Approximating the full likelihood for marginal $2 \times J$ contingency tables and case-control data.

Markus Chagas Stein*
Department of Statistics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil; and
Department of Statistics, University of Auckland, Auckland, New Zealand - m.stein@auckland.ac.nz

Chris Wild
Department of Statistics, University of Auckland, Auckland, New Zealand - c.wild@auckland.ac.nz

Alastair Scott
Department of Statistics, University of Auckland, Auckland, New Zealand - a.scott@auckland.ac.nz

Abstract

In this work, we consider likelihood-based approaches for constrained two-way contingency tables. Very often only the marginal distributions of variables of interest are available and inferences for unit-record effects are subject to a range of biases. Examples of such problem appear in so-called "ecological inference" and disclosure limitation. In this case, the only reliable way of reducing bias and the impact of uncheckable assumptions is supplementing the marginal distributions by other, e.g. individual-level, information. Our interest is in case-control sampling as auxiliary information to the group-level distributions. Maximum likelihood estimation can be a very challenging task, because its calculations require the enumeration of all tables consistent with the observed data. However, good approximations to the full likelihood can be carried out by sampling possible tables. We compare inferences via true likelihood in $2 \times J$ tables and estimated likelihoods, through uniform sampling methods and an informative sampling scheme based on independent binomial distributions. Preliminary results show that large improvements in computational time can be obtained for minimal losses in efficiency of estimation.

Keywords: Ecological inference; Contingency tables; Case-control sampling; Maximum likelihood estimation.

1. Introduction

Inference about the association between two or more categorical variables with only the marginal distributions available is often called ecological inference (Wakefield, 2004). Such aggregated data usually arises because census and survey organisations often only publish data as margins and not as unit-record data, but researchers often still want to make inferences about individual-level effects. In many cases, however, additional information beyond the marginals may be able to be obtained. Here we consider marginal totals supplemented by a case-control subsample (e.g. Haneuse and Wakefield, 2008).

The same data structures arise when researchers who have taken case-control samples supplement their research data using available population data. Efficiency in analysis of retrospective samples can often be improved substantially where relevant marginal population information can be obtained. Non-nested multiphase sampling can arise, for example, when multiple subsamples of a cohort are taken either as a result of separate research projects over time, or as part of a deliberate sampling effort targeting multiple variables. It can also arise when full-cohort information is available on sets of low-dimensional table-margins, say from census data or other sources.

Those problems consist of making inference for the joint probability distribution of a partially observed contingency table, using information about population marginal distributions and also a detailed subsample information. One of the main issues is that likelihood calculations involve sums of terms from all possible population tables consistent with given the observed margins and the subsample data. This becomes infeasible as the tables and the counts get larger, due to the huge number of possible tables. Additionally, many tables do not contribute significantly to the full likelihood. We explore the use of approximations obtained by sampling the set of possible tables. Methods include Markov chains and forms of importance sampling.

Sampling methods for two-way contingency tables with given constraints have been studied extensively, some of the first algorithms arise as alternatives to the classical Fisher's exact test (Mehta and Patel (1980)). There has been an increased demand for sampling contingency tables methods as applied problems are involving more challenging tables; some recent main methods include Markov chains (many of these are based on the seminal work of Diaconis and Sturmfels, 1998) and sequential importance sampling methods (Dinwoodie and Chen, 2011).

In the ecological inference context, Wakefield (2004) reviews several methods to perform inference in 2×2 tables for which only the population margins are observed, and states that the solution to the ecological inference problem is to supplement the marginal population data with representative and accurate survey sample information on individual level. Haneuse and Wakefield (2008) show a hybrid design in which observed population margins are supplemented by case-control sample data and focus on maximum likelihood (ML) estimation. They discuss connections between their method and two-phase sampling (Scott and Wild, 1997; Breslow and Holubkov, 1997), and pointed out that marginal exposure data are not used in the two-phase scheme, but their approach needs full enumeration of possible tables. Wakefield et al. (2011) presented a bayesian approach that uses Markov basis for generation of allowable tables, also only for 2×2 tables.

