of interest $S(\tau)$ and nuisance parameters. This reparameterization is exploited in the computation of the profile confidence intervals in all of these examples.

3 Examples of Likelihood-Implicit Estimands

As a rule, examples in the profile confidence interval literature use likelihood-explicit estimands. Rare exceptions to this rule include the estimands in Tsimikas et al. (2002) and Lang (2005). Tsimikas et al. considered estimands of the form $S(\tau) = \tau_1^T \mathbf{G} \tau_2$ that arise in a receiver operating characteristic (ROC) curve setting. Here, $\mathbf{G}$ is an upper-triangular matrix of constants, the table probabilities in τ_1 and τ_2 give the distributions of ordinal rating variables, and estimand $S(\tau)$ is an area under a non-parametric ROC curve. This estimand does not afford a simple explicit reparameterization of the data likelihood; that is, the estimand is likelihood-implicit. Thus, the computational methods used for the likelihood-explicit examples of the previous section are not applicable. Tsimikas et al. (2002) address this problem and describe an algorithm that uses feasible sequential quadratic programming optimization for computing a profile likelihood confidence interval for the estimand. To illustrate the utility of maximum likelihood estimation for homogeneous linear predictor models, Lang (2005) computed profile likelihood intervals for two different likelihood-implicit estimands, a model-based conditional probability and a difference between marginal conditional probabilities.

This section gives several more examples that use likelihood-implicit estimands.

3.1 Measures of Association for Case-Control Data.

Example 3.1. Agresti (2002) describes a case-control study of the relationship between smoking and myocardial infarction. Let A be the disease-status variable (1=control, 2=case) and let B be the level-of-smoking variable (1="no cigarettes", 2="1-24 cigarettes per day", 3="over 24 cigarettes per day"). Define the 2×3 table probabilities $\tau_{ij} = P(B = j | A = i), i = 1, 2; j = 1, 2, 3$. In tabular form,

		B			
		1 "no cigs"	2 "1-24 cigs"	3 "over 24 cigs"	
A	1 "control"	τ_{11}	τ_{12}	τ_{13}	1.0
	2 "case"	τ_{21}	τ_{22}	τ_{23}	1.0

A Modified Gamma Measure of Association. Let $B_i \sim B|\{A = i\}, i = 1, 2$, be inde-

pendent random variables. The estimand $\gamma^* \equiv P(B_2 > B_1 | B_1 \neq B_2)$ is a reasonable measure of association between A and B. In words, if a case and a control subject are independently sampled, given that their smoking levels are different, the chances that the case patient smokes more than the control patient is γ^*. The estimand γ^* can be written as

$$\gamma^* = S(\tau) \equiv \frac{\tau_{11}(\tau_{22} + \tau_{23}) + \tau_{12}\tau_{23}}{1 - \sum_{i=1}^{3} \tau_{2i}\tau_{1i}}.$$

This estimand is a relatively complicated function of the table probabilities in τ. It can be considered a likelihood-implicit estimate because it is not obvious how to find an explicitly invertible one-to-one function of the form $\tau \mapsto (S(\tau), \eta)$, where η is a vector of nuisance parameters. Moreover, this estimand is not of the quadratic form considered in Tsimikas et al. (2002). Thus, the available methods for computing profile intervals are not directly applicable.

Global Odds Ratios. The estimand γ^* is a measure of association that exploits the ordinality of smoking level B. The global odds ratios are alternative measures of association that exploit ordinality. Let

$$\Omega_1 \equiv \frac{odds(A = 2 | B \geq 2)}{odds(A = 2 | B < 2)} \text{ and } \Omega_2 \equiv \frac{odds(A = 2 | B \geq 3)}{odds(A = 2 | B < 3)}.$$

That is, for example, the odds of myocardial infarction ($A = 2$) for those smoking at least one cigarette a day ($B \geq 2$) is Ω_1 times that of the odds of myocardial infarction for those not smoking at all ($B < 2$). We can re-express these odds ratios as functions of the table probabilities. Specifically, we can "invert" the conditional probabilities to obtain

$$\Omega_1 = \frac{odds(B \geq 2 | A = 2)}{odds(B \geq 2 | A = 1)} = \frac{\tau_{11}(\tau_{22} + \tau_{23})}{\tau_{21}(\tau_{12} + \tau_{13})} \equiv S_1(\tau)$$

$$\Omega_2 = \frac{odds(B \geq 3 | A = 2)}{odds(B \geq 3 | A = 1)} = \frac{(\tau_{11} + \tau_{12})\tau_{23}}{(\tau_{21} + \tau_{22})\tau_{13}} \equiv S_2(\tau).$$

Like γ^*, these global odds-ratio estimands are likelihood-implicit.

Agresti (p 98, 2002) gives data for a random sample of $n_1 = 62$ controls and $n_2 = 4$ cases. The counts in tabular form are

		B			
		1 "no cigs"	2 "1-24 cigs"	3 "over 24 cigs"	
A	1 "control"	25	25	12	62
	2 "case"	0	1	3	4

These counts are viewed as realizations of two independent multinomial vectors, $Y_1 \sim mult(62, \tau_{11}, \tau_{12}, \tau_{13})$ and $Y_2 \sim mult(4, \tau_{21}, \tau_{22}, \tau_{23})$. In Section 9.1, these data are used to compute Wald, profile score, and profile likelihood confidence intervals for each of the three estimands $\gamma^* = S(\tau)$, $\Omega_1 = S_1(\tau)$, and $\Omega_2 = S_2(\tau)$.

3.2 Marginal Mean and Dispersion Estimands

Example 3.2. A recent issue of a popular golf magazine gave expert ratings of several different golf clubs. Let R_1 and R_2 be a randomly sampled expert's ratings (1=low, 2, 3, 4, 5=high) for two particular clubs. Define the 5×5 table probabilities as $\tau_{ij} \equiv P(R_1 = i, R_2 = j), i, j = 1, 2, 3, 4, 5$. An estimand of primary interest is the difference between the mean ratings, $E(R_1) - E(R_2) = S(\boldsymbol{\tau}) \equiv \sum_{i=1}^{5} i\tau_{i+} - \sum_{j=1}^{5} j\tau_{+j}$. Other estimands of interest include the mean ratings, $E(R_1) = S_1(\boldsymbol{\tau}) \equiv \sum_{i=1}^{5} i\tau_{i+}$ and $E(R_2) = S_2(\boldsymbol{\tau}) \equiv \sum_{j=1}^{5} j\tau_{+j}$, and the dispersions of ratings, $D(R_1) = S_3(\boldsymbol{\tau}) \equiv 1 - \sum_{i=1}^{5} \tau_{i+}^2$, and $D(R_2) = S_4(\boldsymbol{\tau}) \equiv 1 - \sum_{j=1}^{5} \tau_{+j}^2$. It is reasonable to classify each of these estimands as likelihood-implicit.

For a random sample of 25 experts, the ratings can be summarized as counts in tabular form

		R_2					
		1	2	3	4	5	
R_1	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
	3	1	0	0	0	0	1
	4	0	0	1	1	0	2
	5	0	0	0	2	20	22
		1	0	1	3	20	25

These counts are viewed as realizations of $\mathbf{Y} \sim \ mult(25, \tau_{11}, \tau_{12}, \ldots, \tau_{55})$. In Section 9.2, these data are used to compute Wald, profile score, and profile likelihood confidence intervals for each of the estimands, $S(\boldsymbol{\tau})$, $S_1(\boldsymbol{\tau}), \ldots, S_4(\boldsymbol{\tau})$.

4 Test-Inversion Confidence Regions

Suppose that data vector $\mathbf{y}$ is a realization of random vector $\mathbf{Y}$, which has a distribution that depends on parameters in $\boldsymbol{\theta}$. A reasonable approach for computing a confidence region for estimand $S(\boldsymbol{\theta})$ is based on inverting tests of $H_\Delta : S(\boldsymbol{\theta}) = \Delta$ versus $K_\Delta : S(\boldsymbol{\theta}) \neq \Delta$. Specifically, if, for each candidate Δ, test statistic $D(\mathbf{Y}, \Delta)$ has a null limiting distribution that is free of $\boldsymbol{\theta}$ and $P(D(\mathbf{Y}, \Delta) > cv | H_\Delta) \approx \alpha$, then $C(\mathbf{y}) \equiv \{\Delta : D(\mathbf{y}, \Delta) \leq cv\}$ is an approximate level $1 - \alpha$ confidence region for $S(\boldsymbol{\theta})$. Confidence region $C(\mathbf{y})$ for $S(\boldsymbol{\theta})$ comprises all those values Δ for which the hypothesis $H_\Delta : S(\boldsymbol{\theta}) = \Delta$ would *not* be rejected. Casella and Berger (Sec 9.2, 1990) gives a nice introduction to test-inversion confidence regions.

For the contingency table settings considered in this paper, the model parameter $\boldsymbol{\theta}$ is composed of $(\boldsymbol{\sigma}, \boldsymbol{\tau})$, where $\boldsymbol{\sigma}$ is a collection of expected sample sizes and $\boldsymbol{\tau}$ is a collection of table

probabilities. The estimands we consider have the form $S(\tau)$; they do not involve σ.

There are many candidate test statistics D for testing $H_\Delta : S(\tau) = \Delta$ versus $K_\Delta : S(\tau) \neq \Delta$ in the contingency table setting. Three of the more common candidates are the Wald, Pearson's X^2, and the G^2 (Bishop et al. 1975) statistics. Other examples include the power divergence statistics of Read and Cressie (1988).

Wald statistics have the form $W_g^2(\mathbf{Y}, \Delta) = (g(S(\hat{\tau})) - g(\Delta))^2 / \hat{\sigma}^2_{g \circ S}$, where g is some differentiable one-to-one function, $\hat{\sigma}^2_{g \circ S}$ is an approximate variance of $g(S(\hat{\tau}))$, and $\hat{\tau}$ is the unrestricted MLE; i.e. $\hat{\tau}$ is the vector of sample table proportions. Note that there is no single Wald statistic because different choices of g lead to different Wald statistics.

The statistics X^2 and G^2 can be be expressed in terms of (σ, τ) or in terms of the expected table counts $\mu \equiv E(\mathbf{Y})$. This follows because there is an explicit one-to-one correspondence between μ and (σ, τ), as shown in Section 5.3 below. Here we choose the simpler μ parameterization. Pearson's X^2 statistic has the form $X^2(\mathbf{Y}, \hat{\mu}(\Delta))$ and the G^2 statistic has the form $G^2(\mathbf{Y}, \hat{\mu}(\Delta))$, where $\hat{\mu}(\Delta)$ is the MLE under $H_\Delta : S(\tau) = \Delta$,

$$X^2(\mathbf{Y}, \mu) = (\mathbf{Y} - \mu)^T diag^{-1}(\mu)(\mathbf{Y} - \mu), \quad \text{and} \quad G^2(\mathbf{Y}, \mu) = 2\mathbf{Y}^T \log(\mathbf{Y}/\mu).$$

For testing H_Δ vs. unrestricted K_Δ, it can be shown that Pearson's X^2 is the score statistic and G^2 is the likelihood ratio statistic for all of the sampling models considered in this paper.

Section 8.1 below argues that when the estimand function $S(\cdot)$ satisfies mild conditions, and when H_Δ holds, the Wald, score, and likelihood ratio statistics W_g^2, X^2, and G^2 have approximate central $\chi^2(1)$ distributions when the expected sample sizes are large. These results hold for the wide variety of sampling plans considered in this paper. It follows that

$$\begin{aligned} WCI_g(\mathrm{y}) &\equiv \{\, \Delta \,:\, W_g^2(\mathrm{y}, \Delta) \,\leq\, \chi^2_\alpha(1)\} \\ PSCI(\mathrm{y}) &\equiv \{\, \Delta \,:\, X^2(\mathrm{y}, \hat{\mu}(\Delta)) \,\leq\, \chi^2_\alpha(1)\} \\ PLCI(\mathrm{y}) &\equiv \{\, \Delta \,:\, G^2(\mathrm{y}, \hat{\mu}(\Delta)) \,\leq\, \chi^2_\alpha(1)\} \end{aligned} \tag{1}$$

are approximate $1 - \alpha$ level confidence intervals for $S(\tau)$. Here, $\chi^2_\alpha(1)$ is the $1 - \alpha$ quantile of the $\chi^2(1)$ distribution.

