

Exact Inference on Estimating a Linear Combination of Multinomial Probabilities.

Haben Michael¹, Angelina Chen², and Lu Tian³

¹University of Massachusetts

²affil2

SUMMARY: It is often needed to perform inference on several proportions based on dependent data. For example, an employer may want a confidence interval for the proportion of man-hours lost due to illness within a given time-frame, based on employee's attendance records. It is equivalent to making inference on a linear combination of a multinomial probability vector. The conventional likelihood-based approach relies on a large sample approximation and may have a poor performance when the multinomial probability vector lies close to the boundary of the unit simplex, the parameter space of the probability vector. On the other hand, the validity of our method is always guaranteed since it is based on the inversion of exact tests. We find substantial improvements in its robustness over the conventional approach with small sample size, particularly when the multinomial parameter is sparse. We illustrate our method by applying it to analyze data on the impact of air pollution data on breathing difficulty.

KEYWORDS: Binary outcomes; Longitudinal data; Rare events.

1 Introduction

We consider the problem of forming a confidence interval for a linear combination of a multinomial parameter based on a single observation X . That is, given an observation X distributed as multinomial with count n and probabilities $\mathbf{p} = (p_1, \dots, p_m)$, we describe methods for obtaining a confidence interval (CI) for

$$\theta_0 = \mathbf{c}^t \mathbf{p} = \sum_{i=1}^m \mathbf{c}_i p_i. \quad (1)$$

where $\mathbf{c} = (\mathbf{c}_1, \dots, \mathbf{c}_m)$ are given known constants. Our CI is exact in the sense that it is based on a test statistic, whose distribution is known up to easily controlled monte carlo error, rather than asymptotically or otherwise approximated. An exact CI is important for applications involving rare events, small sample sizes, or both these conditions.

1.1 Motivating example

As a motivating example, suppose a researcher from a hospital or insurance company is interested in the cumulative incidence rate of an event within a time window in a population of interest. For example, patients are given a preliminary screening test at regular time interval and there is health-related or economic cost associated with each positive test result. Then the total cost is proportional to the cumulative incidence rate. The data are modeled as n vectors of length $m - 1$, $m > 1$, consisting of binary values,

$$(X_{11}, X_{12}, \dots, X_{1(m-1)}), (X_{21}, X_{22}, \dots, X_{2(m-1)}), \dots, (X_{n1}, X_{n2}, \dots, X_{n(m-1)}), X_{ij} \in \{0, 1\}.$$

The target parameter of inference is

$$\theta_0 = E \left(\frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^{m-1} X_{ij} \right).$$

The dependence structure among the repeated observations $(X_{i1}, X_{i2}, \dots, X_{i(m-1)})$ is left unspecified while the n vectors are assumed identically independently distributed (IID), such as one might obtain by administering the tests to randomly selected subjects from a population. Since each X_{ij} is 0 or 1, the sums $\sum_{j=1}^{m-1} X_{ij}, i = 1, \dots, n$, are IID random variables each taking an integer value in $\{0, 1, \dots, m-1\}$. Viewing $0, 1, \dots, m-1$, as m categories, we identify the value of $\sum_{j=1}^{m-1} X_{ij}$ as a choice from these m categories. Therefore, $\sum_{j=1}^{m-1} X_{ij}, i = 1, \dots, n$ are IID, and their sum is multinomial with parameters n and \mathbf{p}_0 , where the i^{th} component of \mathbf{p}_0 is

$$p_{0i} = P \left(\sum_{j=1}^{m-1} X_{1j} = i-1 \right), i = 1, \dots, m.$$

It follows that

$$\theta = \sum_{i=1}^m \frac{(i-1)}{(m-1)} p_{0i}.$$

1.2 Literature

Many studies in the health sciences regularly measure a rarely occurring event over time. A standard method of analysis is generalized estimating equations, and variants that take into account the rarity of the event (Schaefer, 1983; Cordeiro and McCullagh, 1991; Bull et al., 1997; Cordeiro and Cribari-Neto, 1998; Leung and Wang, 1998; Anderson and Blair, 1982; Self and Liang, 1987). These methods are mainly intended to ascertain the relation between the events and available covariates, which in turn usually requires imposing modeling assumptions. As we are interested only in prevalence, we can use non-parametric methods. There is also a long line of research into exact tests and CIs for contingency table data. Overviews are given in Mehta (1994); Agresti (2001). These exact methods are typically based on test inversion like ours. These methods, however, are not designed for dependent outcomes, as our data require.

The remainder of the paper is organized as follows. In Section 2 we describe the construction of the proposed CI, first from a theoretical standpoint in Section 2.1, then considering practical aspects in Section 2.2. In Section 3 we examine the coverage and power of the proposed CI using synthetic data, comparing it to a standard CI. In Section 4 we apply the proposed method to form a CI for the prevalence of wheezing in a population of children. We conclude in Section 5 with suggestions for future research. Software implementing the proposed method and the routines used in the simulation section of the paper are publicly available at the website of the corresponding author.

2 Method

The general problem is forming a CI for a real function of a multinomial parameter vector. One solution is to find a level $1 - \alpha$ CI for the multinomial parameter vector and transform it to obtain a level $1 - \alpha$ CI for the function of the multinomial parameter vector. Since a multinomial parameter is a probability mass function, this solution involves a type of density estimation. We take advantage of the assumption that the function is a linear combination.

