Posterior Distributions for Rare Events in Multivariate Categorical Data

DOUGLAS H. JONES
Management Science and Information Systems
Rutgers Business School—Newark and New Brunswick
Piscataway, NJ 08854
USA

Abstract: This study proposes a method to estimate the posterior distribution of multidimensional categorical data. This methodology enables Bayesian analysis of rare events by borrowing strength from a large database. Once the posterior distributions are profiled, further analysis can be performed and/or decisions made about importance of the occurrence of a particular rare event. For example, the occurrence of a rare event can signal an unusual or undesirable activity in a supply chain and lead to instability in vendors or suppliers and other chain components, possibly leading to the failure of the entire supply chain. Some supply chains are critical for a stable economy and national security, thus early and efficient detection of disruptions of these supply chains are essential.

Keywords: Early Detection, Rare Events, Bayesian Data Analysis, Supply Chain Security, Datamining Techniques, Signal Detection, Active Surveillance Methods.

1. Introduction

With the advent of powerful hardware, interconnectivity, OLAP and of data mining software, many more users are capable of manipulating datasets in ways that were almost impossible before. The availability of huge amounts of data and the increase in accessibility comes together with the promise information for early decision analysis in a dynamic environment.

Microdata are files that consist of individual records that contain values of variables for a single person, a business establishment, or another individual unit. This study considers the joint probability distribution of rare events under full-file analysis and record-level metrics. It is assumed that low cell counts and sampling zeroes correspond to combinations of attributes with a potential of several rare events that could represent some threat to the stability of the supply chain.

A related area is the publishing of data aggregates and the risk of micro-record disclosure. An occurrence of a rare combination of attributes can lead to the identification of individual records; see [3] for a description of the issues. This is the problem of *disclosure risk*. It is a "complementary" problem to ours, in the sense that records in a database under high risk of disclosure are equivalent to low probability rare events. We will consider the methods developed for this problem as background information for our problem. Define the following:

Population unique. A record within a dataset which is unique within the population on a given key [1]

Sample unique. A record within a dataset which is unique within that dataset on a given key [1].

Several methods have been proposed to assess disclosure risks from population uniques in microdata files. These methods vary from sorting files by all attributes in the key fields, using algorithms to find the frequency of small count sets, to data modeling. Most methods provide with an overall measure of disclosure risk for the whole data set.

The existence of sample uniques increases the likelihood of re-identification of individual records; hence the releasing entity should focus on these records. If a subset of variables leads to uniqueness in the population; then, by matching records, an intruder can get access to additional information about the unique individual. There is the possibility of inferring information about population uniques from sample data and from profiles with no

representation in the sample (sampling zeroes). Sample uniqueness does not imply population uniqueness. The existence of sample uniques, which are also population uniques, in data files increases the likelihood of disclosure. Statisticians propose a model to estimate the number of population uniques using sample data [2]. These models provide a general disclosure risk measure for the data set. The problem where, given a dataset, there may be *P*-sets of data that are unique has been addressed [3]. This *P*-set is unique in the set of *N* records in the following two cases: 1) there is but one individual with that profile; 2) there is a small number *k* of individuals, say $k = 2$ or 3, with that specific profile. An approach is aimed at finding rare combinations of any attribute in the dataset [3]. The Special Unique Detection Algorithm (SUDA) was introduced in [5], which determines the uniqueness of a record, within a dataset, based on the uniqueness of a subset of the complete attribute set. The complexity of these types of algorithms is discussed in literature [5, 17]. The algorithm in [17] has a complexity of $O(2^q N)$, where q is the number of fields and N is the number of records.