The confidence intervals $PSCI(\mathrm{y})$ and $PLCI(\mathrm{y})$ in (1) are referred to as the profile score confidence interval and the profile likelihood confidence interval, respectively. The adjective *profile* is used for the following reason. Let $\mathcal{T}$ be the set of all possible table probability vectors τ. Both test statistics have the form $D(\mathrm{y}, \Delta) = D^*(\mathrm{y}, \hat{\mu}(\Delta))$, where $D^*(\mathrm{y}, \mu_0(\tau_0))$

is the test statistic for testing the simple null $H_0 : \tau = \tau_0$ vs. $H_1 : \tau \neq \tau_0$. The one-dimensional curve $\{(\hat{\tau}(\Delta), D^*(\mathrm{y}, \hat{\mu}(\Delta)) : \Delta \in S(\mathcal{T})\}$ lies on the $dim(\mathcal{T})$-dimensional surface $\{(\tau, D^*(\mathrm{y}, \mu_0(\tau))) : \tau \in \mathcal{T}\}$. Thus, the set $\{(\Delta, D^*(\mathrm{y}, \hat{\mu}(\Delta)) : \Delta \in S(\mathcal{T})\}$, which determines the confidence region for $S(\tau)$, gives a profile of test statistic D^* with respect to Δ. In contrast, the test-inversion confidence interval WCI based on a Wald statistic is not generally a profile confidence interval.

Remark 1. In the likelihood-ratio case, the $D^*(\mathrm{y}, \hat{\mu}(\Delta))$ profile curve is an affine transformation of the *profile log likelihood* $\ell_p(\Delta|\mathrm{y})$ curve (cf. McCullagh and Nelder 1989:254-5). Specifically $\ell_p(\Delta|\mathrm{y}) = \ell(\hat{\mu}|\mathrm{y}) - \frac{1}{2}D^*(\mathrm{y}, \hat{\mu}(\Delta))$, where $\ell(\cdot|\mathrm{y})$ is the log likelihood and $\hat{\mu}$ is the unrestricted ML estimate of μ.

Remark 2. So as to avoid confusion with terminology used in the profile likelihood literature, the reader should note that in this paper the profile score confidence interval is based on the profile of the score statistic $X^2(\mathrm{y}, \hat{\mu}(\Delta))$ as described above. It is *not* based on the profile score function (cf. DiCiccio et al. 1996), which has the form $s_p(\Delta|\mathrm{y}) = \partial\ell_p(\Delta|\mathrm{y})/\partial\Delta$.

Remark 3. Confidence regions for scalar estimands obtained by inverting tests are most often, but not always, intervals. For this reason, we simply refer to the regions as intervals. (See Casella and Berger, Sec 9.2, 1990).

In contrast to profile score and profile likelihood intervals, Wald intervals can always be analytically inverted. Specifically, it is easy to see that

$$WCI_g(\mathrm{y}) = \{\ \Delta\ : g(S(\hat{\tau})) - z^*\hat{\sigma}_{g\circ S} \leq g(\Delta) \leq g(S(\hat{\tau})) + z^*\hat{\sigma}_{g\circ S}\} = g^{-1}[g(S(\hat{\tau})) \pm z^*\hat{\sigma}_{g\circ S}], \tag{2}$$

where z^* is the $1 - \alpha/2$ quantile of the standard normal distribution. Moreover, because the Wald statistics do not use the ML estimate of $S(\tau)$ under H_Δ, the computation is simple for both likelihood-explicit and likelihood-implicit estimands. This computational advantage explains the popularity of the Wald intervals, which are the default for most statistical packages. However, this advantage is less of an issue with today's computing power and generally does not outweigh the disadvantages such as poor coverage properties, lack of range-preserving property, and dependence on choice of parameterization.

In contrast to Wald intervals, and owing to the invariance property of MLE's, profile score and likelihood intervals are invariant to the choice of parameterization. With Wald intervals, one must be wary of the choice of parameterization, because many do not work very well. For

example, a poor choice of g in W_g^2 can give a Wald interval that includes values that fall outside the range of candidate $S(\tau)$ values. Also, a poor choice can lead to a Wald interval with coverage probability that is very far from the nominal probability. Moreover, it has been argued (Meeker and Escobar, Section 2.3, 1995) that profile score and likelihood intervals generally work as well as, or better than, the Wald interval based on the best choice of parameterization .

5 Multinomial-Poisson Sampling Plans

Confidence intervals obviously depend on the sampling scheme and corresponding sampling distribution of the data. Contingency table counts can be the result of many different sampling schemes. As examples, stratified random sampling with fixed sample sizes leads to product multinomial counts; and simple random sampling with random sample size that follows a Poisson distribution leads to independent Poisson counts. This section gives a formal description of a broad class of sampling plans and corresponding sampling distributions.

5.1 Sampling Plans and Sufficient Table Counts

Let composite categorical variable C have support on c distinct values; without loss of generality, we label these values $\{1, 2, \ldots, c\}$. Let π denote the joint distribution of C. That is, $C \sim \pi$ means that $P(C = i) = \pi_i,\ i = 1, \ldots, c$. As an example, C could represent two categorical variables, viz. $C \equiv (A, B)$, and have support on the 4 distinct values $\{(1,1), (1,2), (2,1), (2,2)\}$. Without loss of generality, the 4 distinct values could be re-labeled as $\{1, 2, 3, 4\}$. As is the case in this example, we will follow the convention of always listing support point labels in lexicographical order.

Sampling Plan. Let $\{C \in \phi_k\}, k = 1, \ldots, K$ correspond to $K \geq 1$ disjoint strata. Specifically, $\phi_k \cap \phi_{k'} = \emptyset$ and $\cup_{k=1}^{K} \phi_k = \{1, 2, \ldots, c\}$. A stratified random sample is taken from these strata. The sample size for stratum k, say N_k, is either *a priori* fixed, $N_k = n_k$, or is a Poisson random variable, $N_k \sim Poisson(\delta_k)$. It is assumed that the sample sizes are independent of each other and of the responses.

The model parameters corresponding to this sampling plan include the expected sample sizes $\sigma_k \equiv E(N_k)$ and the conditional probabilities that describe the distributions of $C|\{C \in \phi_k\}$, for $k = 1, \ldots, K$. More specifically, let $k(i)$ indicate the stratum in which C value i resides; so

that i falls in $\phi_{k(i)}$. Define the *table probabilities* as

$$\tau_i \equiv P(C = i | C \in \phi_{k(i)}), \quad i = 1, \ldots, c.$$

It follows that the model parameters include the expected sample sizes σ and the table probabilities τ.

Sufficient Table Counts. Let random variable Y_i be the number of sampled units with C value equal to i. For any of the sampling plans considered herein, the *table counts* in $\mathbf{Y} \equiv (Y_1, \ldots, Y_c)$ are sufficient for (σ, τ). By the sufficiency principle (e.g. Casella and Berger, p. 247, 1990), we base our inference on the sampling distribution and observed value of $\mathbf{Y}$.

5.2 A Formalization: The Sampling Plan Triple.

Every sampling plan described in the previous subsection can be uniquely identified by a *sampling plan triple* $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$, which comprises a *population matrix*, a *sampling constraint matrix*, and a *vector of fixed sample sizes.*

The $c \times K$ *population matrix* $\mathbf{Z}$ indicates stratum membership in that the k^{th} column of $\mathbf{Z}$ indicates which values of C are included in strata $\{C \in \phi_k\}$. In symbols, the $(i,k)^{th}$ element of $\mathbf{Z}$ is $(\mathbf{Z})_{ik} = I(i \in \phi_k)$.

Let $\mathbf{Q}(\{a_1, \ldots, a_g\}) \equiv [\mathbf{e}_{a_1}, \ldots, \mathbf{e}_{a_g}]_{K \times g}$, where $1 \leq a_1 < a_2 < \cdots < a_g \leq K$ and $\mathbf{e}_a$ is the elementary vector with a '1' in the a^{th} position and '0's everywhere else. Define $\mathbf{Q}(\emptyset) = 0$. Suppose that sample sizes in strata $k_1 < \cdots < k_f$ are fixed and sample sizes in strata $k_1^* < \cdots < k_r^*$ are random; here, $f + r = K$. Let $\mathbf{n} = (n_{k_1}, \ldots, n_{k_f})$ be the corresponding *vector of fixed sample sizes.* Define

$$\mathbf{Q}_F \equiv \mathbf{Q}(\{k_1, \ldots, k_f\}) \text{ and } \mathbf{Q}_R \equiv \mathbf{Q}(\{k_1^*, \ldots, k_r^*\})$$

The *sampling constraint matrix* $\mathbf{Z}_F$ is defined as $\mathbf{Z}_F = \mathbf{Z}\mathbf{Q}_F$. That is, $\mathbf{Z}_F$ comprises those columns in $\mathbf{Z}$ that correspond to strata with *fixed* sample sizes. By convention, when no sample sizes are fixed, $\mathbf{Z}_F$ and $\mathbf{n}$ are set equal to 0. It is also useful to label as $\mathbf{Z}_R$ the complementary collection of columns in $\mathbf{Z}$ that correspond to strata with *random* sample sizes. Specifically, let $\mathbf{Z}_R = \mathbf{Z}\mathbf{Q}_R$. By convention, when no sample sizes are random, $\mathbf{Z}_R$ is set equal to 0.

The components in the sampling plan triple $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ are useful for theoretical, computational, and notational reasons (see Lang 2004). Some examples include: (i) The table probabilities can be expressed as a simple function of the joint probabilities, viz. $\tau = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\pi)\pi \equiv$

$\mathbf{t}(\pi)$. (ii) The structural constraints on the table probabilities can be expressed as $\mathbf{Z}^T\tau = \mathbf{1}_K$. (iii) The expected counts have the form $E(\mathbf{Y}) = \mathbf{D}(\mathbf{Z}\sigma)\tau$. (iv) The variance of the counts is $var(\mathbf{Y}) = \mathbf{D}(\mathbf{Z}\sigma)[\mathbf{D}(\tau) - \mathbf{D}(\tau)\mathbf{Z}_F\mathbf{Z}_F^T\mathbf{D}(\tau)]$. (v) The vector of sample sizes (fixed and random) is $\mathbf{Z}^T\mathbf{Y} = (N_1, \ldots, N_K)^T$. (vi) The vector of fixed samples sizes is $\mathbf{Z}_F^T\mathbf{Y} = \mathbf{n}$.

