Exact, nonparametric confidence interval for incidence given repeated measurements

literature: overviews of exact, small sample methods for contingency table data: 1. Mehta CR. The exact analysis of contingency tables in medical research. Statistical Methods in Medical Research 1994; 3: 135–56. 2. Agresti A. Exact inference for categorical data: Recent advances and continuing controversies. Statistics in Medicine 2001; 20: 2709–22

1 Method

1.1 Data

The data are n IID vectors of length m consisting of binary values,

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

We don't make any assumptions about the dependence structure within each vector $(X_{i1}, X_{i2}, \ldots, X_{im})$. The sums $\sum_{j=1}^{m} X_{ij}, i = 1, \ldots, n$ are IID. Since each X_{ij} is 0 or 1, the sums are IID random variables each taking a value in $0, 1, \ldots, m$. Viewing $0, 1, \ldots, m$ as $m+1$ categories, we can identify each vector sum $\sum_{j=1}^{m} X_{ij}$ as a choice of one of thse $m+1$ categories. These choices are IID, so their sum is multinomial. ((rewrite m to $m-1$ to be consistent with the remainder?))

1.2 Multinomial estimation

Let X_0 be an observaiton from the multinomial distribution with sample size n and parameter p_0. Let $c = (0, \ldots, m-1)/(m-1)$. The goal is a confidence interval for $\theta_0 = c^t p_0$ based on X_0.

One CI is given by maximum likelihood. The MLE $c^t \hat{p}$ is asymptotically normal with variance $c^t(\mathrm{diag}(p) - pp^t)c$, which may be approximated by $c^t(\mathrm{diag}(\hat{p}) - \hat{p}\hat{p}^t)c$. For a given finite sample size, the coverage of this CI deteriorates as the multinomial parameter p approaches the boundary of the parameter space, the simplex in R^m. We therefore look for a more efficient CI.

We may obtain an exact CI by inverting a hypothesis test.

$T = T(X, p)$ is a function of the data X and a parameter value p. Choices of T discussed below. A level α test of the null that X_0 follows p rejects for large values of T, i.e., $T(X_0, p) \geq t_{p,\alpha}$, where $t_{p,\alpha} = \inf\{t : P(T(X, p) \geq t) \geq \alpha\}, X \sim p$, is the $1-\alpha$ quantile of $T(X, p)$ under p. A level α test that X_0 follows a distribution in the composite null $\theta = \{p \in \Delta^{m-1} : c^t p = \theta\}$ rejects when $\sup_{p:c^t p=\theta} T(X_0, p) - t_{p,\alpha} > 0$. The set of parameters θ at which the test fails to reject,

$$A(X_0) = \{\theta : \sup_{p:c^t p=\theta} T(X_0, p) - t_{p,\alpha} \geq 0\} \tag{1}$$

contains θ_0 with probability $\geq 1 - \alpha$,

$$\begin{aligned} & P(\theta_0 \in A(X_0)) && (2)\\ & = P(\sup_{p:c^t p=\theta_0} T(X_0, p) - t_{p,\alpha} \geq 0\} && (3)\\ & \geq P(T(X_0, p_0) \geq t_{p_0,\alpha}\} && (4)\\ & \geq 1 - \alpha && (5) \end{aligned}$$

((equality when the $1-\alpha$ quantile is unique eg CDF is continuous, when p0 is worse case)) The set ((ref)) may therefore serve as a level $1-\alpha$ CI for θ_0.

$T = T(X, p)$ is a function of the data X and a

chioce of T...

Applying this procedure requires the distribution of the test statistic ($T(X, p)$ at p, which is often unavailable. We approximate, to arbitrary accuracy, by using a sample observed distribution, i.e., sampling $T(X, \theta)$ under p.

This sampling at each p in turn requires a choice of p, and therefore θ values to be selected as candidates. We approximate the supremum in ... by choosing a subset of $\{p : c^t p = \theta\}$, for a subset of the possible values

of $\theta \in [\min_i c_i, \max_i c_i]$. Two further tuning parameters are introduced, controlling the number of θ and the number of p at each θ. ((may vary by θ to reflect prior knowledge)) Obtaining a subset of $\{p : c^t p = \theta\}$ is described below.