2.1 Inference on a linear combination of the multinomial parameter by test inversion

Let X be an observation from the multinomial distribution with sample size n and parameter \mathbf{p}_0 , i.e., $MN(n, \mathbf{p}_0)$. Let \mathbf{c} be a vector of length m , not necessarily $\mathbf{c}_0 = (0, \dots, m-1)^t / (m-1)$ as above, though we continue to assume the components of \mathbf{c} are nonnegative. The goal is a confidence interval for $\theta_0 = \mathbf{c}^t \mathbf{p}_0$ based on X_0 .

One CI is given by maximum likelihood estimator (MLE) of \mathbf{p}_0 , $\hat{\mathbf{p}} = X/n$. The MLE of θ_0 is $\mathbf{c}^t \hat{\mathbf{p}}$, whose distribution can be approximated by

$$N \left(\theta_0, \frac{1}{n} \mathbf{c}^t (\text{diag}(\mathbf{p}) - \mathbf{p} \mathbf{p}^t) \mathbf{c} \right).$$

It's variance may be approximated by

$$\frac{1}{n}\hat{\sigma}^2 = \frac{1}{n}\mathbf{c}^t(\text{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}^t)\mathbf{c}.$$

Therefore, a Wald-type 95% CI for θ_0 can be constructed as

$$\left[\mathbf{c}^t\hat{\mathbf{p}} - \frac{1.96}{\sqrt{n}}\hat{\sigma}, \mathbf{c}^t\hat{\mathbf{p}} + \frac{1.96}{\sqrt{n}}\hat{\sigma} \right].$$

One drawback of this CI is that it need not lie in the parameter space for θ . For example, for $\mathbf{c} = \mathbf{c}_0$, θ is a proportion but the CI need not lie within $[0, 1]$. Another drawback is that for a given finite sample size, the coverage of this CI deteriorates as the multinomial parameter \mathbf{p}_0 approaches the boundary of the parameter space, the simplex in \mathbb{R}^m . We therefore look for a more reliable CI.

We may obtain an exact CI by inverting an exact hypothesis test. Let $T = T(X, \mathbf{p})$ be a function of the data X and a parameter value \mathbf{p} . Choices of T are discussed below. A level α test of the null that $X \sim MN(n, \mathbf{p})$ rejects for large values of T , i.e., $T(x_0, \mathbf{p}) \geq t_{\mathbf{p}, \alpha}$, where x is the observed data, $t_{\mathbf{p}, \alpha}$ is the $1 - \alpha$ quantile of $T(X, \mathbf{p})$, where $X \sim MN(n, \mathbf{p})$. A valid level α test of the composite null that

$$H_0(\theta) : \mathbf{p}_0 \in \{\mathbf{p} \in \Delta_+^{m-1} : \mathbf{c}^t\mathbf{p} = \theta\},$$

rejects when

$$\inf_{\mathbf{p} \in \Omega_\theta} \{T(x_0, \mathbf{p}) - t_{\mathbf{p}, \alpha}\} > 0, \quad (2)$$

where $\Omega(\theta) = \{\mathbf{p} \in \Delta_+^{m-1} : \mathbf{c}^t\mathbf{p} = \theta\}$, $\Delta_+^{m-1} = \{\mathbf{p} = (p_1, p_2, \dots, p_m)^t \in \mathbb{R}^m \mid \mathbb{1}^t\mathbf{p} = 1, p_i \geq 0\}$ and $\mathbb{1}$ is the vector consisting of all ones. The set of parameters θ at which the test fails to reject,

$$CI(x_0) = \left\{ \theta : \inf_{\mathbf{p} \in \Omega_\theta} [T(x_0, \mathbf{p}) - t_{\mathbf{p}, \alpha}] \leq 0 \right\}, \quad (3)$$

contains θ_0 with probability $\geq 1 - \alpha$. It is because

$$\begin{aligned} & P(\theta_0 \in CI(X_0)) \\ &= P\left(\inf_{\mathbf{p} \in \Omega_{\theta_0}} [T(X_0, \mathbf{p}) - t_{\mathbf{p}, \alpha}] \leq 0\right) \\ &\geq P(T(X_0, \mathbf{p}_0) \leq t_{\mathbf{p}_0, \alpha}), \quad \text{since } \mathbf{c}^t\mathbf{p}_0 = \theta_0 \\ &\geq 1 - \alpha. \end{aligned}$$

The set $CI(x_0)$ may therefore serve as a level $1 - \alpha$ CI for θ_0 . Another perspective is to view

$$\Omega_{\mathbf{p}}(x_0) = \{\mathbf{p} \mid T(x_0, \mathbf{p}) < t_{\mathbf{p}, \alpha}\}$$

as a $100(1 - \alpha)\%$ confidence region for probability vector \mathbf{p} and $CI(x_0)$ as its projection onto $\mathbf{c}^t\mathbf{p}$:

$$CI(x_0) = \left[\inf_{\mathbf{p} \in \Omega_{\mathbf{p}}(x_0)} \mathbf{c}^t\mathbf{p}, \sup_{\mathbf{p} \in \Omega_{\mathbf{p}}(x_0)} \mathbf{c}^t\mathbf{p} \right].$$

Computing the quantiles $t_{\mathbf{p}, \alpha}$ requires the distribution of the test statistic $T(X, \mathbf{p})$, where $X \sim MN(n, \mathbf{p})$, whose analytic form is often complex. In such cases, the distribution may be approximated, to arbitrary accuracy, by simulation. As the quantiles are then only computed at a finite number of select values \mathbf{p} , the minimization of $T(x_0, \mathbf{p}) - t_{\mathbf{p}, \alpha}$ over the set $\Omega(\theta)$ in (3) is in turn approximated by taking the minimum over a finite grid of $\Omega(\theta)$. To construct the CI for θ_0 , one needs to repeat this minimization for a set of $\theta \in [L, U]$. Further details on the algorithm are given below. This CI is exact, i.e., its mean coverage no smaller than the nominal coverage, subject to provisos:

1. There is monte carlo error, which may be reduced arbitrarily by increasing the tuning parameters: The numbers n_θ and $n_{\mathbf{p}}$ of points θ and $\mathbf{p} \in \Omega_\theta$ selected, and the size B of the empirical distribution used in computing the quantiles $t_{\mathbf{p},\alpha}$.
2. The null hypothesis $H_0 : \mathbf{p}_0 \in \Omega(\theta)$ is a composite null hypothesis, so that the test statistic on which the CI is based is typically conservative. That is, the null consists of multiple \mathbf{p} s and the corresponding distributions of test statistics and (2) suggests the least favorable p value. This conservativeness is part of the definition of a p-value for a composite null, and, due to the equivalence of CIs and hypothesis testing, unavoidable. The difference is the gap between largest and smallest p-values for different \mathbf{p} s with the same $\mathbf{c}^t \mathbf{p}$ value. The difference depends on how sensitive the distribution of the test statistic to the value of \mathbf{p} . If the distribution is approximately pivotal, i.e., independent of \mathbf{p} , then the result is less conservative. In simulations below, the effect is to inflate the coverage by about 1 – 2%. See Figure 2 for an illustration, where, among the values $\mathbf{p}_0 \in \Omega(\theta)$, the p-values near the boundary of the simplex are larger.
3. Discreteness: There are m^n possible values for X_0 sampled as multinomial of size n with m categories, so at most m^n possible values for a test statistic $T(X, \mathbf{p})$. Still fewer values may be observed when \mathbf{p} is close to the boundary of the simplex Δ^{m-1} , where some categories are rarely observed. There are then at most $m^n + 1$ possible values for a p-value. The nominal level of the test may not be among these p-values, in which case the p-value obtained under (2) will be larger than the nominal level. This issue may be addressed by introducing randomness to the test statistic, though in practice doing so has been described as “unacceptable,” the preference being to specify p-value cutoffs that lie among those made available by the data (Agresti, 2003).

The first point above is that the CI should have coverage equal to or exceeding the nominal level modulo monte carlo error, while the second and third show that the coverage may be strictly larger than the nominal level.

2.2 Algorithm

We propose the following algorithm for forming a CI for $\theta_0 = \mathbf{c}^t \mathbf{p}_0$ based on test inversion.

1. Select $\theta_1, \dots, \theta_{n_\theta}, n_\theta \geq 1$, in the interval $[L, U] = [\min_i \mathbf{c}_i, \max_i \mathbf{c}_i]$. The selection may be deterministic or sampled from a continuous distribution on the interval. E.g., if one is only interested in testing the null of a specific value θ , rather than forming a CI, only that value need be considered.
2. At each θ among $\theta_1, \dots, \theta_{n_\theta}$:
 - (a) Select $\mathbf{p}_1, \dots, \mathbf{p}_{n_{\mathbf{p}}}$, from $\{\mathbf{p} \in \mathbb{R}^m : \mathbf{c}^t \mathbf{p} = \theta\} \cap \Delta_+^{m-1}$, where $\Delta_+^{m-1} = \{\mathbf{p} \in \mathbb{R}^m : \mathbb{1}^t \mathbf{p} = 1, \mathbf{p} \geq 0\}$ is the probability simplex in \mathbb{R}^m . The number $n_{\mathbf{p}} = n_{\mathbf{p}}(\theta)$ may depend on θ , and the distribution of the points may reflect prior knowledge or interests. Methods for obtaining the points in this intersection are discussed below.
 - (b) At each \mathbf{p} among $\mathbf{p}_1, \dots, \mathbf{p}_{n_{\mathbf{p}}}$:
 - i. Sample $X_1^*, \dots, X_B^* \sim MN(n, \mathbf{p})$
 - ii. Set

$$\hat{q}(\mathbf{p}) = B^{-1} \sum_{i=1}^B I\{T(X_i^*, \mathbf{p}) \geq T(x_0, \mathbf{p})\},$$

an estimate of the p-value of the observed data at \mathbf{p} , where $I\{\cdot\}$ is an indicator function.

- (c) Set $\hat{q}(\theta) = \max_i \hat{q}(\mathbf{p}_i)$, an estimate of the p-value of the observed data at θ . This estimate is conservative as the maximum of $\hat{q}(\mathbf{p}_i)$ is used to estimate the maximum of $q(\mathbf{p}_i) = E\{\hat{q}(\mathbf{p}_i)\}$. In practice this conservativeness may be rendered negligible by ensuring the tuning parameter B is big relative to $n_{\mathbf{p}}$.
3. Take $\widehat{CI}(X_0) = \{\theta_i, 1 \leq i \leq n_\theta : \hat{q}(\theta_i) > \alpha\}$, an approximate of an exact level $1 - \alpha$ CI for θ_0 , where the approximation is in the sense discussed in Section 2.1.

In the following, we will discuss the operational details of (a), (b), (c) in this algorithm.