5.3 Sampling Distribution of Table Counts

Expected Sample Size, Table Probability Parameterization. For sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$, the vector of counts $\mathbf{Y}$ comprises independent blocks of multinomial and/or Poisson components. Specifically, counts that correspond to a stratum with fixed sample size follow a multinomial distribution; counts that correspond to a stratum with Poisson sample size are independent with Poisson distributions; and stratum-specific blocks of counts are independent. The probability density of $\mathbf{Y}$ can be written as

$$P(\mathbf{Y} = \mathbf{y}) = c^*(\mathbf{y}) \exp\{\mathbf{y}^T \log \tau + \mathbf{y}^T \mathbf{Z}_R \log(\mathbf{Q}_R^T\sigma) - \mathbf{1}^T\mathbf{Q}_R^T\sigma\} I(\mathbf{y} \in \mathcal{Y}), \tag{3}$$

where $c^*(\mathbf{y}) \equiv \mathbf{n}!/\mathbf{y}!$ if $\mathbf{Z}_F \neq 0$, and $c^*(\mathbf{y}) \equiv 1/\mathbf{y}!$ if $\mathbf{Z}_F = 0$. Here, $(x_1, \ldots, x_m)! \equiv x_1!x_2! \cdots x_m!$ The support of $\mathbf{Y}$ is $\mathcal{Y} = \{\mathbf{y} : \mathbf{Z}_F^T\mathbf{y} = \mathbf{n},\ y_i = 0, 1, \ldots\}$. We use $\mathbf{Y} \sim MP(\sigma, \tau | \mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ to denote a random vector with this probability density function. Candidate values of (σ, τ) lie in $\mathcal{S} \times \mathcal{T}$, where

$$\mathcal{S} \equiv \{\sigma : \sigma = \mathbf{Q}_F\mathbf{n} + \mathbf{Q}_R\delta,\ \delta > 0\} \quad \text{and} \quad \mathcal{T} \equiv \{\tau :\ \mathbf{Z}^T\tau = 1,\ \tau > 0\}. \tag{4}$$

Mean Parameterization. If $\mathbf{Y} \sim MP(\sigma, \tau | \mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ then the mean is

$$\mu \equiv E(\mathbf{Y}) = \mathbf{m}(\sigma, \tau) \equiv \mathbf{D}(\mathbf{Z}\sigma)\tau.$$

The *mean function* $\mathbf{m}$ has several useful properties, including (i) $\mathbf{m} : \mathcal{S} \times \mathcal{T} \to \mathbf{m}(\mathcal{S} \times \mathcal{T})$ is one-to-one; (ii) $\mathbf{m}^{-1} : \mathbf{m}(\mathcal{S} \times \mathcal{T}) \to \mathcal{S} \times \mathcal{T}$ is defined as $\mathbf{m}^{-1}(\mu) = \left(\mathbf{Z}^T\mu,\ \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mu)\mu\right)$; and (iii) $\mathbf{m}(\mathcal{S} \times \mathcal{T}) = \{\mu :\ \mathbf{Z}_F^T\mu = \mathbf{n},\ \mu > 0\}$.

Because the mean vector is a one-to-one function of the expected sample sizes and table probabilities, we can re-parameterize the table count model in terms of μ. Specifically, $P(\mathbf{Y} = \mathbf{y}) = c(\mathbf{y}) \exp\{\mathbf{y}^T \log \mu - \mathbf{1}^T\mathbf{Z}_R^T\mu\} I(\mathbf{y} \in \mathcal{Y})$, where $c(\mathbf{y}) = \mathbf{n}! \exp\{-\mathbf{n}^T \log \mathbf{n}\}/\mathbf{y}!$ if $\mathbf{Z}_F \neq 0$ and $c(\mathbf{y}) = 1/\mathbf{y}!$ if $\mathbf{Z}_F = 0$. The support of $\mathbf{Y}$ is $\mathcal{Y} = \{\mathbf{y} : \mathbf{Z}_F^T\mathbf{y} = \mathbf{n},\ y_i = 0, 1, \ldots\}$. We use $\mathbf{Y} \sim MP(\mu | \mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ to denote a random vector with this probability density function. Candidate values of μ lie in $\mathbf{m}(\mathcal{S} \times \mathcal{T}) = \{\mu :\ \mathbf{Z}_F^T\mu = \mathbf{n},\ \mu > 0\}$.

6 Computing Profile Confidence Intervals

For a likelihood-explicit estimand $S(\tau)$ like those in Section 2 there exists an explicitly invertible function of the form $\tau \mapsto (S(\tau), \eta)$. More formally, there exists an explicit f such that $S(\tau) = \Delta$ iff $\exists \eta \in \mathcal{N} \ni \tau = \mathrm{f}(\Delta, \eta)$, where η is a $(dim(\mathcal{T}) - 1) \times 1$ vector of nuisance parameters and $\mathcal{N}$ is a set that contains a $dim(\mathcal{T}) - 1$ dimensional rectangle. Here, $\mathcal{T}$ is the set defined in (4).

The existence of f implies that the score and likelihood ratio statistics can be easily evaluated. In particular, the log likelihood corresponding to density (3) can be reparameterized and the following identity can be exploited:

$$\max_{\sigma,\ S(\tau)=\Delta} \ell(\sigma, \tau) = \max_{\sigma,\eta} \ell(\sigma, \mathrm{f}(\Delta, \eta)).$$

The examples in the literature typically use standard unrestricted gradient methods (e.g. Newton-Raphson) to find the ML estimates under H_Δ.

Remark 4. Actually, the examples in the literature considered sampling plans that do not include any unknown expected sample sizes. That is, there were no σ parameters to estimate. The current paper allows for unknown σ to accommodate Poisson counts.

There are two main limitations of the existing computational approach for likelihood-explicit estimands. i) The approach is case specific: For each distinct estimand $S(\tau)$, the explicit f must be determined. This rules out general fitting algorithms. ii) The approach does not work for likelihood-implicit estimands like γ^* and $E(R_1) - E(R_2)$ of Examples 3.1 and 3.2. It would not work for estimands like the correlation coefficient and Goodman and Kruskal's gamma.

With regard to likelihood-implicit estimands, a limitation of the algorithm of Tsimikas et al. (2002) is that it is case-specific; it is specifically designed for estimands of the quadratic form $S(\tau) = \tau_1^T \mathrm{G} \tau_2$ along with product-multinomial sampling. Similarly, the algorithm implicitly used in Lang (2005) was designed for the specific estimands in that paper.

The approach advocated in this paper avoids the existing limitations. It is related to the constrained maximization approach of Tsimikas et al. (2002) and is a natural extension of the algorithm used in Lang (2005). The idea is simple: Profile confidence intervals involve the inversion of score or likelihood ratio tests of $H_\Delta : S(\tau) = \Delta$. Under mild conditions on estimand function S, as described in the next section, the H_Δ model with constraints $h_\Delta(\tau) \equiv S(\tau) - \Delta = 0$ can be expressed as a *multinomial-Poisson homogeneous* (MPH) model

(Lang 2004). In this case, established MPH modeling results and fitting algorithms can be used to compute unrestricted and restricted (under H_Δ) ML estimates of $(\boldsymbol{\sigma}, \boldsymbol{\tau})$, as well as the Wald, score, and likelihood-ratio goodness-of-fit statistics. The MPH method also leads to simple descriptions of the asymptotic distributions of these statistics. It follows that when the MPH method is applicable, the inversion of the score and likelihood ratio tests of $H_\Delta : S(\boldsymbol{\tau}) = \Delta$, or equivalently the computation of the profile confidence intervals, is straightforward.

There are several advantages to using the MPH model approach, including: (i) It is applicable for both likelihood-explicit and likelihood-implicit estimands; (ii) There is no need to specify an explicit f for reparameterization; (iii) The score and likelihood ratio statistics are easily evaluated by solving the Lagrangian restricted likelihood equations; (iv) There is no need to analytically evaluate derivatives of $S(\boldsymbol{\tau})$ as they can be numerically approximated; and (v) There is no need to work on a case-by-case basis.

The practical advantage of the MPH model approach is that it lends itself to general algorithms for computing profile confidence intervals for a broad class of estimands and a broad class of sampling plans. As an example, consider the following...

General Algorithm for Computing a Profile Score Confidence Interval.

Step 1. Compute a Wald interval for $S(\boldsymbol{\tau})$, as in (2). Let $\Delta_{W,low}$ and $\Delta_{W,up}$ be the endpoints.

Step 2. Compute the lower profile confidence bound.

Step 2a. Using $\Delta_{W,low}$ as a reference point, find Δ_1 and Δ_2 satisfying $X^2(\mathbf{y}, \widehat{\boldsymbol{\mu}}(\Delta_1)) > \chi^2_\alpha(1) > X^2(\mathbf{y}, \widehat{\boldsymbol{\mu}}(\Delta_2))$. Here, $\widehat{\boldsymbol{\mu}}(\Delta)$ is the ML estimate of $\boldsymbol{\mu}$ under H_Δ, which can be found using the MPH model fitting algorithm described in Lang (2004).

Step 2b. With Δ_1 and Δ_2 as starting estimates, use a bisection algorithm to solve for Δ_{low} in $X^2(\mathbf{y}, \widehat{\boldsymbol{\mu}}(\Delta_{low})) = \chi^2_\alpha(1)$. Each bisection iteration involves fitting an MPH model.

Step 3. Using the same approach as in Step 2, determine an upper confidence bound Δ_{up} that satisfies $X^2(\mathbf{y}, \widehat{\boldsymbol{\mu}}(\Delta_{up})) = \chi^2_\alpha(1)$.

In Step 2a, it often works to choose Δ_1 and Δ_2 close to, but on either side of, $\Delta_{W,low}$. Similarly, in Step 3, it often works to choose Δ_1 and Δ_2 close to, but on either side, of $\Delta_{W,up}$. A

judicious choice of parameterization for the Wald interval can lead to better starting estimates of the profile interval bounds.

Essentially the same algorithm works for profile likelihood confidence intervals; one need only replace X^2 by G^2 in Steps 2 and 3. The algorithm has been coded in R (Ihaka and Gentleman 1996; see also `http://cran.r-project.org`) and is available from the author upon request.

7 Applicability of the Interval Estimation Approach

The MPH-based computational approach outlined in the previous section is applicable provided the models under H_Δ, viz.

$$\mathbf{Y} \sim \ MP(\boldsymbol{\mu}|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n}), \quad h_\Delta(\boldsymbol{\tau}) \equiv S(\boldsymbol{\tau}) - \Delta = 0$$

are multinomial-Poisson homogeneous (MPH) models. Lang (2004) defines an MPH model as $\mathbf{Y} \sim \ MP(\boldsymbol{\mu}|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n}), \quad \mathbf{h}(\boldsymbol{\mu}) = \mathbf{0}$, where the constraint function $\mathbf{h}$ is $\mathbf{Z}$ homogeneous (i.e. homogeneous with respect to the sampling plan) and sufficiently smooth and the constraints $\mathbf{h}(\boldsymbol{\mu}) = \mathbf{0}$ are non-redundant. More formally, $\mathbf{h}$ must satisfy the four conditions H_0–H_3 of Lang (2004). The next two subsections give conditions under which the constraint function h_Δ satisfies these MPH requirements.

7.1 Homogeneous Estimands and Constraints

Lang (2004) gives the following definition.

Definition 1: *The function* $\mathbf{h}$ *defined on* $\{\mathbf{x} \in R^c : \mathbf{x} > 0\}$ *is* $\mathbf{Z}$ *homogeneous, or, equivalently, homogeneous with respect to the sampling plan* $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$, *if*

$$\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) = \mathbf{G}(\boldsymbol{\gamma})\mathbf{h}(\mathbf{x}), \ \forall \boldsymbol{\gamma} > 0, \ \forall \mathbf{x} > 0,$$

where $\mathbf{G}(\boldsymbol{\gamma})$ *is a diagonal matrix with* j^{th} *diagonal element equal to* $\gamma_{k_j}^{p_j}$, *for some* $k_j \in \{1, \ldots, K\}$ *and some real number* p_j. *In the special case when* $p_j \equiv 0$ *so* $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) = \mathbf{h}(\mathbf{x})$, *we say that* $\mathbf{h}$ *is 0-order* $\mathbf{Z}$ *homogeneous.*

Both papers, Lang (2004) and Lang (2005), make use of several useful properties of homogeneous functions. Below, we give three different properties that are more directly useful for the purposes of the current paper. Proofs are given in the Appendix.

Property 7.1: *The function* $\mathbf{t}$ *defined as* $\mathbf{t}(\mathbf{x}) \equiv \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{x})\mathbf{x}$ *is a 0-order* $\mathbf{Z}$ *homogeneous mapping of* $\{\mathbf{x} : \mathbf{x} > 0\}$ *onto* $\mathcal{T} = \{\boldsymbol{\tau} : \boldsymbol{\tau} > 0, \mathbf{Z}^T\boldsymbol{\tau} = 1\}$.