We are interested in inference about constrained contingency tables, where you have marginal population distribution supplemented by sample information about the interior of the table, specially for binary outcomes and categorical/discrete exposures. Specifically our interest is in likelihood calculation and approximations to this.

2. Likelihood function

We consider a $2 \times J$ table with a binary response variable (outcome), Y, and a $J-$level categorical explanatory variable (exposure), X. Let N_{ij} be the total number of individuals in a finite population (or cohort) with $(Y = i, X = j)$, and $\boldsymbol{N_{yx}} = \{N_{ij}\}$ the set of all entries in this population. Also suppose that only the margins $\boldsymbol{N_{y+}} = (N_{0+}, N_{1+})$ and $\boldsymbol{N_{+x}} = (N_{+1}, \ldots, N_{+J})$ are known in the population (aggregated level).

In addition to the marginal population distribution we observe complete information from a case-control subsample, where we take n_{0+} controls and n_{1+} cases and observe their X values, denote $\boldsymbol{n_{y+}} = (n_{0+}, n_{1+})$. This gives us a set of internal counts in the sample table, which we denote $\boldsymbol{n_{yx}}$. Table 1 shows the data strucure described.

Table 1: Population margins and case-control subsample.

Population				Case-control subsample			
X	$Y = 0$	$Y = 1$	N_{+x}	X	$Y = 0$	$Y = 1$	n_{+x}
1			N_{+1}	1	n_{01}	n_{11}	
2			N_{+2}	2	n_{02}	n_{12}	
$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$	
J			N_{+J}	J	n_{0J}	n_{1J}	
N_{y+}	N_{0+}	N_{1+}	N	n_{y+}	n_{0+}	n_{1+}	n

Likelihood construction

We will assume that if the internal cells $\boldsymbol{N_{yx}} = \{N_{ij}\}$ of the population table were observed, then conditional on the population total N, they would follow a multinomial distribution. In the case-control table, the number of controls and cases at each level follow two independent hypergeometric distributions. Also we have constraints $n_{1j} \leq N_{1j} \leq N_1 - n_{1+} + n_{1j}$ (corresponding constraints of the controls are satified automatically).

Our interest is in models for the partially observed population table. The subsample table provides information about the unobserved internal structure of the population table. We first suppose we could observe the interior of population table. Then the observed data would consist of $(\boldsymbol{N_{yx}}, \boldsymbol{n_{yx}})$, and conditioning on the population size, N, the probability of the observed data is

$$
\begin{aligned}
P\left(\boldsymbol{N_{yx}}, \boldsymbol{n_{yx}} \mid N\right) &= P\left(\boldsymbol{N_{yx}} \mid N; \boldsymbol{\theta}\right) \times P\left(\boldsymbol{n_{y+}} \mid \boldsymbol{N_{yx}}; \boldsymbol{\delta}\right) \times P\left(\boldsymbol{n_{yx}} \mid \boldsymbol{N_{yx}}, \boldsymbol{n_{y+}}\right) \\
&= P\left(\boldsymbol{N_{yx}} \mid N; \boldsymbol{\theta}\right) \times P\left(\boldsymbol{n_{y+}} \mid \boldsymbol{N_{y+}}; \boldsymbol{\delta}\right) \times P\left(\boldsymbol{n_{yx}} \mid \boldsymbol{N_{0x}}, \boldsymbol{N_{1x}}, \boldsymbol{n_{y+}}\right) \qquad (1)
\end{aligned}
$$