Another way to detect unique profiles is by cross-classifying all observations in a file using a k-dimensional contingency table, where k is the number of attributes in the file. Cells with a count of one correspond to joint categories with one observation: a unique observation. Cell probabilities can be estimated with loglinear models, as suggested in [6] and [14]. It is possible to infer population uniqueness from sample zeros and there is risk of disclosure related to small cell counts larger than one [6]. It is proposed the use a Bayesian framework as a natural way to find population uniques via an approach called model averaging [6]. Small cell counts carry small amounts of information and it is difficult to conclude based on this amount of information. The problem of finding joint events with small frequencies in datasets and the need for a strong assumption on the distribution of the cell counts in the population is presented in [15]. A Poisson-Gamma model is proposed in [2]. However, the Poisson-Gamma model and related models tend to underestimate the number of uniques in the population [8]. Others use the multinomial-Dirichlet model [9], [15]. Some researches feel fitting a loglinear model to contingency tables is computationally expensive and propose to fit a Lancaster-type additive model of interaction terms for cell probabilities of contingency tables [15]. A disadvantage of the additive probability model is that it assumes no *structural zeros*. A full Bayesian approach to evaluate the number of population uniques, in terms of the posterior probability distribution of population uniqueness was proposed in [13]. Super population models provide an overall estimate of the number of population uniques in the sample but they do not suggest which records have higher identification risk than others.

2. Methodology

The point estimates for the cell counts assist in identifying profiles with low cell counts. From the posterior distribution, cells with expected low cell counts can be identified. Define the following:

Low cell count. A cell count such that ($m_\theta \leq \delta$), where δ is a threshold established by the user and m_θ is the expected count for cell θ.

Potential disclosure risk set (ϕ). The set constituted by those records with the profile corresponding to the joint attributes described by low cell counts and sampling zeroes.

It is assumed that combinations of attributes with a potential of several rare events that could represent some threat to the stability of the supply chain correspond low cell counts and sampling zeroes correspond.

To model the data, the first step is to categorize the observations by all attributes, treating the dataset as a multi-way contingency table and fitting a complete, hierarchical log-linear model to the contingency table. A discrete data model is fitted using the Bayesian iterative proportional fitting (BIPF) [7] to obtain the joint posterior distribution of all cell counts. This approach allows the researchers to sample the posterior distributions of the elementary cells of the multidimensional contingency table to explore the following: 1) the posterior mean, median and mode of the distribution for each elementary cell count; 2) the posterior variation of the distribution for each elementary cell count; and 3) the posterior distribution of the minimum values for each elementary cell count.

A prior Dirichlet distribution for the prior is assumed. Sampling zeros are treated as suggested in literature that deals with sparseness and model selection [12]. The methods are: 1) adding a very small constant to each cell (i.e., 10^{-8}); 2) adding one count to each cell (the Lidstone correction [11]); or

3) adding the *minimax* amount, $N^{1/2}/K$, to each cell where N is the total count (the Trybula correction [16]).

A standard rectangular microdata file is used. The population (N = 2,000,000) for this experiment is known and distributed Dirichlet. The samples sizes used are 1,000 and 10,000 for sampling ratios of 0.0005 and 0.005 respectively; which are smaller sampling ratios than what is mentioned in literature [20]. Ideally, there is a value δ such that $\min(Np) \leq \delta$ indicates a risk of disclosure; where Np is the estimated cell count and p is the posterior point estimate for the proportion of observations in the population that corresponds to profile θ. Arbitrarily, $\delta = 50$. A four-step experiment was performed: 1) simulate a population that has a Dirichlet distribution; 2) sample from the population; 3) fit a complete log-linear model using BIPF (the burn-in time is 500 iterations and 10,000 iterations post-convergence.); 4) described the posterior distribution of $\min(Np)$; 5) assess the effect of flattening constants based on the coefficient of skewness.

The selection of a flattening constant has a repercussion in the resulting posterior distribution. Two methods to treat sampling zeros will be used: Lindstone and Trybula. Types of cells, those for profiles with small probability of occurrence and those for profiles with very small probability of occurrence (which could represent a rare event), were selected to observe the effect of the flattening constants on the posterior distribution of cell counts.

Let ε be the set of elementary cells corresponding sampling ones but relatively large number of observations in the population; thus representing a rare event. Let ϕ be the set of elementary cells corresponding to sampling ones and a relatively small number of observations in the population; thus representing rare events. It is of interest to compare the posterior distribution of $\min_{\varepsilon}(Np)$ versus the posterior distribution of $\min_{\phi}(N\pi)$, where π are the proportion estimates obtained from the sample data, and $X \sim Dirichlet(\beta)$ are the observations for the underlying population..