Property 7.2: *To argue that* $\mathbf{h}$ *is* $\mathbf{Z}$ *homogeneous, we need only show the defining condition holds for all* $\gamma > 0$ *and for* $\mathbf{x} \in \mathcal{T} = \{\mathbf{x} : \mathbf{Z}^T\mathbf{x} = \mathbf{1}, \mathbf{x} > \mathbf{0}\}$.

Property 7.3: *Let* f *be any function defined on* $\mathcal{T}$. *Define* $\mathrm{f}_0(\mathbf{x}) \equiv \mathrm{f}(\mathbf{t}(\mathbf{x}))$, *where* $\mathbf{t}(\mathbf{x}) = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{x})\mathbf{x}$. *Then* (i) f_0 *is 0-order* $\mathbf{Z}$ *homogeneous and* (ii) $\mathrm{f}_0(\mathbf{x}) = \mathrm{f}(\mathbf{x})$, $\forall \mathbf{x} \in \mathcal{T}$. *We say* f_0 *is a 0-order* $\mathbf{Z}$ *homogeneous version of* f [*over* $\mathcal{T}$].

Property 7.3 implies that every estimand $S(\boldsymbol{\tau})$ has a 0-order $\mathbf{Z}$ homogeneous version, namely, $S_0(\boldsymbol{\tau}) \equiv S(\mathbf{t}(\boldsymbol{\tau}))$. Therefore, the model under H_Δ

$$\mathbf{Y} \sim MP(\boldsymbol{\mu}|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n}), \quad h_\Delta(\boldsymbol{\tau}) \equiv S(\boldsymbol{\tau}) - \Delta = 0$$

is equivalent to

$$\mathbf{Y} \sim MP(\boldsymbol{\mu}|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n}), \quad h_{\Delta,0}(\boldsymbol{\mu}) \equiv S_0(\boldsymbol{\mu}) - \Delta = 0, \tag{5}$$

where $h_{\Delta,0}$ is 0-order $\mathbf{Z}$ homogeneous. That is, the homogeneity condition $\mathtt{H}_3$ in Lang (2004) is satisfied for every possible estimand $S(\boldsymbol{\tau})$. Here, we used the fact that $\boldsymbol{\mu} = \mathbf{D}(\mathbf{Z}\boldsymbol{\sigma})\boldsymbol{\tau}$, so by 0-order $\mathbf{Z}$ homogeneity, $S_0(\boldsymbol{\tau}) = S_0(\boldsymbol{\mu})$ and hence $h_{\Delta,0}(\boldsymbol{\tau}) = h_{\Delta,0}(\boldsymbol{\mu})$.

7.2 Smooth, Non-Redundant Estimand and Constraint Functions

The previous subsection showed that the homogeneity condition $\mathtt{H}_3$ in Lang (2004) is satisfied for every possible estimand $S(\boldsymbol{\tau})$. The compatibility condition $\mathtt{H}_0$ of Lang (2004) is also always satisfied because Δ is in $S_0(\mathcal{T})$. Therefore, to argue that the model under H_Δ is an MPH model, one need only show that the constraint function $h_{\Delta,0}$ defined in (5) satisfies the remaining two conditions in Lang (2004)–the smoothness condition $\mathtt{H}_1$ and the full-rank condition $\mathtt{H}_2$.

Exploiting the simple form of the constraint function $h_{\Delta,0}$, it is easy to see that it will satisfy the two conditions $\mathtt{H}_1$ and $\mathtt{H}_2$ in Lang (2004) if 0-order $\mathbf{Z}$ homogeneous S_0 satisfies the following two conditions:

$$\begin{aligned} &\text{(S1)} \quad S_0 \text{ has continuous second-order derivatives at any } \mathbf{x} > \mathbf{0}; \text{ and} \\ &\text{(S2)} \quad \frac{\partial S_0(\mathbf{x})}{\partial \mathbf{x}^T} \neq \mathbf{0}, \ \mathbf{x} > \mathbf{0}. \end{aligned} \tag{6}$$

Condition (S2) implies that the set of constraints $\{h_{\Delta,0}(\boldsymbol{\mu}) = 0, \mathbf{Z}^T\boldsymbol{\tau} = \mathbf{1}\}$ are non-redundant (see Proposition 7 in Lang 2004). It also implies that the estimator of $S_0(\boldsymbol{\tau})$ admits a non-degenerate linear approximation.

It is important that the full-rank condition (S2) holds for a 0-order $\mathbf{Z}$ homogeneous version of the estimand function, not just any version. As an example, consider a two-celled contingency

table with probabilities τ_1 and τ_2 that satisfy $\mathbf{Z}^T\boldsymbol{\tau} \equiv \tau_1 + \tau_2 = 1$; that is, $\mathbf{Z} = 1$. The estimand function defined as $S_1(\boldsymbol{\tau}) \equiv \tau_1^2 + \tau_2^2 + 2\tau_1\tau_2$ satisfies condition (S2), but the 0-order $\mathbf{Z}$ homogeneous version $S_{10}(\boldsymbol{\tau}) = 1$ does not. For this example, the model constraint $S_1(\boldsymbol{\tau}) = \Delta$ imposes no additional constraint beyond the structural constraint $\tau_1 + \tau_2 = 1$. As another example, the estimand function defined as $S_2(\boldsymbol{\tau}) \equiv \tau_1\tau_2$ satisfies (S2), but the 0-order $\mathbf{Z}$ homogeneous version $S_{20}(\boldsymbol{\tau}) = \tau_1\tau_2/(\tau_1+\tau_2)^2$ does not; it has vanishing derivative at $(1/2, 1/2)$. For this example, $S_2(\boldsymbol{\tau}) = \Delta$ does impose an additional constraint, but the linear approximation for the estimator is degenerate when $\boldsymbol{\tau} = (1/2, 1/2)$. In this case, higher-order approximations are required and the MPH model results of Lang (2004) are not directly applicable.

Conditions (S1) and (S2) along with the implicit function theorem imply that $S(\boldsymbol{\tau}) = \Delta$ can be locally expressed as $\boldsymbol{\tau} = \mathbf{f}(\Delta, \boldsymbol{\eta})$ for some function f. For asymptotic purposes, conditions (S1) and (S2) really only need to be satisfied in a neighborhood of the true $\boldsymbol{\tau}$. Assuming smoothness and the full rank condition over the larger regions ensures that ML fitting algorithms based on iterative gradient methods are applicable. As an example, for the estimand $S(\boldsymbol{\tau}) = \tau_1\tau_2$ of the previous paragraph, the theory is asymptotically applicable provided $\boldsymbol{\tau} \neq (1/2, 1/2)$. Even when $\boldsymbol{\tau} \neq (1/2, 1/2)$, however, the ML fitting algorithms would only be applicable if the profile intervals do not include the problematic $S_2(\boldsymbol{\tau})$ value 0.25, which corresponds with $\boldsymbol{\tau} = (1/2, 1/2)$.

8 Useful Theoretical Results based on MPH Modeling

The results of the previous section are summarized in Theorem 8.1. This theorem gives conditions under which the MPH results of Lang (2004) apply. The remainder of the section gives other relevant corollaries and theorems.

Theorem 8.1. *Suppose that counts* $\mathbf{y}$ *are realizations of* $\mathbf{Y} \sim MP(\boldsymbol{\mu}|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$. *For every estimand function* S, *there exists a 0-order* $\mathbf{Z}$ *homogeneous version* S_0. *If this 0-order version* S_0 *satisfies conditions* (S1) *and* (S2), *then, for every candidate* Δ, *the model with constraints* $h_\Delta(\boldsymbol{\tau}) \equiv S(\boldsymbol{\tau}) - \Delta = 0$ *can be re-expressed as a multinomial-Poisson homogeneous model with 0-order constraints* $h_{\Delta,0}(\boldsymbol{\mu}) \equiv S_0(\boldsymbol{\mu}) - \Delta = 0$.

In practice, to automatically re-express the model with constraint function h_Δ in terms of a 0-order $\mathbf{Z}$ homogeneous version, one can simply use $h_{\Delta,0}(\boldsymbol{\mu}) \equiv h_\Delta(\mathbf{t}(\boldsymbol{\mu}))$, where $\mathbf{t}(\boldsymbol{\mu}) = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\boldsymbol{\mu})\boldsymbol{\mu}$. That is, the estimand $S(\boldsymbol{\tau})$ can be re-expressed as $S(\mathbf{t}(\boldsymbol{\mu}))$, where $\boldsymbol{\mu} = \mathbf{D}(\mathbf{Z}\boldsymbol{\sigma})\boldsymbol{\tau}$

is the vector of expected counts. That $S(\mathbf{t}(\mu))$ equals $S(\tau)$ follows because $\mathbf{t}(\mu) = \mathbf{t}(\tau)$ by Property 7.1 and $\mathbf{t}(\tau) = \tau$ for $\tau \in \mathcal{T}$.

8.1 Approximate Chi-Square Result

Corollary 8.1. *Under the conditions of Theorem 8.1 and for every candidate value of* Δ*, the goodness-of-fit statistics* $W_g^2(\mathbf{Y}, \Delta)$*,* $X^2(\mathbf{Y}, \widehat{\mu}(\Delta))$*, and* $G^2(\mathbf{Y}, \widehat{\mu}(\Delta))$ *for testing* $H_\Delta : S(\tau) = \Delta$ *have a central* $\chi^2(1)$ *null limiting distribution.*

Corollary 8.1 implies that $WCI_g(\mathbf{y})$, $PSCI(\mathbf{y})$, and $PLCI(\mathbf{y})$ of (1) are indeed approximate $1 - \alpha$ level confidence intervals when the conditions of Theorem 8.1 are satisfied.

8.2 Estimability Result

It is of interest to know when an estimand $S(\pi)$ can be estimated using data that do not come from the joint distribution π, but rather from conditional distributions with table probabilities $\tau \equiv \mathbf{t}(\pi) \equiv \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\pi)\pi$. For example, consider $(A, B) \sim \{\pi_{ij}, i = 1, 2,\ j = 1, 2, 3\}$ and the global odds ratio estimand

$$S(\pi) \equiv \frac{odds(A = 2|B \geq 2)}{odds(A = 2|B < 2)} = \frac{\pi_{11}(\pi_{22} + \pi_{23})}{\pi_{21}(\pi_{12} + \pi_{13})}.$$

An interesting question is this: If data were collected via stratified random sampling from the $A|B = j$ distributions could $S(\pi)$ be estimated? As we shall see in Example 8.2 below, the answer is 'no' in that the $A|B = j$ table probabilities $\mathbf{t}(\pi)$ do not identify a unique value of $S(\pi)$.

The following definition of estimability is based on the idea that an estimand $S(\pi)$ should be considered estimable using table probabilities only if each possible table probability vector identifies a unique value of $S(\pi)$. The definition can be shown to be equivalent to the one given earlier in Lang (2004).

Definition 2. *Estimand* $S(\pi)$ *is* $\mathbf{Z}$ *estimable* [*over* $\Omega \equiv \{\pi : \pi > 0, 1^T\pi = 1\}$] *if* $S(\pi)$ *is a function of the table probabilities* $\mathbf{t}(\pi)$ [*over* Ω].

Another way to state the definition is that estimand $S(\pi)$ is estimable using data from the unrestricted model $MP(\mu|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ if $S(\pi)$ is a function of the table probabilities $\mathbf{t}(\pi)$. The model $MP(\mu|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ is unrestricted in that the table probabilities $\mathbf{t}(\pi)$ are free to take on any value in the set $\mathcal{T}$ defined in (4); equivalently the joint probabilities in π are free to take on any value in Ω.

The phrases in brackets in the definition can be omitted for the purposes of the current paper. They are included to anticipate a generalized version of estimability that will be described in a subsequent paper that discusses model-based estimation of $S(\pi)$, which uses restricted versions of $MP(\mu|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$.

Using the definition to show that $S(\pi)$ is $\mathbf{Z}$ estimable is not always so straightforward because we must show that $\mathbf{t}(\pi_1) = \mathbf{t}(\pi_2)$ implies that $S(\pi_1) = S(\pi_2)$. To address this difficulty, Lang (2004) gave a relatively simple sufficient condition for estimability. Here we also give a relatively simple necessary condition.

