Counting Tables using the Double Saddlepoint Approximation

Vadim Zipunnikov[*], James G. Booth[†]and Ruriko Yoshida[‡]

November 2007

Abstract

A double saddlepoint approximation is proposed for the number of contingency tables with counts satisfying certain linear constraints. Computation of the approximation involves fitting a generalized linear model for geometric responses which can be accomplished almost instantaneously using the iterated weighted least squares algorithm. The approximation is far superior to other analytical approximations that have been proposed, and is shown to be highly accurate in a range of examples, including some for which analytical approximations were previously unavailable. A similar approximation is proposed for tables consisting of only zeros and ones based on a logistic regression model. A higher-order adjustment to the basic double saddlepoint further improves the accuracy of the approximation in almost all cases.

Keywords: Contingency tables; Darwin's finch data; Generalized linear model; Uniform association; Quasi-independence.

1. Introduction

Let $\{y_{ij}\}$ denote the counts in an $r \times c$ contingency table. How many tables are there with the same row and column margins? Gail & Mantel

[*] Ph.D. Student, Department of Statistical Science, Cornell University

[†] Professor, Biological Statistics and Computational Biology, Cornell University. Email: Jim.Booth@Cornell.edu

[‡] Assistant Professor, Department of Statistics, University of Kentucky

(1977) proposed the following approximation obtained via an application of the central limit theorem.

Let $(Y_{i1},\ldots,Y_{ic})$ denote a random vector that assigns equal probability to every ordered set of c non-negative integers summing to $y_{i\cdot}$, independently for $i=1,\ldots,r$. Then

$$\begin{aligned} E(Y_{ij}) &= y_{i\cdot}/c\,, \\ \mathrm{var}(Y_{ij}) &= y_{i\cdot}(y_{i\cdot}+c)(c-1)/(c+1)c^2\,, \\ \mathrm{cov}(Y_{ij},Y_{ik}) &= -y_{i\cdot}(y_{i\cdot}+c)/(c+1)c^2\,. \end{aligned}$$

It follows that the column sums, $Y_{\cdot 1},\ldots,Y_{\cdot c}$, are identically distributed and equicorrelated with

$$\begin{aligned} E(Y_{\cdot j}) &= y_{\cdot\cdot}/c\,, \\ \sigma^2 = \mathrm{var}(Y_{\cdot j}) &= \sum_{i=1}^{r} y_{i\cdot}(y_{i\cdot}+c)(c-1)/(c+1)c^2\,, \\ \mathrm{cov}(Y_{\cdot j},Y_{\cdot k}) &= -\sigma^2/(c-1)\,. \end{aligned}$$

Hence, the multivariate normal approximation to the conditional probability of the observed vector of column marginal totals given the row totals is

$$p(y_{\cdot 1},\ldots,y_{\cdot c}) = ((c-1)/2\pi\sigma^2 c)^{(c-1)/2} c^{1/2} \exp(-Q/2)\,,$$

where $Q = ((c-1)/\sigma^2 c)(\sum_{j=1}^{c} y_{\cdot j}^2 - y_{\cdot\cdot}^2/c)$. The total number of tables with unrestricted column totals is

$$N = \prod_{i=1}^{r} \binom{y_{i\cdot}+c-1}{c-1}\,. \qquad (1)$$

The Gail and Mantel approximation to the number of tables with the same row and column margins as $\{y_{ij}\}$ is then $N \times p$.

As an example, Gail and Mantel consider the 4×3 table with row and column margins $\{20, 10, 5, 5\}$ and $\{11, 10, 19\}$ respectively. Then the approximation gives 21,469 tables, which is in good agreement with the exact number, 22,245. However, the approximation is not symmetric in the rows and columns. If the approximation is applied to the transpose of the contingency table the approximation to the number of tables is 11,933, which is far from the correct answer. To be fair, the approximation can be expected to work well when the number of rows (or columns) being averaged over is large relative to the number of columns (or rows), and so it is not surprising that the approximation based on averaging over rows is better in this instance. As another example, consider the 5×5 table of pathologists ratings from Holmquist et al. (1967) (see also Agresti, 1990, p.368). In this case the row and column margins are $\{26, 26, 38, 22, 6\}$, and $\{27, 12, 69, 7, 3\}$, respectively. The normal approximation gives 12.5 billion tables, and 261 billion after transposing the rows and columns. The correct answer in this case is 193,316,293,000, which was computed using exact algebraic methods.

An alternative analytical approximation for the number of two-way contingency tables with fixed margins is given in Diaconis & Efron (1985). However, this approximation can also be quite inaccurate. Holmes & Jones (1996) give an example of a 5×4 table with row margins, $\{9, 49, 182, 478, 551\}$, and column margins, $\{9, 309, 355, 596\}$. In this case the exact number of possible tables is 33,819,042,818,100,768 or 3.382×10^{16} to four significant figures. Applying the Diaconis-Efron formula results in the approximations, 1.319×10^{17}, and 4.126×10^{16}, after switching rows and columns. Thus, the Diaconis-Efron formula is in error by at least 20%.