2.2.1 Details for (a): Sampling on $\{\mathbf{c}^t \mathbf{p} = 1\} \cap \Delta_+^{m-1}$

The key step of the algorithm is to sample a sufficient number of “representatives” from $\{\mathbf{p} \in \mathbb{R}^m \mid \mathbf{c}^t \mathbf{p} = \theta\} \cap \Delta_+^{m-1}$ to capture the range of possible values of the probability vector implied by the composite null $H_0(\theta)$. In this section, we describe two methods for obtaining points in $\{\mathbf{p} \in \mathbb{R}^m : \mathbf{c}^t \mathbf{p} = \theta\} \cap \Delta_+^{m-1}$, the intersection of a hyperplane with the probability simplex in \mathbb{R}^m . Assume for the moment that $\theta \neq 0$. Fixing θ and renaming \mathbf{c}/θ as \mathbf{c} , we rewrite the intersection as

$$\mathcal{S} = \{\mathbf{p} \in \mathbb{R}^m : \mathbf{c}^t \mathbf{p} = 1\} \cap \Delta_+^{m-1} = \{\mathbf{p} \in \mathbb{R}^m : \mathbf{c}^t \mathbf{p} = \mathbb{1}^t \mathbf{p} = 1, p_i \geq 0\}. \quad (4)$$

We first describe a fast method that samples non-uniformly on \mathcal{S} , then a slower method that samples uniformly.

• Approach 1.

Up to a constant factor, points in \mathcal{S} satisfy $(\mathbf{c} - \mathbb{1})^t \mathbf{p} = 0$. Therefore a simple method is

- Let $\mathbf{d} = \mathbf{c} - \mathbb{1} = (d_1, \dots, d_m)^t$ and $I_+ = \{i \mid d_i > 0\}$ and $I_- = \{i \mid d_i \leq 0\}$ denote the indices of the nonnegative and negative elements of \mathbf{d} , respectively. Assuming that \mathcal{S} is non-empty, the I_- is non-empty. Sample $\mathbf{u} = (u_1, \dots, u_m)^t$ from a continuous distribution with a support on the unit cube $[0, 1]^m$, and let $\tilde{\mathbf{p}} = (\tilde{p}_1, \dots, \tilde{p}_m)^t$, where

$$\tilde{p}_j = \begin{cases} u_j, & \text{if } j \in I_+ \\ -u_j \frac{\sum_{k \in I_+} d_k u_k}{\sum_{k \in I_-} d_k u_k}, & \text{if } j \in I_- \end{cases}.$$

Then, $\mathbf{d}^t \tilde{\mathbf{p}} = 0$, for $\tilde{\mathbf{p}} \in R_+^m$.

- Normalize the sampled $\tilde{\mathbf{p}}$

$$\tilde{\mathbf{p}} \leftarrow \frac{\tilde{\mathbf{p}}}{\mathbb{1}^t \tilde{\mathbf{p}}}$$

such that $\mathbf{c}^t \tilde{\mathbf{p}} = \mathbb{1}^t \tilde{\mathbf{p}} = 1$.

This sampling approach involves only simple calculation and fast. A drawback of this sampling approach is that the resulting $\tilde{\mathbf{p}}$ does not uniformly distributed over \mathcal{S} .

• Approach 2.

The intersection of \mathcal{S} is a convex polytope in \mathbb{R}_+^m , i.e., the convex hull of a finite set of vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_k\} \subset \mathbb{R}_+^m$. If these vertices are available, we may sample $\mathbf{w} = (w_1, \dots, w_k)^t$ from the simplex Δ_+^{k-1} and apply a linear transformation

$$\tilde{\mathbf{p}} \leftarrow \sum_{i=1}^k w_i \mathbf{v}_i,$$

to map the sample onto \mathcal{S} . If the initial sample $\tilde{\mathbf{p}}$ is sampled uniformly on the simplex, whether stochastically or deterministically, then, since non-degenerate linear transformations preserve uniformity (e.g., Devroye (2006)), the image will be uniformly distributed on \mathcal{S} . Other sampling schemes on the probability simplex such as a Dirichlet may be used to reflect prior knowledge about the location of the true parameter. This approach requires the vertices $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, which is equivalent to solving a linear programming problem. Avis and Fukuda (1991) give an algorithm that runs in time on the order of km . For a natural number \bar{k} A grid of $\binom{k+\bar{k}-1}{\bar{k}-1}$ points w_1, \dots, w_k , partitioning the simplex into identical small simplices is given by

$$\mathcal{P} = \left\{ \bar{k}^{-1} \mathbf{z} \mid \mathbf{z} = (z_1, \dots, z_m)^t, \sum_{i=1}^m z_i = \bar{k}, z_i \in \{0, 1, \dots, m\}, 1 \leq i \leq m \right\}.$$

For example, when $(\bar{k}, m) = (3, 3)$, it is not difficult to see that

$$\mathcal{P} = \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1/3 \\ 2/3 \end{pmatrix}, \begin{pmatrix} 0 \\ 2/3 \\ 1/3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/3 \\ 0 \\ 2/3 \end{pmatrix}, \right. \\ \left. \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}, \begin{pmatrix} 1/3 \\ 2/3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2/3 \\ 0 \\ 1/3 \end{pmatrix}, \begin{pmatrix} 2/3 \\ 1/3 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\}.$$

The value \bar{k} controls the density of samples on the probability simplex.