Theorem 8.2: *Suppose that S is defined on $\Omega \equiv \{\pi : \pi > 0, \mathbf{1}^T\pi = 1\}$ and let S_0 be a 0-order 1 homogeneous version of S. Then the estimand $S(\pi)$ is $\mathbf{Z}$-estimable if and only if S_0 is 0-order $\mathbf{Z}$-homogeneous.*

The necessary condition of Theorem 8.2 gives a relatively simple way to establish non-estimability. It also lends itself to a computer-based approach to detecting that an estimand $S(\pi)$ is *not* $\mathbf{Z}$ estimable. Specifically, one could compute $|S_0(\mathbf{D}(\mathbf{Z}\gamma)\mathbf{x}) - S_0(\mathbf{x})|$ for several randomly selected γ and $\mathbf{x}$ values. If the difference is ever non-zero then S is not $\mathbf{Z}$ estimable.

Theorem 8.2 implies that if $S(\pi) = S(\mathbf{t}(\pi))$, where $\mathbf{t}(\pi) \equiv \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\pi)\pi$ is a vector of table probabilities, then $S(\pi)$ is $\mathbf{Z}$ estimable. This follows because S is both 0-order 1 and 0-order $\mathbf{Z}$ homogeneous. Theorem 8.2 also leads directly to the following corollary.

Corollary 8.2: *Under the conditions of Theorem 8.2, if $S(\pi)$ is $\mathbf{Z}$-estimable then $S(\pi) = S_0(\tau)$, where $\tau = \mathbf{t}(\pi)$ is the vector of table probabilities and S_0 is any 0-order 1 homogeneous version of S.*

Corollary 8.2 shows explicitly how $\mathbf{Z}$-estimable $S(\pi)$ is a function of the table probabilities $\mathbf{t}(\pi)$. By invariance of MLE's, Corollary 8.2 implies that the unrestricted MLE of $S(\pi)$ is $S_0(\widehat{\tau})$, where $\widehat{\tau}$ is the MLE of the table probabilities.

Example 8.1. Let A and B be two dichotomous random variables with joint distribution $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$. Two estimands of interest are the odds ratio $S_1(\pi) \equiv \pi_{11}\pi_{22}/(\pi_{12}\pi_{21})$ and the relative risk $S_2(\pi) \equiv (\pi_{11}/\pi_{1+})/(\pi_{21}/\pi_{2+})$. Consider the simple random sampling plan, row stratified $(B|A)$ random sampling plan, and column stratified $(A|B)$ random sampling plan.

The corresponding population matrices are, respectively,

$$\mathbf{Z}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{Z}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{Z}_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

It is easy to verify that the odds ratio is 0-order homogeneous with respect to all three of the sampling plans. It follows that $S_1(\boldsymbol{\pi})$ is estimable using data from any of these three sampling plans; this is a well known result. Because S_1 is a 0-order 1 homogeneous version of itself and because $S_1(\boldsymbol{\pi})$ is estimable for all three sampling plans, Corollary 8.2 implies that $S_1(\boldsymbol{\pi}) = S_1(\boldsymbol{\tau})$, where $\boldsymbol{\tau}$ is the vector of table probabilities for any of the three sampling plans.

The relative risk estimand function S_2 is 0-order $\mathbf{Z}_1$ and 0-order $\mathbf{Z}_2$ homogeneous. It is not, however, 0-order $\mathbf{Z}_3$ homogeneous because $\mathbf{D}(\mathbf{Z}_3\boldsymbol{\gamma})\mathbf{x} = (\gamma_1 x_{11}, \gamma_2 x_{12}, \gamma_1 x_{21}, \gamma_2 x_{22})$ implies that

$$S_2(\mathbf{D}(\mathbf{Z}_3\boldsymbol{\gamma})\mathbf{x}) = \frac{\gamma_1 x_{11}/(\gamma_1 x_{11} + \gamma_2 x_{12})}{\gamma_1 x_{21}/(\gamma_1 x_{21} + \gamma_2 x_{22})}$$

and as an example, for $\boldsymbol{\gamma} = (1, 2)$ and $\mathbf{x} = (2, 1, 1, 1)$,

$$S_2(\mathbf{D}(\mathbf{Z}_3\boldsymbol{\gamma})\mathbf{x}) = \frac{2/(2+2)}{1/(1+2)} \neq \frac{2/3}{1/2} = S_2(\mathbf{x}).$$

It follows that the relative risk estimand $S_2(\boldsymbol{\pi})$ is estimable using data from a simple random sample or a row stratified sample. The necessary condition of Theorem 8.2 implies that it is *not* estimable using data from the column stratified sample.

Example 8.2. Consider $(A, B) \sim \{\pi_{ij}, i = 1, 2, \ j = 1, 2, 3\}$ and the global odds ratio estimand

$$S(\boldsymbol{\pi}) \equiv \frac{odds(A = 2|B \geq 2)}{odds(A = 2|B < 2)} = \frac{\pi_{11}(\pi_{22} + \pi_{23})}{\pi_{21}(\pi_{12} + \pi_{13})}.$$

For the row stratified $(B|A)$ sampling plan with table probabilities $\boldsymbol{\tau}$, it is straightforward to directly show that $S(\boldsymbol{\pi}) = S(\boldsymbol{\tau})$ and hence $S(\boldsymbol{\pi})$ is estimable. Alternatively, it is easy to verify that Theorem 8.2's sufficient condition for estimability holds.

In contrast, the global odds ratio $S(\boldsymbol{\pi})$ is *not* estimable using data from a column stratified $(A|B)$ sampling plan. To see this, first note that S is 0-order 1 homogeneous. By Theorem 8.2, the non-estimability will be proven once we show that S is not 0-order homogeneous with respect to the column stratified sampling plan. The population matrix for column stratified sampling plan has the form

$$\mathbf{Z}^T = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

Thus, $\mathbf{D}(\mathbf{Z}\gamma)\mathbf{x} = [\gamma_1 x_{11}, \gamma_2 x_{12}, \gamma_3 x_{13}, \gamma_1 x_{21}, \gamma_2 x_{22}, \gamma_3 x_{23}]^T$ and

$$S(\mathbf{D}(\mathbf{Z}\gamma)\mathbf{x}) = \frac{\gamma_1 x_{11}(\gamma_2 x_{22} + \gamma_3 x_{23})}{\gamma_1 x_{21}(\gamma_2 x_{12} + \gamma_3 x_{13})}.$$

But, as an example, for $\gamma = (1, 1, 2)$ and $\mathbf{x} = (1, 1, 1, 1, 2, 3)$,

$$S(\mathbf{D}(\mathbf{Z}\gamma)\mathbf{x}) = \frac{1(2+6)}{1(1+2)} \neq \frac{1(2+3)}{1(1+1)} = S(\mathbf{x}),$$

so S is not 0-order $\mathbf{Z}$ homogeneous.

Example 8.3. Let A and B be ordinal variables with joint distribution probabilities in π. The Pearson correlation estimand $S(\pi)$ is *not* estimable under row or column stratified sampling. The non-estimability can be shown using the necessary condition of Theorem 8.2. In particular, it is straightforward to show that S_0, a 0-order 1 homogeneous version of S, is not 0-order homogeneous with respect to the row or column sampling plan. Alternatively, it is easy to see by example that table probabilities for the row or column sampling plan do not uniquely identify a correlation, $S(\pi)$, value. As an example, in the 2×2 table setting, the two joint distributions $\pi^{(1)} = (4, 1, 1, 4)/10$ and $\pi^{(2)} = (4, 1, 10, 40)/55$ correspond with the same row conditional distributions, viz. $\mathbf{t}(\pi^{(1)}) = \mathbf{t}(\pi^{(2)}) = (0.8, 0.2, 0.2, 0.8)$. However, for $\{0, 1\}$ row and column scores, $S(\pi^{(1)}) = 0.6 \neq 0.396 = S(\pi^{(2)})$.

8.3 Sampling Plan Invariance Results:

Suppose that data $\mathbf{y} \leftarrow MP(\sigma, \tau | \mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ are to be used to estimate $S(\tau)$. Unless stated otherwise, assume throughout this section that S_0 is a 0-order $\mathbf{Z}$ homogeneous version of S that satisfies conditions (S1) and (S2) of (6). This means that the model under $H_\Delta : S(\tau) = \Delta$ can be expressed as an MPH model.

Definition 3 (MPH Estimation Object). *The* MPH estimation object, *denoted* $est[S(\tau), \mathbf{y}, (\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})]$, *is composed of the following pieces of information:* (i) *Estimand* $S(\tau)$; (ii) *ML estimate* $S(\widehat{\tau})$ *and approximate standard error* $ase(S(\widehat{\tau}))$; *and* (iii) *Confidence intervals* $WCI_g(\mathbf{y})$, $PSCI(\mathbf{y})$, *and* $PLCI(\mathbf{y})$ *as given in* (1).

Theorem 8.3 (Sampling Plan Invariance). *Consider the MPH estimation object* $est[S(\tau), \mathbf{y}, (\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})]$ *and an alternative data model* $\mathbf{y} \leftarrow MP(\sigma^*, \tau^* | \mathbf{Z}^*, \mathbf{Z}_F^*, \mathbf{n}^*)$. *If* S_0 *is 0-order* $\mathbf{Z}^*$ *homogeneous then*

$$est[S(\tau), \mathbf{y}, (\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})] \;=\; est[S_0(\tau^*), \mathbf{y}, (\mathbf{Z}^*, \mathbf{Z}_F^*, \mathbf{n}^*)].$$

The proof of Theorem 8.3, which is outlined in the Appendix, is based on general equivalence results for MPH models as described in Lang (2004). The equivalence of the MPH estimation objects implies that (i) $S(\tau) = S_0(\tau^*)$; (ii) $S(\widehat{\tau}) = S_0(\widehat{\tau}^*)$, $ase(S(\widehat{\tau})) = ase(S_0(\widehat{\tau}^*))$; and (iii) $WCI_g(\mathrm{y}) = WCI_g^*(\mathrm{y})$, $PSCI(\mathrm{y}) = PSCI^*(\mathrm{y})$, and $PLCI(\mathrm{y}) = PLCI^*(\mathrm{y})$.

Corollary 8.3. *For convenience, the MPH estimation object* $est[S(\tau), \mathrm{y}, (\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})]$ *can be computed using a full-multinomial or an independent Poisson data model. In symbols,*

$$est[S(\tau), \mathrm{y}, (\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})] = est[S_0(\pi), \mathrm{y}, (1, 1, 1^T\mathrm{y})] = est[S_0(\pi), \mathrm{y}, (1, 0, 0)].$$

Corollary 8.3 follows immediately from Theorem 8.3 because when S_0 is 0-order $\mathbf{Z}$ homogeneous it must also be 0-order 1 homogeneous.

Example 8.4. If $S(\pi)$ is $\mathbf{Z}$ estimable then Theorem 8.2 implies that S_0, a 0-order 1 homogeneous version of S, is also 0-order $\mathbf{Z}$ homogeneous. Corollary 8.3 implies that $est[S(\pi), \mathrm{y}, (\mathbf{Z}, \mathbf{Z}_F, \mathrm{n})]$ can be more simply computed using the full-multinomial estimation object $est[S_0(\pi), \mathrm{y}, (1, 1, 1^T\mathrm{y})]$ or the independent-Poisson estimation object $est[S_0(\pi), \mathrm{y}, (1, 0, 0)]$.

Example 8.5. Let A and B be dichotomous random variables with $P(A = i, B = j) = \pi_{ij}, i, j = 1, 2$. Suppose we wish to estimate the odds ratio $S(\pi) = \pi_{11}\pi_{22}/(\pi_{12}\pi_{21})$ based on data coming from a row $(B|A)$ stratified sampling plan. Because $S(\pi)$ is estimable with these data, Example 8.4 implies that for computing the Wald or profile confidence intervals, we can treat the data as though the components were independent Poisson realizations. We also would get the same results if we assumed the data came from a single four-celled multinomial.

