Statistics and Its Interface Volume 2 (2009) 223–225

Some issues concerning disclosure risk in contingency tables*

Arthur Cohen† and Harold B. Sackrowitz

Inference issues in connection with disclosure risk in contingency tables are considered. For a model which makes no assumptions regarding relationships among cells, Bayes estimators of disclosure risk and Bayes tests for "uniques" are found. Formulas for sample size determination also ensue.

Keywords and phrases: Bayes estimators, Bayes tests, hypergeometric distribution, uniques.

1. INTRODUCTION

Classification of a sample of individuals according to a set of attributes is often times displayed in a contingency table. Such tables can sometimes pose a risk of disclosure in that it may be possible to identify an individual (or small group of individuals) that possess specific attributes. A simple example is as follows:

Suppose a company offers free cholesterol testing for employees. The distribution of those taking the test and found to have high cholesterol is given in Table 1 below.

Suppose these, seemingly anonymous, results are released. If it turns out that only one woman remained in the group hired 3–5 years ago then the personnel manager of the company would be able to identify her and know she had high cholesterol.

A practical consideration for an agency collecting data in contingency table form is whether there is some degree of "disclosure risk" if the data is released. Based on some measures or estimates of this risk the agency may decide not to release the data, at least in its current form.

There have been a variety of measures of disclosure risk and a number of studies concerned with estimating this risk based on a sample contingency table. These include Bethlehem, Keller, and Pannekock (BKP) (1990), Fienberg and Makov (1998), Fienberg, Makov, and Steele (1998), Rinott (2003), Rinott and Shlomo (2007), and Zhang (2005). A variety of models are considered in these references including log linear models, regression models, with a variety of assumptions about sampling frequencies and population frequencies. Often, assumptions are made about population frequencies, thus casting the problem into a Bayesian formulation. Sometimes the priors chosen in the Bayes approach depend on additional parameters that are related to one another. Should the factors of the contingency table be ordered, this sometimes leads to sparseness assumptions or to regression assumptions about the population cell frequencies. The approach here is to assume a realistic hypergeometric distribution model for the sample frequencies. Priors on the population frequencies are chosen so that they do not depend on additional hyperparameters that might be related. Furthermore the priors are permutation invariant so that each population cell frequency has the same marginal distribution.

Table 1. Individuals with high cholesterol

	Years with Company				
	0–1	1–3	3–5	5–7	> 7
Women	4	8	1	15	12
Men	6	7	5	11	9

To describe the approach in this paper it will be helpful to introduce some notations and definitions. Let $\mathbf{f} = \{f_k\}$, $k = 1, \ldots, K$ be sample cell frequencies of an $m - way$ contingency table with K cells. The corresponding population table is $\mathbf{F} = \{F_k\}$. Let $n = \sum_{k=1}^{K} f_k$, and $N = \sum_{k=1}^{K} F_k$ be fixed sample sizes and population sizes respectively. To start we assume nothing special about the categories and we make no assumptions about any structural relationships among the cells of the table. Cell i is called a unique if $f_i = 1$ and $F_i = 1$. A cell i is called a sample unique if $f_i = 1$. A unique represents an individual whose identity can be compromised. One index measuring the total risk in an $m-way$ table is τ, defined as the total number of uniques. Our objectives in this study are to estimate the random quantity τ and also to decide which sample uniques are uniques. We remark that the notion of uniques is also of keen interest in some genetic studies.

For a basic but realistic sampling model (multivariate hypergeometric) we derive the posterior probabilities that $F_i = 1$ given $f_i = 1$ for two different prior distributions. The posterior probabilities are denoted by Q_i, $Q_i = P\{F_i = 1 \mid f_i = 1\}$ can be utilized in three ways. Namely to estimate τ, the disclosure risk, to test hypotheses as to whether the i^{th} sample unique is a unique, and for sample size determination

*Research supported by NSF Grant DMS-0804547 and NSA Grant H-98230-06-0076.

†Corresponding author.

required to support an inference concerned with estimating τ and with testing for a population unique.

In the next section we state the model and derive posterior probabilities that a cell is a unique, given that the cell is a sample unique. From these we estimate τ, test hypotheses about uniques and discuss sample size determination issues.

2. ESTIMATING DISCLOSURE RISK AND TESTING FOR UNIQUENESS

Assume $\mathbf{f}$, a $k \times 1$ vector of sample frequencies, is distributed according to a hypergeometric distribution with parameters $\mathbf{F}$, such that $\sum_{k=1}^{K} f_k = n$, and $\sum_{k=1}^{K} F_k = N$. Thus the joint probability mass function is

$$h(\mathbf{f}) = \prod_{k=1}^{K} \binom{F_k}{f_k} \Big/ \binom{N}{n}. \tag{2.1}$$

If f_i is a sample unique then the conditional distribution of $\mathbf{f}^{(i)} = (f_1, \ldots, f_{i-1}, f_{i+1}, \ldots, f_K)'$ given $f_i = 1$ is

$$h(\mathbf{f}^{(i)} | f_i = 1) = \prod_{\substack{k=1 \\ k \neq i}}^{K} \binom{F_k}{f_k} \Big/ \binom{\sum_{\substack{k=1 \\ k \neq i}}^{K} F_k}{n-1}. \tag{2.2}$$

If $g(\mathbf{F})$ represents a prior distribution of $\mathbf{F}$, then the posterior probability that $F_i = 1 | (f_i = 1, \mathbf{f}^{(i)})$, say Q_i, is

$$Q_i = \sum_{\{F_i = 1, \mathbf{F}^{(i)}\}} \prod_{\substack{k=1 \\ k \neq i}}^{K} \binom{F_k}{f_k} g(F_i = 1, \mathbf{F}^{(i)}) \Big/ \sum_{\{\mathbf{F}\}} \prod_{k=1}^{K} \binom{F_k}{f_k} g(\mathbf{F}), \tag{2.3}$$

where $\mathbf{F}^{(i)} = (F_1, \ldots, F_{i-1}, F_{i+1}, \ldots, F_K)'$.