Here we assume that choices for n_{0+} and n_{1+} can depend only on population numbers of cases and controls, and the choice mechanism, (parameterized by $\boldsymbol{\delta}$) is unrelated to the model of interest ($\boldsymbol{\theta}$). However we do not observe the internal cells for the population table, $\boldsymbol{N_{yx}}$, we only observe the population margins. Therefore the *full likelihood* function is given by summing the individual probabilities (1) over all population tables consistent with the observed data. Let R^* denote the set of all possible tables and S denote the number of tables in R^*. Then, using (1)

$$
\begin{aligned}
P\left(\boldsymbol{N_{y+}}, \boldsymbol{N_{+x}}, \boldsymbol{n_{yx}} \mid N; \boldsymbol{\theta}\right) &= \sum_{R^*} P\left(\boldsymbol{N_{yx}}, \boldsymbol{n_{yx}} \mid N\right) \\
&\propto \sum_{R^*} P\left(\boldsymbol{N_{yx}} \mid N; \boldsymbol{\theta}\right) \times P\left(\boldsymbol{n_{yx}} \mid \boldsymbol{N_{0x}}, \boldsymbol{N_{1x}}, \boldsymbol{n_{y+}}\right) \\
&= \sum_{R^*} \left(\frac{N!}{N_{00}! \ldots N_{1J}!} \pi_{00}^{N_{00}} \ldots \pi_{1J}^{N_{1J}} \right) \times \left[\frac{\binom{N_{01}}{n_{01}} \ldots \binom{N_{0J}}{n_{0J}}}{\binom{N_{0+}}{n_{0+}}} \frac{\binom{N_{11}}{n_{11}} \ldots \binom{N_{1J}}{n_{1J}}}{\binom{N_{1+}}{n_{1+}}} \right] \\
&= \sum_{k=1}^{S} \frac{N!}{\binom{N_{0+}}{n_{0+}} \binom{N_{1+}}{n_{1+}}} \prod_{j=1}^{J} \left[\binom{N_{+j}}{N_{k1j}} \pi_{1|j}^{N_{k1j}} \pi_{0|j}^{N_{k0j}} \right] \left(\frac{\pi_j^{N+j}}{N_{+j}!} \right) \left[\binom{N_{k0j}}{n_{0j}} \binom{N_{k1j}}{n_{1j}} \right] \\
&= \sum_{k=1}^{S} L_k(\boldsymbol{\theta}) \, . \qquad (2)
\end{aligned}
$$

Here $\pi_{ij} = P\left(\boldsymbol{Y} = i, \boldsymbol{X} = j\right)$, $\pi_{i|j} = P\left(Y = i \mid X = j\right)$, $\pi_j = P\left(X = j\right)$, N_{kij} is the entry of N_{ij} in the k-th table, and L_k's will be referred to as the *individual likelihoods*. If the parameters of interest relate only to the conditional probabilities $\pi_{i|j}$ (as in logistic models for $P\left(Y = i \mid X = j\right)$), then the terms invovling π_j are orthogonal and can be omitted from the likelihood. In our illustrations we use the linear logistic model ($\pi_{1|j}(\boldsymbol{\theta}) = \frac{e^{\theta_0 + \theta_1 \times j}}{1 + e^{\theta_0 + \theta_1 \times j}}$).

3. Approximating the likelihood

Although (2) looks relatively simple its calculation becomes difficult unless N is small because the number of tables in R^* becomes extremely large extremely quickly. The cardinality of R^* is $(N_{1+} - n_{1+} + 1)^{J-1}$, e.g. for a population number of cases $N_{1+} = 500$, a case sample size $n_{1+} = 50$, and $J = 3$, the number of possible tables is already large as 203,401. However, if we take larger values and set $N_{1+} = 10000$, a case sample size $n_{1+} = 1000$, and $J = 10$, there are 3.9×10^{35} possible tables! So we want to get good approximations to the *full likelihood* using a very much smaller number of tables.

Uniform Sampling Approach

If we look at the form of the likelihood it is proportional to $\sum_{k=1}^{S} L_k / S$, which is the population mean over the $\boldsymbol{N_{yx}}$ tables. This suggests approximating the likelihood function using a sample mean where we sample s tables from the the R^* set.

Simple Random Sampling

In our first approach, we draw a simple random sample from the R^* set. This does not really help much as would need to enumerate all tables in R^*, which is unrealistic for large S. But we will keep it as a gold standard for comparison purposes.