Example 8.6. Suppose that 5×2 Poisson table counts $\mathrm{y} \leftarrow MP(\sigma, \pi|1, 0, 0)$ are to be used to estimate the conditional probability $S(\pi) = \pi_{11}/\pi_{1+}$. Theorem 8.3 implies that $est[\pi_{11}/\pi_{1+}, \mathrm{y}, (1, 0, 0)] = est[\tau_{11}^*, \mathrm{y}, (\mathbf{Z}^*, \mathbf{Z}^*, \mathbf{n}^*)]$, where $\mathbf{Z}^*$ corresponds with the row stratified sampling plan. In words, the independent Poisson estimation object for the conditional probability is identical to the product-binomial estimation object. The result follows because $S_0 \equiv S$ is 0-order 1 and $\mathbf{Z}^*$ homogeneous and $S_0(\tau^*) = \tau_{11}^*/\tau_{1+}^* = \tau_{11}^*$. Note that the product-binomial estimation is arguably more straight-forward for this example.

Example 8.7. Suppose that subjects are accrued over a period of one year, and each subject is cross-classified on sex and two other ordinal variables A and B. The resulting counts can be modeled as $\mathrm{y} \leftarrow MP(\sigma, \pi|1, 0, 0)$, where σ is the expected number of subjects accrued annually and $\pi_{kij} = P(SEX = k, A = i, B = j)$. The estimand of interest is the difference between the conditional Pearson correlations, $S(\pi) \equiv \rho_1 - \rho_2$, where $\rho_k = corr(A, B|SEX = k)$. More

specifically, if $\{a_i\}$ and $\{b_j\}$ are the A and B scores, then

$$\rho_k = \frac{\sum_i \sum_j a_i b_j \left(\frac{\pi_{kij}}{\pi_{k++}}\right) - \left[\sum_i a_i \left(\frac{\pi_{ki+}}{\pi_{k++}}\right)\right]\left[\sum_j b_j \left(\frac{\pi_{k+j}}{\pi_{k++}}\right)\right]}{\sqrt{\left(\sum_i a_i^2 \left(\frac{\pi_{ki+}}{\pi_{k++}}\right) - \left[\sum_i a_i \left(\frac{\pi_{ki+}}{\pi_{k++}}\right)\right]^2\right)\left(\sum_j b_j^2 \left(\frac{\pi_{k+j}}{\pi_{k++}}\right) - \left[\sum_j b_j \left(\frac{\pi_{k+j}}{\pi_{k++}}\right)\right]^2\right)}}, \quad k = 1, 2.$$

It is straightforward to see that $S(\boldsymbol{\pi})$ is both 0-order 1 homogeneous and 0-order $\mathbf{Z}$ homogeneous, where $\mathbf{Z}$ corresponds with the SEX-specific stratified sampling plan. Theorem 8.3 implies that $est[S(\boldsymbol{\pi}), \mathbf{y}, (1, 0, 0)] = est[S(\boldsymbol{\tau}), \mathbf{y}, (\mathbf{Z}, \mathbf{Z}, \mathbf{n})]$, where the table probabilities are defined as $\tau_{kij} \equiv \pi_{kij}/\pi_{k++}$. In words, asymptotic inference about the difference between the conditional correlations is the same whether we treat the data as independent Poisson or as product-multinomial.

9 Numerical Examples

9.1 Example 3.1 Revisited

Consider the data displayed in tabular form in Example 3.1. These data can be collected in the vector $\mathbf{y} = (25, 25, 12, 0, 1, 3)$ and modeled as product-multinomial:

$$\mathbf{y} \leftarrow \mathbf{Y} \sim MP(\boldsymbol{\sigma}, \boldsymbol{\tau} | \mathbf{Z}, \mathbf{Z}_F, \mathbf{n}), \text{ where } \mathbf{Z}^T = \mathbf{Z}_F^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \text{ and } \mathbf{n} = (62, 4).$$

Note that the table probabilities are defined as $\tau_{ij} = \pi_{ij}/\pi_{i+}$, where $\pi_{ij} \equiv P(A = i, B = j)$, and both expected sample sizes are known *a priori*, viz. $\boldsymbol{\sigma} = \mathbf{n} = (62, 4)$.

Because the unrestricted ML estimate of $\boldsymbol{\tau}$ is the vector of sample proportions $\widehat{\boldsymbol{\tau}} = (25/62, 25/62, 12/62, 0/4, 1/4, 3/4)$, the ML estimate of $\gamma^* = S(\boldsymbol{\tau})$ is $S(\widehat{\boldsymbol{\tau}}) = 0.9358$. The maximum likelihood estimate of $\Omega_1 = S_1(\boldsymbol{\tau})$ does not exist because the sample proportion $\widehat{\tau}_{21} = 0$. An asymptotically equivalent estimate is $S_1(\widetilde{\boldsymbol{\tau}}) = 6.71$, where $\widetilde{\boldsymbol{\tau}}$ is the vector of sample proportions based on the original counts plus 0.5. The maximum likelihood estimate of $\Omega_2 = S_2(\boldsymbol{\tau})$ is $S_2(\widehat{\boldsymbol{\tau}}) = 12.50$.

Table 1 gives profile score, profile likelihood, and Wald confidence intervals for the three estimands $\gamma^* = S(\boldsymbol{\tau})$, $\Omega_1 = S_1(\boldsymbol{\tau})$, and $\Omega_2 = S_2(\boldsymbol{\tau})$. For illustrative purposes, two Wald intervals are computed for each estimand. The first Wald interval is based on $W_{g_1}^2$, where g_1 is the identity function and the other Wald interval is based on $W_{g_2}^2$, where g_2 is the logit function for γ^* and the log function for the global odds ratios. The Wald intervals for Ω_1 used modified data of the form '*original count* + 0.5.' In contrast, the profile intervals for Ω_1 required no such data modification because the unrestricted ML estimates were not used in the computation.

Table 1. Case-Control Data: Estimates and Nominal 95% Confidence Intervals

Estimand	Estimate	Profile Score	Profile Likelihood	Wald (g_1)	Wald (g_2)
γ^*	0.936	[0.525, 0.990]	[0.610, 0.996]	[0.813, 1.058]	[0.655,0.991]
Ω_1	6.711	[0.661, ∞)	[1.042, ∞)	[-13.086, 26.507]†	[0.351, 128.221]†
Ω_2	12.500	[1.598, 93.771]	[1.460, 265.326]	[-16.865, 41.865]	[1.193, 130.967]

† These intervals used modified data of the form '*original count* + 0.5.'

For each of the estimands, the four confidence intervals are quite different, which is hardly unexpected in this sparse-data setting. All three of the Wald (g_1) intervals include out-of-range values; the intervals include γ^* values bigger than 1 and global odds ratio values less than 0. This highlights one of the effects of poor choice of parameterization. The Wald (g_2) intervals use better parameterizations that do not lead to out-of-range values.

In contrast to the Wald (g_1) interval, the profile intervals and the Wald (g_2) interval for γ^* are asymmetric about the point estimate 0.9358; this is intuitively reasonable because the ML estimator of γ^* undoubtedly has a distribution that is skewed to the left, away from the upper bound of 1.0. Although the four intervals for γ^* are quite different, they all lead to the conclusion that the estimate 0.9358 is statistically higher than 0.50; that is, there is a statistically significant association between smoking and myocardial infarction.

In contrast to the Wald (g_1) interval, the profile intervals and the Wald (g_2) interval lead to the same conclusion that the estimate of the global odds ratio Ω_2 is statistically bigger than 1; that is, there is a statistically significant association between smoking and myocardial infarction. The conclusions regarding the statistical significance of the Ω_1 estimate are mixed. The score and Wald intervals lead to a conclusion of non-significance, whereas the likelihood ratio interval leads to a conclusion of significance.

Estimability of a Modified Gamma Measure. The estimand γ^* was referred to as a *modified gamma measure of association.* To see why this is a reasonable label consider the following: The estimand γ^* can be written as

$$\gamma^* = S(\tau) = \frac{C(\tau)}{C(\tau) + D(\tau)},$$

where $C(\tau) \equiv 2\sum_i \sum_j \tau_{ij} \left(\sum_{h>i} \sum_{k>j} \tau_{hk}\right)$ and $D(\tau) \equiv 2\sum_i \sum_j \tau_{ij} \left(\sum_{h>i} \sum_{k<j} \tau_{hk}\right)$. If $\pi = (\pi_{11}, \pi_{12}, \ldots, \pi_{23})$ is the joint probability distribution of (A, B), then $C(\pi)$ and $D(\pi)$ are the probabilities of concordance and discordance used in the computation of Goodman and

Kruskal's gamma. For Example 3.1, $S(\pi) = C(\pi)/[C(\pi) + D(\pi)]$ is estimable because S is 0-order homogeneous with respect to the sampling plan. The modified gamma label is reasonable because Corollary 8.2 implies that $\gamma^* \equiv S(\tau) = S(\pi)$.

Interestingly, using the necessary condition of Theorem 8.2, it is straightforward to show that modified gamma $S(\pi)$ would *not* be estimable if there were more than two rows in the row stratified sampling scheme.

Invariance Result. Let $S_0(\tau) \equiv S(\mathbf{t}(\tau))$ be the 0-order $\mathbf{Z}$ homogeneous version of $S(\tau) = \gamma^*$. Corollary 8.3 implies that the estimation object $est[\gamma^*, \mathbf{y}, (\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})]$ is identical to the independent Poisson estimation object $est[S_0(\pi), \mathbf{y}, (\mathbf{1}, 0, 0)]$. This sampling plan invariance result was exploited in the actual computations. Similarly, the estimation objects for the other estimands were also computed using the independent Poisson version.

9.2 Example 3.2 Revisited

Consider the data displayed in tabular form in Example 3.2. These data can be collected in the 25×1 vector $\mathbf{y} = (0, 0, 0, \ldots, 2, 20)$ and modeled as multinomial:

$$\mathbf{y} \leftarrow \mathbf{Y} \sim MP(\sigma, \tau | \mathbf{Z}, \mathbf{Z}_F, n), \text{ where } \mathbf{Z} = \mathbf{Z}_F = \mathbf{1}_{25} \text{ and } n = 25.$$

Note that the table probabilities are defined as $\tau_{ij} = \pi_{ij}$, where $\pi_{ij} \equiv P(A = i, B = j)$, and the expected sample size is known *a priori*, viz. $\sigma = n = 25$.

Because the unrestricted ML estimate of τ is the vector of sample proportions $\hat{\tau} = (0/25, 0/25, 0/25, \ldots, 2/25, 20/25)$, the unrestricted maximum likelihood estimates of the estimands are as follows: $\widehat{E(R_1)} - \widehat{E(R_2)} = S(\hat{\tau}) = 0.20$, $\widehat{E(R_1)} = S_1(\hat{\tau}) = 4.84$, $\widehat{E(R_2)} = S_2(\hat{\tau}) = 4.64$, $\widehat{D(R_1)} = S_3(\hat{\tau}) = 0.2176$, and $\widehat{D(R_2)} = S_4(\hat{\tau}) = 0.3424$.

Table 2 gives profile score, profile likelihood, and Wald confidence intervals for the five estimands. The Wald intervals are based on the Wald statistic W_g^2, where g is the identity transformation.

Table 2. Golf Club Ratings: Estimates and Nominal 95% Confidence Intervals

Estimand	Estimate	Profile Score	Profile Likelihood	Wald
$E(R_1) - E(R_2)$	0.20	[-0.3594, 0.7061]	[-0.1331, 0.5193]	[0.0079, 0.3920]
$E(R_1)$	4.84	[4.3286, 4.9455]	[4.5209, 4.9601]	[4.6584, 5.0215]
$E(R_2)$	4.64	[4.0412, 4.8501]	[4.1417, 4.8776]	[4.2915, 4.9885]
$D(R_1)$	0.2176	[0.0806, 0.4619]	[0.0610, 0.4421]	[0.0103, 0.4249]
$D(R_2)$	0.3424	[0.1659, 0.5691]	[0.1451, 0.5583]	[0.1187, 0.5661]

Because $E(R_1)$ and $E(R_2)$ are evidently close to the upper limit of 5.0, it is intuitively appealing that the profile intervals are asymmetric about the point estimate. In contrast, the Wald intervals for the means are symmetric about the point estimates. Not surprisingly, the Wald interval for $E(R_1)$ includes out-of-range values.