This CI is exact, i.e., its mean coverage equals the nominal coverage, subject to provisos:

1. monte carlo error, which may be reduced arbitrarily by increasing the tuning parameters ...

2. the null hypothesis $\theta = \theta_0 = \{p : c^t p = \theta_0\}$ is a composite null hypothesis, so that the test statistic on which the CI is based is conservative. That is, the null consists of multiple distributions and the p-value is the least favorable ((ref supremum above)). This is part of the definition of a p-value and unavoidable due to the equivalence of CIs and hypothesis testing. Difference is the gap between largest and smallest p-values within a θ section, which in turn depends on the choice of test statistic, sample size, number of categories. In simulations below, the effect is to inflate the coverage by $< 1\%$.

3. discreteness. There are m^n possible values for X sampled as multinomial of size n with m categories, so at most m^n possible values for a test statistic T. Often fewer observed values when p is close to the boundary of the simplex and some categories are rarely observed. Therefore 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 that is used will be larger than the nominal level. This issue may be addressed by introducing randomness to the test statistic ((Stevens WL. Fiducial limits of the parameter of a discontinuous distribution. Biometrika 1950; 37: 117–29)), though in practice doing so is "considered unacceptable" ((lehmann romano)), instead using p-values which are among those made available by the data. ((agresti 2003 stat meth for for overview of discreteness leading to conservative CIs for count data.))

((agresti 2003 definition of "exact" is that the CI is based on a test stat the distirbution of which is known exactly and not asymptotically or otherwise approximate. This definition applies here other up to the monte carlo error, which is under the control of the analyst. Under the Agresti definition, an exact CI for count data is typically conservative, as in our case, due to discreteness. A further issue in our case is that the null hypotheses, on which the CI is constructed, is compound.))

1.3 Algorithm

algorithm:

1. select theta values. remark: can take a grid, or sample. can choose to reflect prior knowledge. if only interested in testing the null of a specific theta value, can just take one value.

2. at each theta value:

2a. sample p values in preimage of theta. remark similar to above. dirichlet sparsity/concentration.

2a. at each p value: for empirical cdf of test stat. use to get a p-value at this p.

2b. take max to get a pval associated with this theta. remark: conservative. max of the observations (over time: expectation of the max) being used as an estimate of the max of the expectations. requires coordination between number of smaples in 2b, controlling how close the p-value estimates are to the true pvalue, and the number of p values selected in 2a.

3. $1 - \alpha$ CI for theta can be constructed as those theta for which the associated p-value exceed α.

((maybe mention original algorithm, sampling p first, add sims in appendix))

details for 2a:

describe sampling procedure. The hyperplane $c^t x = 1$ in $\{x \geq 0\} \subset \mathbb{R}^m$, for constant $c \geq 0$, is contained in $[0, c_1^{-1}] \times \ldots \times [0, c_m^{-1}]$. So can use rejection sampling to sample points uniformly on its intersection with the solid simplex. Then transform ... to probability simplex. Acceptance probability $O(1/m!)$ (volume of solid simplex in $\mathbb{R}^m$).