When $\theta = 0$, the set $\{\mathbf{p} \in \mathbb{R}^m : \mathbf{c}^t \mathbf{p} = \theta\} \cap \Delta_+^{m-1}$ is a lower-dimensional standard probability simplex. Let m' denote the number of zeros among the entries of \mathbf{c} . If all entries are positive then the intersection is empty. When $\mathbf{c}_0 = (0, \dots, m-1)^t / (m-1)$ is the vector discussed in motivating example, or any other choice of \mathbf{c} with a single 0, the intersection consists of a single point. Generally, the intersection is the probability simplex $\Delta^{[m'-1]}$ embedded in the m' -dimensional subspace of \mathbb{R} given by the zero entries of \mathbf{c} . Sampling on a standard probability simplex may be carried out deterministically as described above or randomly using the Dirichlet distribution.

2.2.2 Details for (b): The choice of the test statistic

Any choice of the test statistic T should, subject to monte carlo error, produce CIs with coverage equal to or exceeding the nominal level. However, some choices will offer narrower intervals. One choice is the studentized observation, centered at the test null,

$$T_1(X, \mathbf{p}) = \frac{|\mathbf{c}^t \hat{\mathbf{p}} - \mathbf{c}^t \mathbf{p}|}{\sqrt{\mathbf{c}^t (\text{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}} \hat{\mathbf{p}}^t) \mathbf{c}}}, \quad \hat{\mathbf{p}} = \frac{X}{n}.$$

One drawback of this statistic is that the denominator may vanish, leading to poor power. The denominator is especially likely to vanish near the boundary or when the multinomial count n is low. This difficulty may be addressed by a test statistic that shrinks the data toward the center of the simplex,

$$T_1(X, \mathbf{p}) = \frac{|\mathbf{c}^t \hat{\mathbf{p}} - \mathbf{c}^t \mathbf{p}|}{\sqrt{\mathbf{c}^t \hat{\Sigma}_1 \mathbf{c}}}$$

where $\hat{\Sigma}_1 = \text{diag}(\bar{\mathbf{p}}) - \bar{\mathbf{p}} \bar{\mathbf{p}}^t$, $\bar{\mathbf{p}} = \frac{X + m^{-1}}{n+1}$. Power against alternatives near the boundaries may be achieved by regularizing this test statistic,

$$T_2(X, \mathbf{p}) = \frac{|\mathbf{c}^t \hat{\mathbf{p}} - \mathbf{c}^t \mathbf{p}|}{\sqrt{\mathbf{c}^t \hat{\Sigma}_1 \mathbf{c}}} + \lambda (\hat{\mathbf{p}} - \mathbf{p})^t \hat{\Sigma}_2^{-1} (\hat{\mathbf{p}} - \mathbf{p}),$$

where $\hat{\Sigma}_2$ is obtained by omitting a single row and column pair from $\text{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}} \hat{\mathbf{p}}^t$, and $\lambda > 0$ is a tuning parameter selected a priori.

2.2.3 Details for (c): Coordinating tuning parameters

The sample size of the empirical distribution at each parameter \mathbf{p} is an additional tuning parameter. For a given θ , let $\mathbf{p}_1, \dots, \mathbf{p}_{n_{\mathbf{p}}}$ be the points sampled in the cross-section $\mathcal{S}_{\theta} = \{\mathbf{p} \mid \mathbf{c}^t \mathbf{p} = \theta\} \cap \Delta_+^{(m-1)}$. Let $q_1 = q(\mathbf{p}_1), \dots, q_{n_{\mathbf{p}}} = q(\mathbf{p}_{n_{\mathbf{p}}})$ denote the associated p-values, obtained as in step (b). If the p-values $q(\mathbf{p})$ depend continuously on the points \mathbf{p} , and if as $n_{\mathbf{p}} \rightarrow \infty$ the points $\mathbf{p}_1, \dots, \mathbf{p}_{n_{\mathbf{p}}}$ become dense in the intersection \mathcal{S}_{θ} , such as through sampling from a continuous distribution in the stochastic approaches or increasing the grid density in a deterministic approach, then

$$\max_{1 \leq i \leq n_{\mathbf{p}}} q(\mathbf{p}_i) \rightarrow \sup_{\mathbf{p} \in \mathcal{S}_{\theta}} q(\mathbf{p})$$

as $n_{\mathbf{p}} \rightarrow \infty$. At the same time, the fast convergence of the empirical to true CDF controls the error in approximating $q(\mathbf{p})$ via its empirical counterpart $\hat{q}(\mathbf{p}) = 1 - \hat{F}_{T(X, \mathbf{p})}(\mathbf{p})$, where $\hat{F}_{T(X, \mathbf{p})}$ is the empirical CDF based on a sample of size B of the test statistic at \mathbf{p} . The Dvoretzky-Kiefer-Wolfvitz inequality gives a universal constant C such that

$$\begin{aligned} P(|\max_{1 \leq i \leq n_{\mathbf{p}}} \hat{q}_i - \max_{1 \leq i \leq n_{\mathbf{p}}} q_i| > \epsilon) &\leq P(\cup_{i=1}^{n_{\mathbf{p}}} \{|\hat{q}_i - q_i| > \epsilon\}) \\ &\leq \sum_{i=1}^{n_{\mathbf{p}}} P(|\hat{q}_i - q_i| > \epsilon) \\ &\leq n_{\mathbf{p}} C \exp(-2B\epsilon^2). \end{aligned}$$

The last expression $\rightarrow 0$ when $B^{-1} \log n_{\mathbf{p}} = o(1)$. So to ensure convergence of the algorithm it is sufficient to have, for example, the number of monte carlo samples B be of the same order as the number of points $\mathbf{p}_1, \dots, \mathbf{p}_{n_{\mathbf{p}}}$, sampled in the intersections S_{θ} for all chosen θ .