The profile intervals are a bit narrower than the respective Wald intervals for the dispersion estimands. In contrast, the profile intervals are quite a bit wider than the Wald interval for the difference between the two mean ratings. A small-scale simulation study of a comparable setting hints that for the difference in mean ratings, the wider profile intervals are in fact a bit conservative in that the coverage probabilities are higher than the nominal 0.95; the Wald interval for the difference had coverage probability that was just a bit lower than the nominal 0.95.

10 Discussion

The MPH interval estimation method described in this paper is directly applicable for many of the contingency table estimands (aka parameters) encountered in practice. Exceptions include estimands that are non-smooth, estimand functions that have vanishing derivatives, and estimands that cannot be expressed as $S(\boldsymbol{\tau})$, where S is a known estimand function. The computational algorithm that carries out the MPH interval estimation only requires the user to input the table counts $\mathbf{y}$, the sampling plan $(\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$, and the estimand function S. The algorithm has been coded in R (Ihaka and Gentleman 1996; see also `http://cran.r-project.org`) and is available from the author upon request.

The examples in Section 9 used very sparse tables to emphasize the potential for large and practically meaningful differences among the three confidence intervals considered herein. Although the profile intervals may usually have better coverage properties than the Wald in-

tervals, especially in these sparse table settings, the researcher still needs to be mindful that the intervals are only as good as the asymptotic approximations upon which they are based.

The estimability results of Section 8.2 are not just of theoretical interest; they are very important in practice. Because all contingency tables of counts look superficially similar and because most contingency table software packages treat them all similarly, it is all too easy to compute and report "estimates" of non-estimable parameters. The estimability results highlight the need to recognize the sampling plan that generated the table counts. For example, most contingency table software will not hesitate to output estimates of both the gamma and Pearson correlation coefficients for any 2×3 table. If the ambiguously labeled "2×3 table" happens to be the result of a single random sample, there is no problem, as both gamma and the Pearson correlation are estimable in this case. However, for row stratified sampling, only gamma is estimable, and this is only because there are only two rows. For column stratified sampling, neither gamma nor the Pearson correlation are estimable.

This paper considered only *non-model-based* estimation of contingency table parameters. In particular, profile intervals were based on inverting tests of $H_\Delta : S(\tau) = \Delta$ vs. unrestricted $K_\Delta : S(\tau) \neq \Delta$. When there is good reason to believe that a parsimonious model, say $\mathcal{M}$, holds or nearly holds, a more efficient model-based estimate of $S(\tau)$ could be considered. In this case, for example, one could compute profile likelihood intervals by inverting tests of $H^*_\Delta : [\mathcal{M}$ and $S(\tau) = \Delta]$ vs. restricted $K^*_\Delta : [\mathcal{M}$ but $S(\tau) \neq \Delta]$. Specific examples of model-based profile interval estimation appear in Gart (1985) and Lang (2005). A more general description of model-based interval estimation will be given in a subsequent paper.

11 Appendix

11.1 Proofs of Properties 7.1-7.3

Proof of Property 7.1: We must show (i) $\forall \mathbf{x} > 0, \mathbf{t}(\mathbf{x}) \in \mathcal{T}$; (ii) $\forall \tau \in \mathcal{T}, \exists \mathbf{x} > 0 \ni \mathbf{t}(\mathbf{x}) = \tau$; and (iii) $\mathbf{t}$ is 0-order $\mathbf{Z}$ homogeneous. The proofs will make repeated use of the following property of population matrices $\mathbf{Z}$: For any $\gamma > 0$, $\mathbf{Z}^T\mathbf{D}(\mathbf{Z}\gamma) = \mathbf{D}(\gamma)\mathbf{Z}^T$.

Proof of (i). For any $\mathbf{x} > 0$, $\mathbf{t}(\mathbf{x}) > 0$ and, letting $\gamma \equiv 1/\mathbf{Z}^T\mathbf{x}$,

$$\mathbf{Z}^T\mathbf{t}(\mathbf{x}) = \mathbf{Z}^T\mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{x})\mathbf{x} = \mathbf{Z}^T\mathbf{D}(\mathbf{Z}\gamma)\mathbf{x} = \mathbf{D}(\gamma)\mathbf{Z}^T\mathbf{x} = \mathbf{D}^{-1}(\mathbf{Z}^T\mathbf{x})\mathbf{Z}^T\mathbf{x} = 1.$$

Thus, $\mathbf{t}(\mathbf{x}) \in \mathcal{T}$. This proves (i).

Proof of (ii). Let $\boldsymbol{\tau} \in \mathcal{T}$. Notice that for $\mathbf{x} = \boldsymbol{\tau}$, $\mathbf{t}(\mathbf{x}) = \boldsymbol{\tau}$ and $\mathbf{x} > 0$. This proves (ii).

Proof of (iii). Let $\boldsymbol{\gamma} > 0$ and $\mathbf{x} > 0$. It follows that

$$\begin{aligned}\mathbf{t}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) &= \mathbf{D}^{-1}\left(\mathbf{Z}\mathbf{Z}^T\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}\right)\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x} = \mathbf{D}^{-1}\left(\mathbf{Z}\mathbf{D}(\boldsymbol{\gamma})\mathbf{Z}^T\mathbf{x}\right)\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x} \\ &= \mathbf{D}^{-1}\left(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{Z}\mathbf{Z}^T\mathbf{x}\right)\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x} = \mathbf{D}^{-1}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{x})\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x} \\ &= \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{x})\mathbf{x} = \mathbf{t}(\mathbf{x}).\end{aligned}$$

This proves (iii) and hence Property 7.1 is proven. ■

Proof of Property 7.2: Suppose that $\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\boldsymbol{\tau}) = \mathbf{G}(\boldsymbol{\gamma})\mathbf{h}(\boldsymbol{\tau}), \forall \boldsymbol{\gamma} > 0, \forall \boldsymbol{\tau} \in \mathcal{T}$. Let $\boldsymbol{\gamma} > 0$ and $\mathbf{x} > 0$. Define $\mathbf{t}(\mathbf{x}) = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{x})\mathbf{x}$. Property 7.1 shows that $\mathbf{t}(\mathbf{x}) \in \mathcal{T}$. Moreover,

$$\begin{aligned}\mathbf{h}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) &= \mathbf{h}\left(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{D}(\mathbf{Z}\mathbf{Z}^T\mathbf{x})\mathbf{t}(\mathbf{x})\right) = \mathbf{G}(\boldsymbol{\gamma} \cdot \mathbf{Z}^T\mathbf{x})\mathbf{h}(\mathbf{t}(\mathbf{x})) \\ &= \mathbf{G}(\boldsymbol{\gamma})\mathbf{G}(\mathbf{Z}^T\mathbf{x})\mathbf{h}(\mathbf{t}(\mathbf{x})) = \mathbf{G}(\boldsymbol{\gamma})\mathbf{h}(\mathbf{D}(\mathbf{Z}\mathbf{Z}^T\mathbf{x})\mathbf{t}(\mathbf{x})) = \mathbf{G}(\boldsymbol{\gamma})\mathbf{h}(\mathbf{x}).\end{aligned}$$

This proves the result. ■

Proof of Property 7.3: Let $\boldsymbol{\gamma} > 0$ and $\mathbf{x} > 0$. Note that

$$\mathbf{f}_0(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) = \mathbf{f}\left(\mathbf{t}(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x})\right) = \mathbf{f}(\mathbf{t}(\mathbf{x})) = \mathbf{f}_0(\mathbf{x}).$$

This proves (i). Now, for any $\mathbf{x} \in \mathcal{T}$, we have $\mathbf{t}(\mathbf{x}) = \mathbf{x}$. Thus, $\mathbf{f}_0(\mathbf{x}) = \mathbf{f}(\mathbf{t}(\mathbf{x})) = \mathbf{f}(\mathbf{x}),\ \forall \mathbf{x} \in \mathcal{T}$. This proves (ii). ■

11.2 Proof of Theorem 8.2.

First note that there always exists a 0-order $\mathbf{1}$ homogeneous version of S. For example, $S_0(\boldsymbol{\pi}) \equiv S(\boldsymbol{\pi}/\mathbf{1}^T\boldsymbol{\pi})$ works. Throughout the proof, S_0 is any 0-order $\mathbf{1}$ homogeneous version of S.

Proof (Sufficiency): S_0 is 0-order $\mathbf{Z}$ homogeneous implies that $S_0(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) = S_0(\mathbf{x}), \forall \boldsymbol{\gamma} > 0, \forall \mathbf{x} > 0$. Let $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \Omega$ satisfy $\mathbf{t}(\boldsymbol{\pi}_1) = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\boldsymbol{\pi}_1)\boldsymbol{\pi}_1 = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\boldsymbol{\pi}_2)\boldsymbol{\pi}_2 = \mathbf{t}(\boldsymbol{\pi}_2)$. Then

$$\begin{aligned}S(\boldsymbol{\pi}_1) &= S_0(\boldsymbol{\pi}_1) = S_0\left(\mathbf{D}(\mathbf{Z}\mathbf{Z}^T\boldsymbol{\pi}_1)\mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\boldsymbol{\pi}_2)\boldsymbol{\pi}_2\right) \\ &= S_0\left(\mathbf{D}(\mathbf{Z}(\tfrac{\mathbf{Z}^T\boldsymbol{\pi}_1}{\mathbf{Z}^T\boldsymbol{\pi}_2}))\boldsymbol{\pi}_2\right) = S_0(\boldsymbol{\pi}_2) = S(\boldsymbol{\pi}_2).\end{aligned}$$

The second last equality follows because S_0 is 0-order $\mathbf{Z}$ homogeneous. It follows that $S(\boldsymbol{\pi})$ is a function of $\boldsymbol{\tau} = \mathbf{t}(\boldsymbol{\pi})$ alone.

Proof (Necessity): Assume that $S(\boldsymbol{\pi})$ is a function of $\mathbf{t}(\boldsymbol{\pi})$ alone, but S_0 is not 0-order $\mathbf{Z}$ homogeneous. We prove the necessity result by showing that this leads to a contradiction.