Now let $R = \binom{N+K-1}{K-1}$ be the number of points in the simplex $\sum_{k=1}^{K} F_k = N$ and let $g(\mathbf{F}) = 1/R$, i.e., g is a uniform prior distribution. Thus our first result is

Theorem 2.1. *The posterior probability that $F_i = 1$ given $\boldsymbol{f}$ with $f_i = 1$ is*

$$Q_i = \frac{(n+K-1)(n+K-2)}{(N+K-1)(N+K-2)}. \tag{2.4}$$

Proof. We utilize a result from Wisniewski (1966) who offered the following probability mass function for positive integers $a_l, l = 1, \ldots, p$.

$$P\{X_l = j_l, l = 1, \ldots, p | \sum_{l=1}^{p} X_l = M\} = \prod_{l=1}^{p} \binom{a_l + j_l - 1}{j_l} \Big/ \binom{\sum a_l + M - 1}{M}, \quad j_l \geq 0, \quad \sum_{l=1}^{p} j_l = M. \tag{2.5}$$

Choose $a_l = f_l + 1$, $j_l = F_l - f_l$ so that $\sum j_l = N - n$. Next note $\binom{F_k}{f_k} = \binom{F_k}{F_k - f_k}$ and apply (2.5) to the denominator of (2.3) to find that the denominator of (2.3) (with $1/R$ canceled out of numerator and denominator) is $\binom{N+K-1}{N-n}$. In the numerator we get the analogue of $\sum j_l$ to be $(N-n)$ once again, while the analogue of $\sum a_l = (n-1) + (K-1)$. Hence the numerator is $\binom{N+K-3}{N-n}$ and (2.4) is established.

Should the prior be the multinominal prior $g(\mathbf{F}) = (N! / \prod_{k=1}^{K} F_k!)(1/K)^N$ we have

Theorem 2.2. *The posterior probability that $F_i = 1$ given $\boldsymbol{f}$ with $f_i = 1$ is*

$$Q_i = [(K-1)/K]^{N-n}. \tag{2.6}$$

Proof. Use (2.5).

Note that neither (2.4) nor (2.6) depends on $\mathbf{f}^{(i)}$. This makes sense since no relationship is assumed to exist among the cells and so every sample unique should be treated in the same way. The probability that a sample unique is also a unique depends only on the population size, sample size, and the number of cells. The probability should be higher as the discrepancy between N and n is smaller and should be higher if K is larger as N, n are fixed.

At this point we let s be the number of sample uniques. Since Q_i is the same for every sample unique, say $Q_i = Q_1$ the Bayes estimator of τ is $\tilde{\tau} = sQ_1$.

One could utilize (2.4) or (2.6) in determining Bayes tests of hypothesis of $H : F_k = 1$ vs $H_A : F_k \neq 1$, when $f_k = 1$.

If losses were c when one decides to accept H when H_A is true, and b when one decides to reject H when H is true then the Bayes test is to reject if $Q_k < c/(b+c)$. For our set up, i.e. no relationship assumed among the cells and for the two priors chosen, it is somewhat reasonable that Q_k does not depend on $\mathbf{f}$, i.e. it is the same for any $\mathbf{f}$.

Note (2.4) or (2.6) can be used to determine sample sizes that would yield posterior probabilities that may be considered, a prior, as desirable for purposes of estimation or testing.

A richer class of priors is available that will yield posteriors that can be calculated. One can focus on a subset of points in the simplex of $\mathbf{F}$ and assign weights of γ to the subset and weights $(1-\gamma)$ to the complementary set. Sometimes such calculations are feasible.

We conclude this section with a numerical example drawn from BKP (1990). The population consists of $N = 46,228$ individuals in a Dutch municipality who are classified according to 4 variables. Namely household composition with 24 categories, age with 14 categories, marital status with 2 categories and sex with 2 categories. The total number of cells $K = 1108 < K' = 24 \times 14 \times 2 \times 2$ since structural zeros are excluded from the analysis but sampling zeros are not. The sample size is $n = 8,399$ and there were 108 sample uniques. Our Bayes estimator of the number of population uniques based

on (2.4) is $\tilde{\tau} = 108Q_1 = (8,399 + 1,108 - 1)(8,399 + 1,108 - 2)/(46,228 + 1,108 - 1)(46,228 + 1,108 - 2) = 4.36$.

Received 5 September 2008

REFERENCES

[1] Bethlehem, J., Keller, W., and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association* **85** 38–45.

[2] Fienberg, S. E., and Makov, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics* **14**(4) 385–397.

[3] Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* **14**(4) 485–502.

[4] Rinott, Y. (2003). On models for statistical disclosure risk estimation. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxemburg*, 275–285.

[5] Rinott, Y. and Shlomo, N. (2006). A smoothing model for sample disclosure risk estimation. *IMS Lecture Notes - Mongraph Series* **54** 161–171.

[6] Wisniewski, T. K. M. (1966). Another statistical solution of a combinatorial problem. *The American Statistician* **20** 25.

[7] Zhang, C.-H. (2005). Estimation of sums of random variables: examples and information bounds. *Ann. Statist.* **33** 2022–2041. MR2211078

Arthur Cohen
Rutgers University
E-mail address: artcohen@rci.rutgers.edu

Harold B. Sackrowitz
Rutgers University