3 Simulation

We use simulated data to verify the coverage of the approximate exact CI and compare it to the CI obtained using the maximum likelihood estimator (MLE). The dimension of the multinomial parameter is $m = 4$, as with the data discussed in Section 4. In the first set of simulation, the multinomial parameter \mathbf{p} underlying the estimand $\theta = \mathbf{c}^t \mathbf{p}$ is of the form $(\delta, (1 - \delta)/3, (1 - \delta)/3, (1 - \delta)/3)'$, where δ ranges between 0, the boundary of the probability simplex, and 1/4, where the parameter is balanced. The coefficient vector is $\mathbf{c} = (0, 1/3, 2/3, 1)^t$, the same choice considered in the motivating example in Section 1. In the second set of simulation, the multinomial probability vector $\mathbf{p} = (1 - \delta - \delta^2 - \delta^3, \delta, \delta^2, \delta^3)$, where δ ranges between 0 and 0.5. The sample size considered are $n = 10, 30, 50$.

Results are summarized in Figure 1. The observed coverage of the CIs are plotted against the distance δ of \mathbf{p} from the boundary of the parameter space. The CI based on the MLE falls below the nominal rate, worse as the sample size decreases or as the estimand is nearer the boundary of the parameter space. The proposed CI controls the FPR, remaining consistently 1-2% above the nominal rate. This gap is expected from the composite nature of the null hypothesis, as discussed above. There is a slight improvement in efficiency in the using the slower method, vertex enumeration, with the proposed CI.

Power. An approximation to the power surface of the hypothesis test on which the CI is based is given in Figure 2. The dimension of the parameter is $m = 3$. Contours are given for the observed rejection rate of the test statistic $\hat{q}(\mathbf{p}) = B^{-1} \sum_{i=1}^B I\{T(X_i^*, \mathbf{p}) \geq T(x_0, \mathbf{p})\}$ at values $\mathbf{p} = (p_1, p_2)$ in the simplex $\Delta^2 = \{(p_1, p_2, 1 - p_1 - p_2) : 0 \leq p_1, p_2 \leq 1\}$. The true multinomial parameter \mathbf{p} is marked as well as the line corresponding to the true value of the estimand, $\{(p_1, p_2) : \mathbf{p} \in \Delta^2, \mathbf{c}_0^t \mathbf{p} = \theta\}$. Following this line to the edge of the simplex, the rejection rate decreases, revealing a cause of the intrinsic conservativeness of the CI for the composite null discussed in Section 2.1.

4 Data Analysis

We consider the prevalence of wheezing observed in a population of repeatedly tested children. This parameter is of interest when the presence of wheezing requires a follow-up procedure and the aggregate resources for the follow-ups is to be estimated. The study consists of 537 children who were checked for wheezing annually in ages 7–10, giving 4 repeated measurements. The observed prevalence of days with wheezing is 0.152. The data is described further in Fitzmaurice and Laird (1993).

An approximate exact 95% confidence interval using the proposed method is (0.130, 0.181). The confidence interval based on the Fisher information is (0.130, 0.174). The length of the proposed CI is 0.055 compared to 0.043 for the MLE. Figure 4 gives the p-values of the hypothesis test inverted to form the proposed CI. The CI given here corresponds to those values θ for which the p-value exceed the level $\alpha = .05$, indicated by a horizontal line.