The objective is to use the posterior distribution of $\min(Np)$ to detect rare events. This decision is presented as a hypothesis test. Defining f_θ as the sample frequency for cell θ, the objective is to test the following hypothesis:

$H_0 : f_\theta \geq \delta$ (no threat concern for profile θ)

$H_1 : f_\theta < \delta$ (threat concern for profile θ)

Where δ is a threshold which represents the largest cell count that is considered a *threat*. Rejecting the null hypothesis would imply that it that cell with profileθ represents an unusual activity and a potential threat.

3. Results

Selection of a Flattening Constant

The BIPF was used to estimate the cell counts using two types of flattening constants: the Lindstone method and the Trybula method. It was found that the posterior distributions using the Trybula method tend to be more positively skewed than the distributions using the Lindstone method. There seem to be a sharp difference in skewness between those distributions corresponding to profiles with a high disclosure risk and those corresponding to profiles with lower disclosure risk. **Error! Reference source not found.** show the coefficients of skewness for the posterior distributions estimated using the Lindstone flattening constant and those obtained using the Trybula flattening constant. The first seven rows correspond to high-risk profile cells. The bottom five rows correspond to low-risk cells.

Table 1: Coefficient of Skewness of Posterior Distributions under Lindstone (left) and Trybula (right)

Cell	Skewness
82	0.9825
79	1.8668
160	1.4519
185	1.5441
197	0.9064
300	1.8824
297	1.6456
92	1.4073
22	1.2325
211	1.1727
354	1.2767
374	0.6747

Cell	Skewness
82	5.0083
79	5.5941
160	6.5518
185	14.0766
197	4.6204
300	6.8075
297	4.1609
92	1.6387
22	1.0729
211	1.4090
354	1.2328
374	1.9075

In this example, the Trybula flattening constant (on the right) seems to provide the information necessary to discriminate between low count cells with medium and very small probabilities of occurrence.

Discriminating Between Lower Risk and Higher Risk

The Trybula method was used to generate posterior distributions for $\min_{\varepsilon}(Np)$ and $\min_{\phi}(N\hat{\pi})$. Figure 1b shows a pronounced positive skewness for the posterior distribution of those profiles representing low and high probability of rare events, $\min_{\phi}(N\hat{\pi})$ when compared against the posterior distribution in Figure 1a of profiles that represent a medium rare event, $\min_{\varepsilon}(Np)$.

Figure 1a: Posterior Distributions: $\min_{\varepsilon}(Np)$

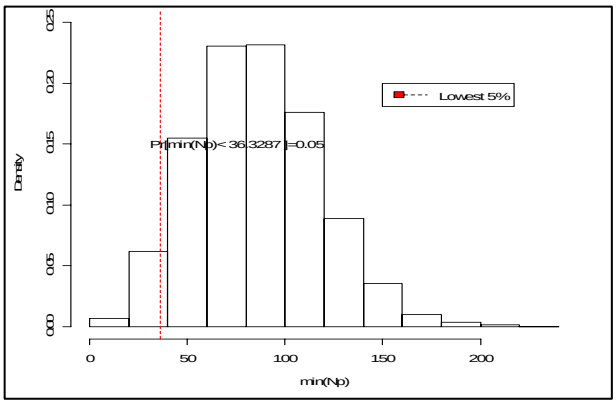
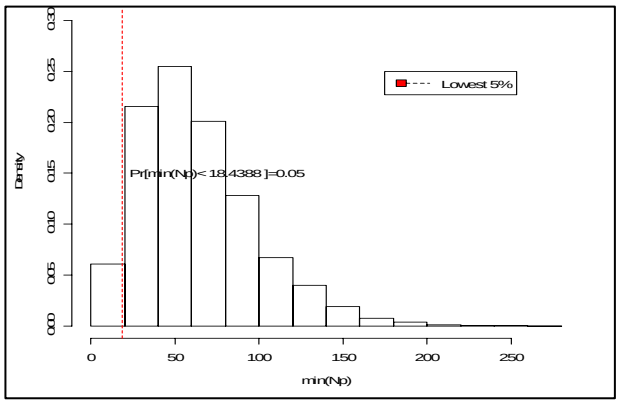


Figure 2b: Posterior Distributions: and $\min_{\phi}(N\hat{\pi})$



Sample of low count cells (including sampling zeros) we can say that, there is a probability of 0.05 of seeing a population count less than 18, given a low probability of a rare event; while there is a probability of 0.05 of seeing a population count less than 36, given high probability of a rare event. Remember that *low and high probabilities* were determined arbitrarily.