Random Walk - Markov Chain

Instead of sampling the set of allowable tables directly, here we use a Markov Chain approach due to Diaconis and Sturmfels(1998). This is a clever method which does not require enumeration of R^* set. We only need to have an initial table and generate a chain by choosing a pair of rows and adding $\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ or $\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ to the selected cells with probability 1/2 each. This generates tables that have the same margins (tables that do not obey the other constraints $N_{ij} \geq n_{ij}$ are eliminated later). One disadvantage is that to obtain convergence of the chain we need to generate a large number of tables to allow for burn-in and thinning.

Importance Sampling Strategy

The large number of tables going into the calculation is one factor. Another is the fact that very few of these tables contribute anything much to the likelihood.

We illustrate using the data in Haneuse and Wakefield (2008). This data consists of a 2×2 table with population margins $\boldsymbol{N_{y+}} = (39875, 125)$ and $\boldsymbol{N_{+x}} = (20000, 20000)$, and case subsample $\boldsymbol{n_{1+}} = (15, 35)$. The range of N_{11} is from 35 to 110. We have 76 possible tables and each of them corresponds to an *individual likelihood* L_k. We plot the 76 L_ks against N_{11} in Figure 1. Approximately 50 of the L_k's are essentially zero.

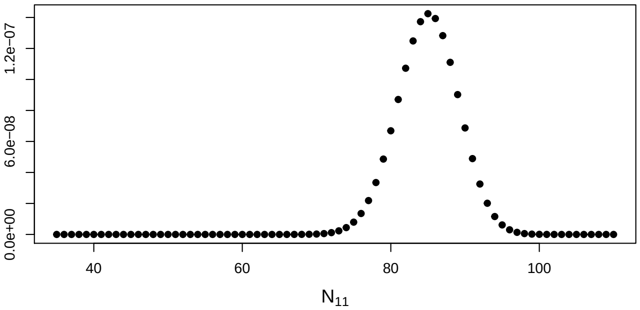

Figure 1: Plot of individual likelihood functions for data in Haneuse and Wakefield (2008).

This suggests that instead of sampling the tables uniformly we should use some form of informative sampling that oversamples tables that makes larger contributions. We draw independent binomial sampling for each row, to do so

1. Estimate the conditional probabilities, $\hat{\pi}_{1|j}$, for $j = 1, \ldots, J-1$;
2. Since $N_{1j} \mid N_{+j} \sim \mathit{Binomial}(N_{+j}, \pi_{1|j})$, we generate each N_{1j} from $\mathit{Binomial}(N_{+j}, \hat{\pi}_{1|j})$;, for $j = 1, \ldots, J-1$; (all other counts can br obtained by subtraction)
3. We discard tables that do not satisfy the constraints (we have strategies for reducing the number of discarded tables), so s^* is the current number of tables left;

Because we are using unequal probability of selections we have to estimate the likelihood, which is a population mean, by a weighted average, $\widehat{L} = \sum_{k=1}^{s^*} \frac{L_k(\boldsymbol{\theta}) w_k}{s^*}$, where

$$w_k^{-1} = P\left\{\text{selecting ``table } k\text{''}\right\} \propto \prod_{j=1}^{J-1} \left[\binom{N_{+j}}{N_{k1j}} \hat{\pi}_{1|j}^{N_{k1j}} \hat{\pi}_{0|j}^{N_{k0j}} \right].$$

Inference from the approximated likelihoods

Maximum likelihood theory has to be adjusted to handle the fact that we are working with an aproximattion, not the likelihood function itself. Our function is like a Monte Carlo maximum likelihood, which was discussed by Geyer and Thompson (1992).