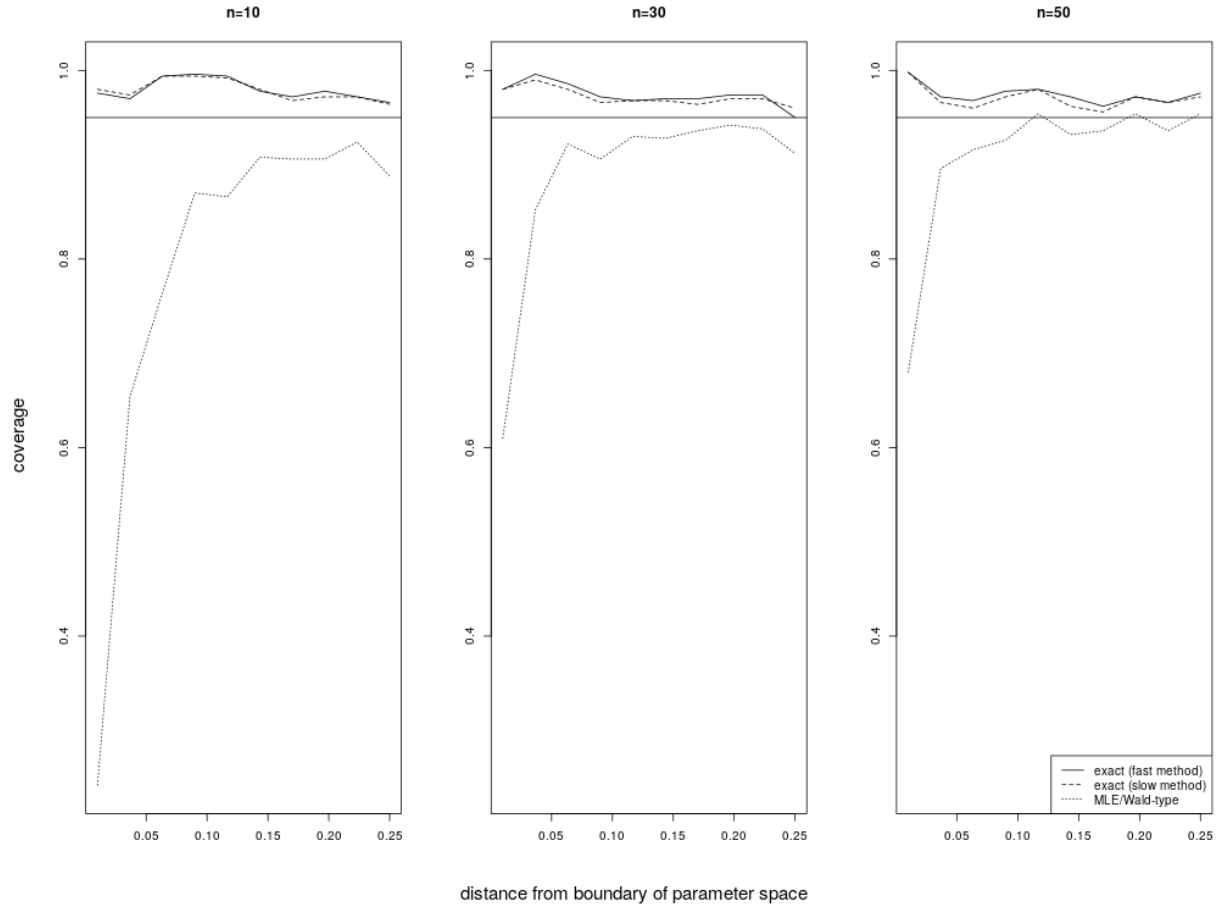


Figure 1: Simulation comparing the coverages of nominal 95% CIs based on the proposed method and the MLE.

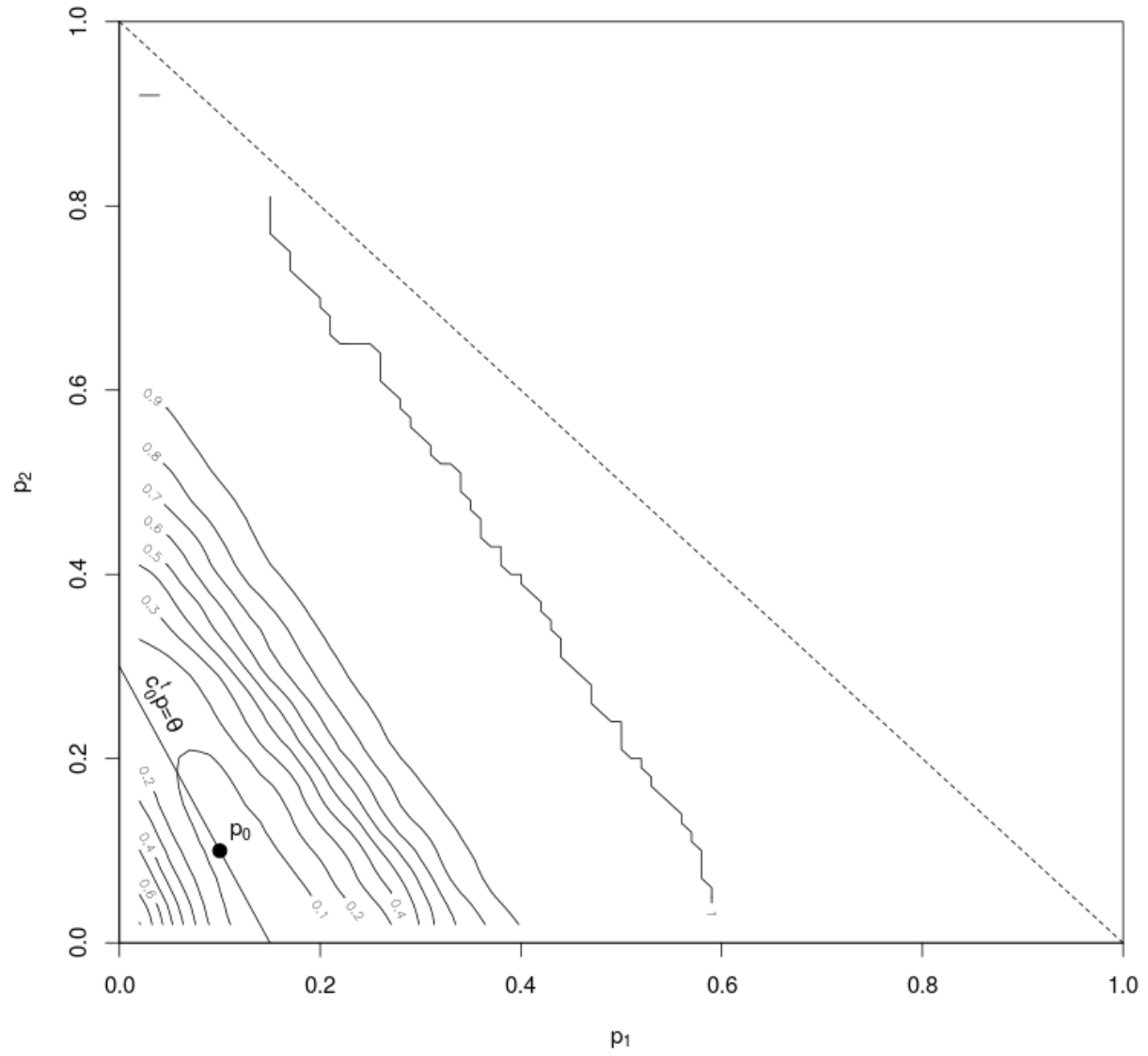


Figure 2: Power surface for $m = 3$. The line $\{\mathbf{p} : \mathbf{c}^t \mathbf{p} = \theta\}$ gives the \mathbf{p} values for the true θ .