4. Conclusion

A method to discover profiles of rare events is being explored. This method uses sample data (including sample zeros by means of a flattening constant), and a Bayesian method, to incorporate a prior distribution for the population into the estimation of the distribution of counts of the population. The method that we explore is intended to identify profiles that are not very frequent in the population using only the information that is gathered through a simple random sample. The posterior distributions for minimum sample counts (or zero) with high probability of a rare event have been compared to the posterior distributions for minimum sample counts (or zero) with low probability of rare event. It was found that this proposed method has potential for identifying and distinguishing between rare events; specifically, it was found that the posterior distribution of the minimum counts for low-probability cells have a skewness that is more positive than the skewness of the posterior distribution of the minimum counts for the low-probability cells; the percentiles of the posterior distribution are distinct enough to distinguish between different profiles of rare events. The objective of a future study will be to develop a single risk measure or hypothesis test to distinguish a high-probability profile from a low-probability profile by means of a metric over probability distribution; this will facilitate risk and threat assessment in automated systems.

Acknowledgement: This research is based on joint research with Dr. Francis A. Méndez Mediavilla, Texas State University—San Marco.

References:

[1] *Glossary of Statistical Disclosure Control*, 2005 October [cited 2007 August 10, 2007]; Available from: http://stats.oecd.org/glossary/detail.asp?ID=6982.

[2] Bethlehem, J.G., W.J. Keller, and J. Pannekoek, Disclosure Control of Microdata, *Journal of the American Statistical Association*, 1990, 85(409).

[3] Dalenius, T., Finding a Needle in a Haystack or Identifying Anonymous Census Records. *Journal of Official Statistics*, 1986, 2(3): p. 329-336.

[4] Doyle, P., et al., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 2001, North-Holland, New York.

[5] Elliot, M.J., A.M. Manning, and R.W. Ford, A Computational Algorithm for Handling the Special Uniques Problem, To appear in *Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

[6] Fienberg, S. and U. Makov, Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data, *Journal of Official Statistics*, 1998, 14(4).

[7] Gelman, A., J. B Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, New York: Chapman & Hall/CRC, 1995.

[8] Hoshino, N., Applying Pitman's sampling formula to microdata disclosure risk assessment. 2000, University of Tokyo: Tokyo, Japan.

[9] Hoshino, N. and A. Takemura, On the relation between logarithmic series model and other superpopulation models useful for microdata disclosure risk assessment, 1998, Graduate School of Economics, University of Tokyo: Tokyo, Japan, page 9.

[10] Manning, A.M. and D.J. Haglin, A new algorithm for finding minimal sample uniques for use in statistical disclosure assessment, *Fifth IEEE International Conference on Data Mining*, 2005. Houston, Texas: IEEE.

[11] Lindstone, G., Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities, *Transactions of the Faculty of Actuaries,* 1920, p. 182-192.

[12] Méndez Mediavilla, F. A., Using discrete multivariate MCMC Bayesian methods for change detection and disclosure control, Rutgers University. Ph.D. Thesis, Rutgers—The State University of New Jersey, 2005.

[13] Omori, Y., Measuring identification disclosure risk for categorical microdata by posterior population uniqueness, *Proceedings of the conference Statistical data protection*, 1999, Lisbon: Office for Official Publications of the European Communities.

[14] Skinner, C.J. and D.J. Holmes, Estimating the Re-identification Risk per Record in Microdata, *Journal of Official Statistics*, 1998, 14(4): p. 361-372.

[15] Takemura, A., Evaluation of per-record identification risk by additive modeling of interaction for contingency table cell probabilities, 2001a, Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo: Tokyo.

[16] Trybula, S., Some Problems of Simultaneous Minimax Estimation, *Annals of Mathematical Statistics*, 1958, 29: p. 245-253.

[17] Winburn, M. and A. Wheeler, Finding Unique Records in Unknown Data, In *International Conference of Integration of Knowledge Intensive Multi-Agent Systems*, 2003, Boston, USA: IEEE