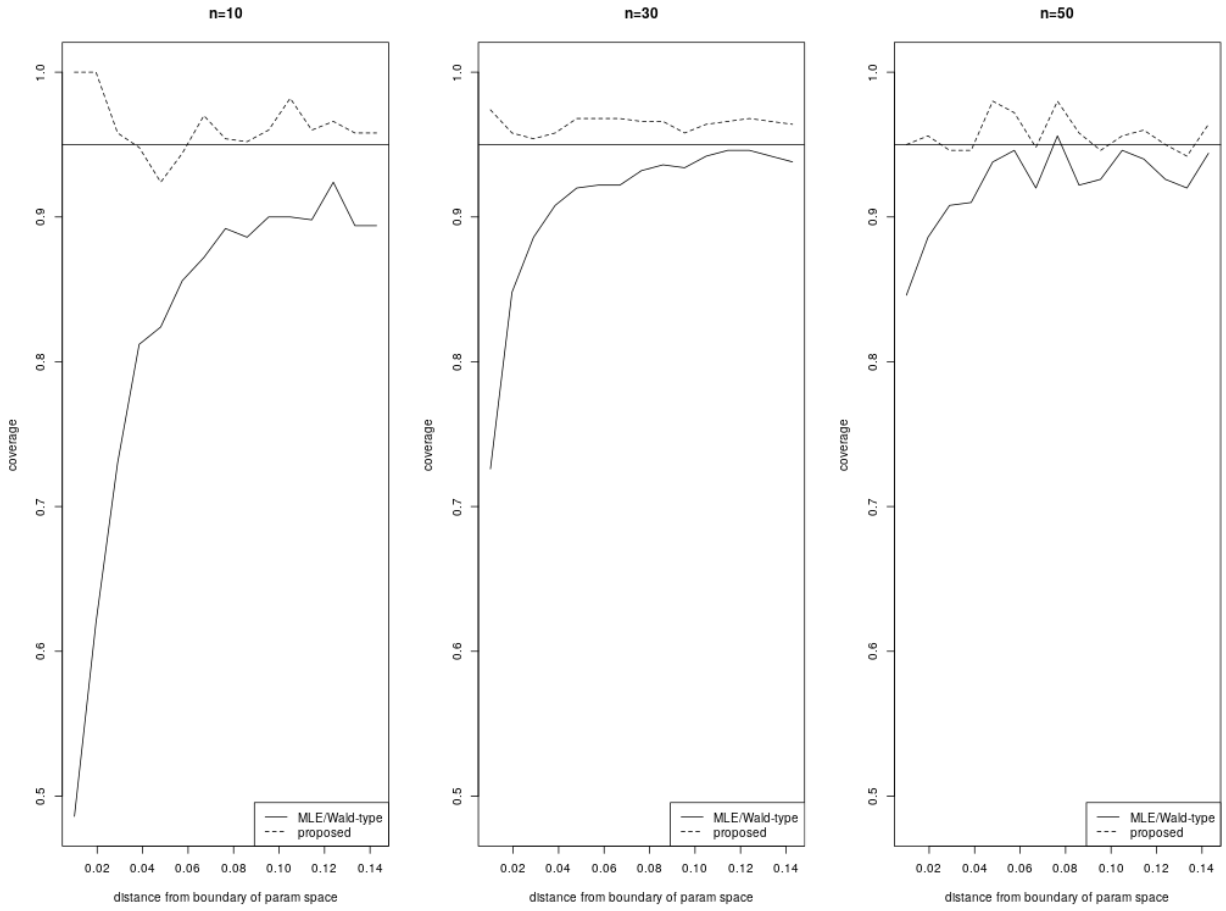
More efficient sampling procedures.

approach 1. fast but not uniform sampling. we want to sample from the intersection of the probability simplex $\mathbb{1}^t x = 1, 0 \leq x \leq 1$, and the hyperplane $c^t x = 1$ for a coefficient vector c. In our application $c = ...$ These points satisfy $c^t x = 1 = \mathbb{1}^t x$ or $(c - \mathbb{1})^t x = 0$ and lie in the unit cube. Therefore we sample u on the orthogonal complement of $c - \mathbb{1}$ and divide u by $\sum_i u_i = \sum_i c_i u_i$. give details. The difficulty is that the central projection, $u \mapsto u/(\sum_i u_i)$ does not preserve uniformity, so sampling u as uniform on delta cross