Considering $\pi_{1|j}(\boldsymbol{\theta}) = P(Y = 1 \mid X = j; \boldsymbol{\theta})$, we have $\widehat{L}(\boldsymbol{\theta}) = \sum_{k=1}^{s} L_k(\boldsymbol{\theta}) w_k / s$ where

$$L_k(\boldsymbol{\theta}) \quad = \quad \frac{N!}{\binom{N_{0+}}{n_{0+}}\binom{N_{1+}}{n_{1+}}} \prod_{j=1}^{J} \left[\binom{N_{+j}}{N_{k1j}} \pi_{1|j}(\boldsymbol{\theta})^{N_{k1j}} \pi_{0|j}(\boldsymbol{\theta})^{N_{k0j}} \right] \left(\frac{\pi_j^{N+j}}{N_{+j}!} \right) \left[\binom{N_{k0j}}{n_{0j}} \binom{N_{k1j}}{n_{1j}} \right].$$

So the estimated log-likelihood function, $\widehat{\ell}(\boldsymbol{\theta})$, can be expressed as

$$\widehat{\ell}(\boldsymbol{\theta}) = \log \widehat{L}(\boldsymbol{\theta}) = \log \left(\sum_{k=1}^{s} L_k(\boldsymbol{\theta}) w_k \right) - \log s.$$

The score function is:

$$\widehat{\boldsymbol{U}}(\boldsymbol{\theta}) = \frac{\partial \widehat{\ell}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\sum_{k=1}^{s} \frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} w_k}{\sum_{k=1}^{s} L_k(\boldsymbol{\theta}) w_k}.$$

The information matrix is:

$$\widehat{\boldsymbol{I}}(\boldsymbol{\theta}) = E\left(-\frac{\partial \widehat{\boldsymbol{U}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \right) = \left(\frac{\sum_{k=1}^{s} \frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} w_k}{\sum_{k=1}^{s} L_k(\boldsymbol{\theta}) w_k} \right) \left(\frac{\sum_{k=1}^{s} \frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} w_k}{\sum_{k=1}^{s} L_k(\boldsymbol{\theta}) w_k} \right)^\top - \frac{\sum_{k=1}^{s} \frac{\partial^2 L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} w_k}{\sum_{k=1}^{s} L_k(\boldsymbol{\theta}) w_k}.$$

From standard results on ratio estimation, conditional on full data, $\widehat{\boldsymbol{U}}(\boldsymbol{\theta})$ is asymptotically normally distributed with $\sqrt{s}(\widehat{\boldsymbol{U}}(\boldsymbol{\theta}) - \boldsymbol{U}(\boldsymbol{\theta})) \sim Normal(\mathbf{0}, V = Cov(\widehat{\boldsymbol{U}}(\boldsymbol{\theta})))$. Here V can be obtained using the results of Cochran (1977; p. 155).

4. Performance

Over the next few months we will be running large-scale simulation studies, and also applying the result to real problems. Here we report on some preliminary observations. These are based on a linear logistic regression model for 2 × 3 tables, with known parameters $\boldsymbol{\theta} = (\alpha, \beta) = (4.0, 0.6)$. We set $N_{+j} = 1000$ for every j, and the sample sizes for cases and controls are both 40. The number of generated tables are $s = (50, 100, 1000, 10000)$ tables, with 1000 simulations each.

In the first set of simulations we fixed the population margins and the case-control subsample, so we can compare the accuracy of the likelihood approximations for β estimation due only to the sampling scheme. The boxplots (Figure 2) show that already for the smaller number of generated table binomial importance sampling tends to be closer to the true maximum likelihood estimate (dashed line), which is obtained from the full likelihood. However markov chain has largest variances, it seems to work similar to the gold standard. We can see a bias with respect to the true value, but our target is the full likelihood.