If S_0 is not 0-order $\mathbf{Z}$ homogeneous then there exists a $\boldsymbol{\gamma} > 0$ and a $\mathbf{x} > 0$ such that $S_0(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\mathbf{x}) \neq S_0(\mathbf{x})$. Because S_0 is 0-order $\mathbf{1}$ homogeneous,

$$\exists \boldsymbol{\gamma} > 0, \mathbf{x} > 0, \ni\ S_0\left(\mathbf{D}(\mathbf{Z}\boldsymbol{\gamma})\frac{\mathbf{x}}{\mathbf{1}^T\mathbf{x}}\right) \neq S_0(\frac{\mathbf{x}}{\mathbf{1}^T\mathbf{x}}).$$

[Note: This follows because $S_0(\frac{1}{1^T\mathbf{x}}\mathbf{D}(\mathbf{Z}\gamma)\mathbf{x}) = S_0(\mathbf{D}(\mathbf{Z}\gamma)\mathbf{x})$ and $S_0(\frac{1}{1^T\mathbf{x}}\mathbf{x}) = S_0(\mathbf{x})$.] Now $\pi_1 \equiv \mathbf{x}/1^T\mathbf{x}$ lies in Ω. Thus,

$$\exists \gamma > 0, \pi_1 \in \Omega, \ni \; S_0(\mathbf{D}(\mathbf{Z}\gamma)\pi_1) \neq S_0(\pi_1).$$

Again, because S_0 is 0-order 1 homogeneous, this implies that

$$\exists \gamma > 0, \pi_1 \in \Omega, \ni \; S_0\left(\frac{\mathbf{D}(\mathbf{Z}\gamma)\pi_1}{1^T\mathbf{D}(\mathbf{Z}\gamma)\pi_1}\right) \neq S_0(\pi_1).$$

Let $\pi_2 \equiv \mathbf{D}(\mathbf{Z}\delta)\pi_1$, where $\delta \equiv \frac{\gamma}{1^T\mathbf{D}(\mathbf{Z}\gamma)\pi_1}$; i.e. π_2 is the argument of S_0 on the left hand side of the previous inequality. Note that π_2 lies in Ω, so we have

$$S_0(\pi_2) \neq S_0(\pi_1), \quad \text{for } \pi_1, \pi_2 \in \Omega. \tag{7}$$

But, owing to the 0-order $\mathbf{Z}$ homogeneity of function t defined as $\mathbf{t}(\mathbf{x}) = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{x})\mathbf{x}$, we have

$$\mathbf{t}(\pi_1) = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\pi_1)\pi_1 = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\mathbf{D}(\mathbf{Z}\delta)\pi_1)\mathbf{D}(\mathbf{Z}\delta)\pi_1 = \mathbf{D}^{-1}(\mathbf{Z}\mathbf{Z}^T\pi_2)\pi_2 = \mathbf{t}(\pi_2),$$

which implies that $S(\pi_1) = S(\pi_2)$ because it was assumed that $S(\pi)$ is a function of $\mathbf{t}(\pi)$ alone. Furthermore, because S_0 is a 0-order 1 homogeneous version of S, we have that $S_0(\pi_1) = S(\pi_1) = S(\pi_2) = S_0(\pi_2)$, which contradicts (7). This proves the necessity result and hence Theorem 8.2 is proven. ■

11.3 Proof of Theorem 8.3.

We need to show that (i) $S(\tau) = S_0(\tau^*)$; (ii) $S(\widehat{\tau}) = S_0(\widehat{\tau}^*)$, $ase(S(\widehat{\tau})) = ase(S_0(\widehat{\tau}^*))$; and (iii) $WCI_g(\mathbf{y}) = WCI_g^*(\mathbf{y})$, $PSCI(\mathbf{y}) = PSCI^*(\mathbf{y})$, and $PLCI(\mathbf{y}) = PLCI^*(\mathbf{y})$.

To prove (i), note that $S(\tau) = S_0(\tau) = S_0(\pi)$ because S_0 is a 0-order $\mathbf{Z}$ homogeneous version of S. But S_0 is also 0-order $\mathbf{Z}^*$ homogeneous and $\tau^* = \mathbf{D}^{-1}(\mathbf{Z}^*\mathbf{Z}^{*T}\pi)\pi$, so it follows that $S_0(\tau^*) = S_0(\pi)$. Therefore, the two estimands $S(\tau)$ and $S_0(\tau^*)$ are identical.

To prove (ii), note that the two unrestricted MPH models $\mathbf{y} \leftarrow MP(\sigma, \tau|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ and $\mathbf{y} \leftarrow MP(\sigma^*, \tau^*|\mathbf{Z}^*, \mathbf{Z}_F^*, \mathbf{n}^*)$ are equivalent in the formal sense described in Lang (2004); specifically, they both are members of the equivalence class denoted by $\mathcal{E}(0, \mathbf{y})$. It follows that numerically $\widehat{\mu} = \widehat{\mu}^*$ and hence

$$S(\widehat{\tau}) = S_0(\widehat{\tau}) = S_0(\widehat{\mu}) = S_0(\widehat{\mu}^*) = S_0(\widehat{\tau}^*).$$

The string of equalities exploit the fact that $\widehat{\mu} = \mathbf{D}(\mathbf{Z}\widehat{\sigma})\widehat{\tau}$ and S_0 is 0-order $\mathbf{Z}$ homogeneous, and $\widehat{\mu}^* = \mathbf{D}(\mathbf{Z}^*\widehat{\sigma}^*)\widehat{\tau}^*$ and S_0 is 0-order $\mathbf{Z}^*$ homogeneous. Now because S_0 satisfies (S1) and

(S2) of (6), it follows that $S_0(\widehat{\mu})$ is a 0-order $\mathbf{Z}$ homogeneous statistic as defined in Lang (2004). Similarly $S_0(\widehat{\mu}^*)$ is a 0-order $\mathbf{Z}^*$ homogeneous statistic. Because $\widehat{\mu}$ and $\widehat{\mu}^*$ are based on equivalent models, Corollary 3 of Theorem 8 in Lang (2004) states that the approximate distributions of $S_0(\widehat{\mu}) = S(\widehat{\tau})$ and $S_0(\widehat{\mu}^*) = S_0(\widehat{\tau}^*)$ are identical. This implies that numerically $ase(S(\widehat{\tau})) = ase(S_0(\widehat{\tau}^*))$. It is worthwhile noting that Corollary 2 of Theorem 8 in Lang (2004) implies that the squares of the numerically identical approximate standard errors, namely $avar(S(\widehat{\tau}))$ and $avar(S_0(\widehat{\tau}^*))$, have particularly simple forms:

$$\begin{aligned} avar(S(\widehat{\tau})) &= \frac{\partial S_0(\widehat{\tau})}{\partial \tau^T}\mathbf{N}^{-1}\mathbf{D}(\widehat{\tau})\frac{\partial S_0^T(\widehat{\tau})}{\partial \tau} \\ avar(S_0(\widehat{\tau}^*)) &= \frac{\partial S_0(\widehat{\tau}^*)}{\partial \tau^{*T}}\mathbf{N}^{*-1}\mathbf{D}(\widehat{\tau}^*)\frac{\partial S_0^T(\widehat{\tau}^*)}{\partial \tau^*}, \end{aligned}$$

where $\mathbf{N} = \mathbf{D}(\mathbf{Z}\mathbf{Z}^T\mathbf{y})$ and $\mathbf{N}^* = \mathbf{D}(\mathbf{Z}^*\mathbf{Z}^{*T}\mathbf{y})$. The derivation of these approximate variance forms used the identity $\partial S_0(\mu)/\partial \mu^T = (\partial S_0(\tau)/\partial \tau^T)\mathbf{D}^{-1}(\mathbf{Z}\sigma)$ and the fact that $\widehat{\sigma} = \mathbf{Z}^T\mathbf{y}$ for MPH models.

To prove (iii), first note that the arguments in the previous paragraph directly generalize as follows: For differentiable g, $g(S(\widehat{\tau})) = g(S_0(\widehat{\tau}^*))$ and $ase(g(S(\widehat{\tau}))) = ase(g(S_0(\widehat{\tau}^*)))$. These equivalences imply that $WCI_g(\mathbf{y}) = WCI_g^*(\mathbf{y})$.

The model $\mathbf{y} \leftarrow MP(\sigma, \tau|\mathbf{Z}, \mathbf{Z}_F, \mathbf{n})$ under $H_\Delta : S(\tau) = \Delta$ can be specified using the 0-order $\mathbf{Z}$ homogeneous constraint $h(\mu) \equiv S_0(\mu) - \Delta = 0$. The model $\mathbf{y} \leftarrow MP(\sigma^*, \tau^*|\mathbf{Z}^*, \mathbf{Z}_F^*, \mathbf{n}^*)$ under $H_\Delta : S_0(\tau^*) = \Delta$ can be specified using the 0-order $\mathbf{Z}^*$ homogeneous constraint $h(\mu^*) \equiv S_0(\mu^*) - \Delta = 0$. Thus, the two restricted models are MPH models that use the identical constraint function. That is, in the formal sense of Lang (2004), these two MPH models are equivalent and are members of the equivalence class $\mathcal{E}(h, \mathbf{y})$. The equivalence results of Lang (2004) imply that the restricted model ML estimates are numerically identical: $\widehat{\mu}(\Delta) = \widehat{\mu}^*(\Delta)$; this means that the two score statistics are numerically identical and the two likelihood ratio statistics are numerically identical. Moreover, the score and likelihood ratio statistics have $\chi^2(1)$ limiting null distributions for both models. It immediately follows that $PSCI(\mathbf{y}) = PSCI^*(\mathbf{y})$ and $PLCI(\mathbf{y}) = PLCI^*(\mathbf{y})$. ■

References

Agresti, A. (2002), *Categorical Data Analysis*, 2nd ed, New York: John Wiley & Sons.

Agresti, A. and Caffo, B. (2000), "Simple and Effective Confidence Intervals for Proportions

and Differences of Proportions Result from Adding Two Success and Two Failures," *The American Statistician,* 54, 280-288.

Agresti, A. and Coull, B.A. (1998), "Approximate is Better than Exact for Interval Estimation of Binomial Parameters," *American Statistician,* 52, 119-126.

Bedrick, E.J. (1987), "A Family of Confidence Intervals for the Ratio of Two Binomial Proportions," *Biometrics,* 43, 993-998.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975), *Discrete Multivariate Analysis,* Cambridge, MA: MIT Press.

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

Casella, G. and Berger, R.L. (1990), *Statistical Inference,* Belmont, CA: Wadsworth.

DiCiccio, T.J., Martin, M.A., Stern, S.E., and Young, G.A. (1996), "Information Bias and Adjusted Profile Likelihoods," *Journal of the Royal Statistical Society–Series B,* 58, 189-203.

Gart, J.J. (1985), "Approximate Test and Interval Estimation of the Common Relative Risk in the Combination of 2×2 Tables," *Biometrika,* 72, 673-677.

Gart, J.J. and Nam, J. (1988a), "Approximate Interval Estimation of the Ratio of Binomial Parameters: A Review and Corrections for Skewness," *Biometrics,* 44, 323-338.

Gart, J.J. and Nam, J. (1988b), "Alternative Confidence Intervals for the Relative Difference," *The Statistician,* 37, 427-431.

Ihaka, R. and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational Graphics and Statistics,* 5, 299-314.

Lang, J.B. (2004), "Multinomial-Poisson Homogeneous Models for Contingency Tables," *The Annals of Statistics,* 32, 340-383.

Lang, J.B. (2005), "Homogeneous Linear Predictor Models for Contingency Tables," *Journal of the American Statistical Association,* 100, 121-134.

Lloyd, C.J. (1990), "Confidence Intervals from the Difference Between Two Correlated Proportions," *Journal of the American Statistical Association,* 85, 1154-1158.

May, W.L. and Johnson, W.D. (1997), "Confidence Intervals for Differences in Correlated Binary Proportions," *Statistics in Medicine,* 16, 2127-2136.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models,* 2nd ed, London: Chapman

and Hall.

Mee, R.W. (1984), "Confidence Bounds for the Difference between Two Probabilities," Correspondence in *Biometrics*, 40, 1175-1176.

Meeker, W.Q. and Escobar, L.A. (1995), "Teaching about Approximate Confidence Regions Based on Maximum Likelihood Estimation," *The American Statistician*, 49, 48-53.

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

Newcombe, R.G. (1998a), "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, 17, 857-872.

Newcombe, R.G. (1998b), "Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods," *Statistics in Medicine*, 17, 873-890.

Newcombe, R.G. (2003), "Confidence Intervals for the Mean of a Variable Taking the Values 0,1 and 2," *Statistics in Medicine*, 22, 2737-2750.

Quesenberry, C.P. and Hurst, D.C. (1964), "Large Sample Simultaneous Confidence Intervals for Multinomial Proportions," *Technometrics*, 6, 191-195.

Read, T.R.C. and Cressie, N.A.C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*, New York: Springer-Verlag.

Tang, N.S. and Tang, M.L. (2003), "Statistical Inference for Risk Difference in an Incomplete Correlated 2×2 Table," *Biometrical Journal*, 45, 34-46.

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

Tsimikas, J.V., Bosch, R.J., Coull, B.A., and El Bharmi, H. (2002), "Profile-Likelihood Inference for Highly Accurate Diagnostic Tests," *Biometrics*, 58, 946-956.