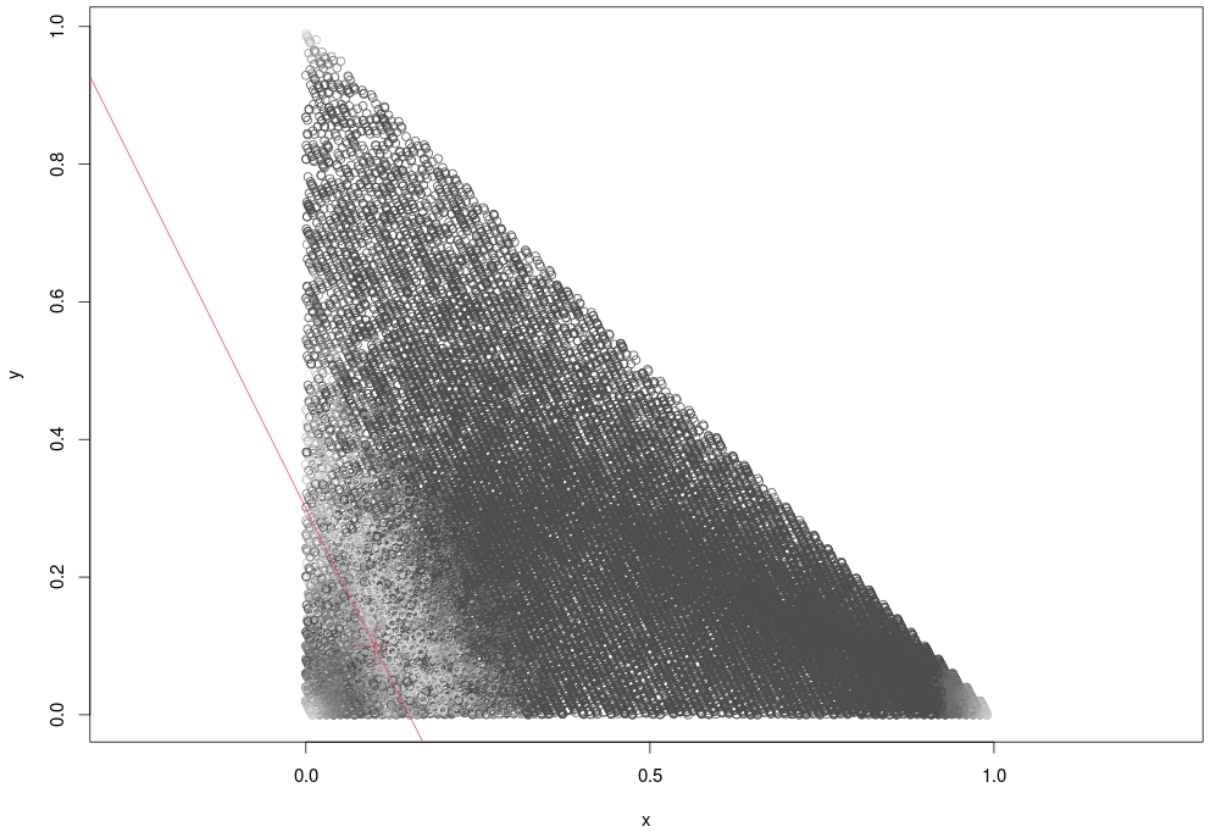
delta will not induce a uniform distribution. seems difficult to even identify the uniform distribution we are want without enumerating the vertices in the intersection of the simplex and linear subspace.

approach 2. slow but uniformly distributed (can get other distributions too). For $c \geq 0, \theta > 0$, the intersection of Δ^{m-1} with $c^t x = \theta$ is a convex polyhedron in R^n_+, i.e., the convex hull of a finite set of vectors $\{v_1, \ldots, v_k\} \subset R^n_+$. These vectors may be enumerated in time .... Therefore sample uniformly on the probability simplex Δ^{k-1} (e.g., normalized exponentials) and apply to the sample the linear transformation mapping the standard basis vector e_i to v_i, $i = 1, \ldots, k$. Since linear transformations preserve uniformity ((e.g., devroye)), the image is uniformly distributed on $\Delta^{m-1} \cap \{c^t x = \theta\}$. Can also use other sampling schemes on the probability simplex to reflect prior knowledge about the location of the true parameter. E.g., Dirichlet.

2 Simulation



give power curves, maybe with different dirichlets



((convert color to bw, set asp ratio to 1))