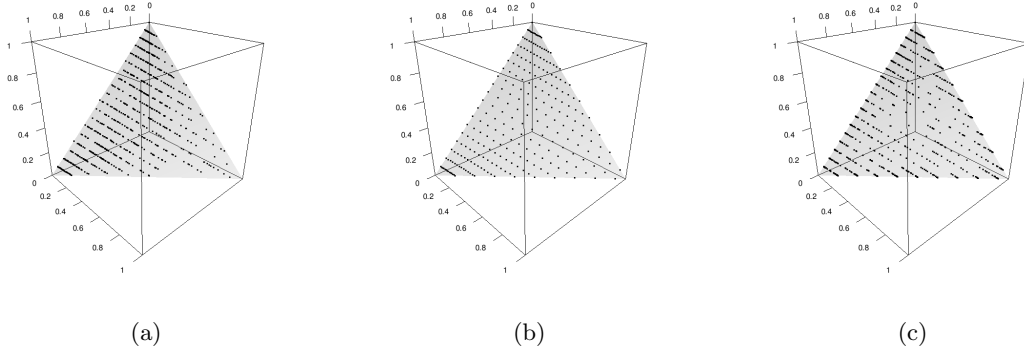


Figure 3: Sampling methods: (a) fast, non-uniform, (b) vertex enumeration using a deterministic choice of points, (c) vertex enumeration using a Dirichlet distribution with increased concentration near the edges.

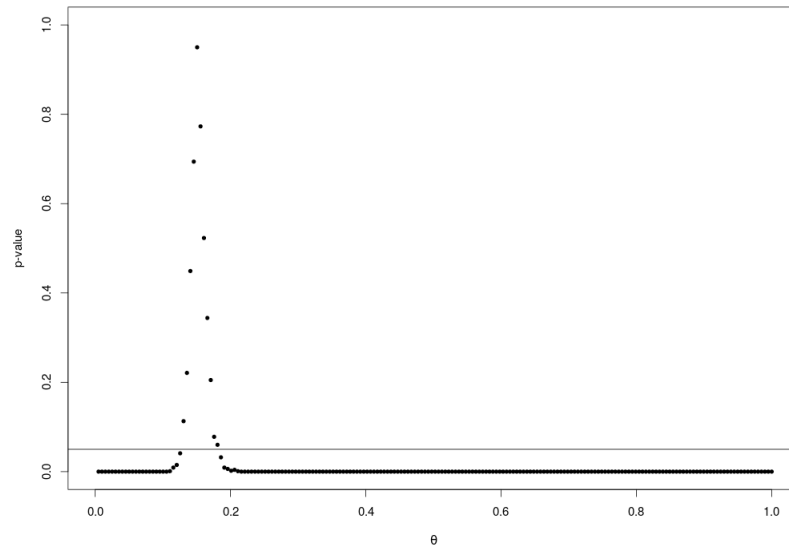


Figure 4: Air pollution data. P-values for the null hypothesis that the observed data follows a distribution in θ , for a grid of θ values.

5 Discussion

We have outlined a procedure for approximately exact CIs for a population prevalence based on repeated outcome measurements. In doing so we solved a more general problem, approximating an exact a CI for a given linear combination of a multinomial parameter $\mathbf{c}^T \mathbf{p}$ based on a single observation $X \sim MN(n, \mathbf{p})$.

Several extensions to the proposed method suggest themselves. First, the method described here can in principle be extended to form a confidence region for several prevalences $\theta_1, \theta_2, \dots$. Such a region is useful for inference on, e.g., contrasts for the prevalence of wheezing under different experimental conditions. Such a grid would require a grid on the cartesian product of several simplices, so that the computational complexity of a direct application of the proposed method would grow exponentially in the number of prevalences θ_i under consideration. However, for the common case of two prevalences, in where a CI for $\theta_1 - \theta_2$ is sought, the computational burden is manageable. A second, more complicated extension is to allow the number m of observations per patient to be random. Under additional assumptions on the nature randomness, the problem may be reformulated as a missing data problem.

References

- Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies. *Statistics in medicine* 20(17-18), 2709–2722.
- Agresti, A. (2003). Dealing with discreteness: Making exact confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical Methods in Medical Research* 12(1), 3–21.
- Anderson, J. and V. Blair (1982). Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika* 69(1), 123–136.
- Avis, D. and K. Fukuda (1991). A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. In *Proceedings of the seventh annual symposium on Computational geometry*, pp. 98–104.
- Bull, S. B., C. M. Greenwood, and W. W. Hauck (1997). Jackknife bias reduction for polychotomous logistic regression. *Statistics in Medicine* 16(5), 545–560.
- Cordeiro, G. M. and F. Cribari-Neto (1998). On bias reduction in exponential and non-exponential family regression models. *Communications in Statistics-Simulation and Computation* 27(2), 485–500.
- Cordeiro, G. M. and P. McCullagh (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 53(3), 629–643.
- Devroye, L. (2006). Nonuniform random variate generation. *Handbooks in operations research and management science* 13, 83–121.
- Fitzmaurice, G. M. and N. M. Laird (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* 80(1), 141–151.
- Leung, D. H.-Y. and Y.-G. Wang (1998). Bias reduction using stochastic approximation. *Australian & New Zealand Journal of Statistics* 40(1), 43–52.
- Mehta, C. R. (1994). The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research* 3(2), 135–156.
- Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine* 2(1), 71–78.
- Self, S. G. and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82(398), 605–610.