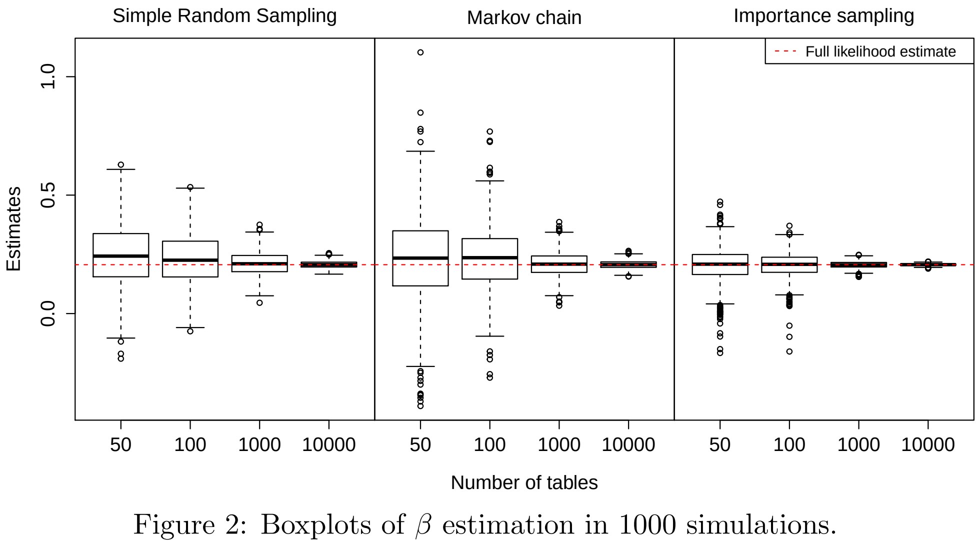

Figure 2: Boxplots of β estimation in 1000 simulations.

We also recorded processing times (Table 2), here by generating different population margins and sampling information. The timings are not giving the real picture because you actually need a lot more tables with Markov Chain to get equivalent precision, although the importance sampling scheme has the smallest times. Table 2 shows the coverage probabilities of 95% confidence intervals as well. The results suggest an advantage in using the importance sampling approach except for a large number of tables.

Table 2: Timings and 95% coverage for β estimates, in 1000 simulations.

	Times (seconds)				Coverage (%)			
Sampling method	50	100	1000	10000	50	100	1000	10000
Simple random sampling	800.4	807.2	904.4	1810.3	62.7	72.1	89.7	94.8
Markov chain	52.4	60.9	213.4	1609.5	55.3	66.2	90.0	94.8
Importance sampling	21.3	27.0	122.1	1032.3	84.0	87.3	91.9	94.3

5. Conclusions

This work consists of an early stage research on the proposed problem. The findings until here suggest that using an informative sampling we can have good approximations for a very small amount of generated tables. It will serve as a base for future studies involving more complex table structures. In addition to likelihood based approaches we are going to explore the efficiency of weighted estimating equations with calibration. Some implications of this research is the possibility of application in ecological inference problems, and official statistics/disclosure limitation as well.

References

Breslow, N. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. Journal of the Royal Statistical Society, Series B 59, 447461.

Cochran, W.G. (1997). Sampling techniques. John Wiley and Sons Inc. New York, Toronto.

Diaconis, P., and Sturmfels, B. (1998), Algebraic Algorithms for Sampling from Conditional Distributions, The Annals of Statistics, 26, 363397.

Dinwoodie, I. H. and Chen, Y. (2011). Sampling Large Tables with Constraints. Statistica Sinica, 21, 1591-1609.

Geyer, C. J., and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. Journal of the Royal Statistical Society, Series B, 54, 657699.

Haneuse, S. J.-P. A. and Wakefield, A. J. C. (2008), The combination of ecological and casecontrol data. Journal of the Royal Statistical Society: Series B, 70: 7393.

Mehta, C.R., Patel, N.R. (1980). A network algorithm for the exact treatment of the $2 \times k$ contingency table. Comm. Statist. Simulation Comput. 9(6), 649664.

Scott, A. and Wild, C. (1997). Fitting regression models to casecontrol data by maximum likelihood. Biometrika 51, 5471.

Wakefield, J. C. Ecological inference for 2×2 tables (2004). Journal of the Royal Statistical Society, Series A, 167: 385-445.

Wakefield J, Haneuse S, Dobra A, Teeple E. (2011) Bayes computation for ecological inference. Statistics in Medicine, 30: 13811396.