In this paper we propose a double-saddlepoint approximation to the number of contingency tables whose counts meet certain linear constraints. The approximation is based on a probabilistic formulation involving a geometric generalized linear model. Computing the approximation involves fitting the generalized linear model (GLM) which can be accomplished essentially instantaneously. The approximation is shown to be extremely accurate in a range of examples, with the relative error generally less than 5%. For example, the approximation to the number of tables with the pathologists ratings table margins is 205 billion. Transposing the table makes little difference resulting in a value of 202 billion. For the data from Gail & Mantel (1977) the corresponding approximations are 20,321 and 21,536, and for the data from Holmes & Jones (1996) the approximations are respectively 3.303×10^{16} and 3.428×10^{16}. However, in almost all cases the approximation is improved by using an easily computed higher-order correction to the double saddlepoint.

An outline of the paper is as follows. In Section 2 we discuss the formulation of the counting problem for two-way tables with fixed margins in terms of a geometric generalized linear model. The double saddlepoint approximation and higher-order correction are described in Section 3. The GLM formulation is then generalized to include multi-way tables and tables with additional constraints in Section 4. Results for several examples are presented in Section 5. A similar approximation for tables containing only zeros and ones based on a logistic GLM probabilistic formulation is presented in Section 6. Exact algebraic and importance sampling methods for table counting are discussed briefly in Section 7. The paper concludes in Section 8 with some discussion.

2. GLM Formulation

Let Y be a geometric random variable with success probability, π. Then, $\mu = E(Y) = (1-\pi)/\pi$, and for $y = 0, 1, \ldots$,

$$\begin{aligned} P(Y=y) &= (1-\pi)^y \pi \\ &= \left(\frac{\mu}{\mu+1}\right)^y \frac{1}{\mu+1} \\ &= \exp\left\{y\theta + \log(1-e^\theta)\right\} , \end{aligned}$$

where $\theta = \log(\mu) - \log(\mu+1)$ is the canonical parameter. If $Y_1, \ldots, Y_c$ are i.i.d. geometric random variables, then their sum, $Y.$, is negative binomial with mass function,

$$P(Y. = y.) = \binom{y. + c - 1}{c-1} (1-\pi)^{y.} \pi^c$$

for $y. = 0, 1, \ldots$. It follows that the conditional distribution of $(Y_1, \ldots, Y_c)$ given $Y. = y.$ is given by

$$P(Y_1 = y_1, \ldots, Y_c = y_c | Y. = y.) = \binom{y. + c - 1}{c-1}^{-1} ,$$

for all non-negative count vectors, $(y_1, \ldots, y_c)$, summing to $y.$.

Now, let $\{Y_{ij}\}$ be a table of counts whose entries are independent geometric random variables with canonical parameters, $\{\theta_{ij}\}$. Consider the generalized linear model,

$$\theta_{ij} = \lambda + \lambda_i^R + \lambda_j^C \tag{2}$$

for $i = 1, \ldots, r$ and $j = 1, \ldots, c$, where R and C denote the nominal-scale row and column factors. Notice that the row and column margins are sufficient statistics for this model. Hence, the conditional distribution of the table counts given the margins is the same regardless of the values of the parameters in the model. In particular, suppose that the column effects are all equal, $\lambda_1^C = \cdots = \lambda_c^C = 0$ say. In this case the counts in each row of the table are i.i.d.. Furthermore, after conditioning on a row margin, the probabilities of all ordered sets of counts summing to the margin are equal.

3. Double-Saddlepoint Approximation

The double-saddlepoint approximation provides an accurate alternative to the normal approximation which can be formulated in terms of the GLM described in the previous section. In general, for a exponential family model (or a GLM with canonical link), the probability density of the sufficient statistic vector, $\mathbf{S}$, can be approximated by the formula

$$\hat{f}_{\mathbf{S}}(\mathbf{s}) = |2\pi\hat{\mathbf{I}}|^{-1/2}\exp(-\hat{l})\,, \tag{3}$$

where $\hat{l}$ is the maximized loglikelihood, and $\hat{\mathbf{I}}$ is the observed information matrix. This formula is originally due to Daniels (1954), although he didn't express it in likelihood notation. In our case we want to approximate a conditional probability for the column margins, $\mathbf{s}_2$, given the row margins, $\mathbf{s}_1$, where $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2)$. This is accomplished by taking a ratio of two saddlepoint approximations of the form (3),

$$\hat{f}(\mathbf{s}_2|\mathbf{s}_1) = \frac{\hat{f}_{\mathbf{S}}(\mathbf{s})}{\hat{f}_{\mathbf{S}_1}(\mathbf{s}_1)} = \left\{\frac{|2\pi\hat{\mathbf{I}}|}{|2\pi\hat{\mathbf{I}}_1|}\right\}^{-1/2}\exp(\hat{l}_1 - \hat{l})\,, \tag{4}$$

where $\hat{l}_1$ is the constrained maximum of the loglikelihood, when the column effects parameters in (2) are all zero.

For the pathologists' ratings data, application of (4) results in the approximation $p = 7.765 \times 10^{-10}$. The formula (1) for the number of tables with the same row margins yields $N = 2.639 \times 10^{+20}$. Multiplying these two numbers leads to an approximate number of tables with the same margins equal to 204.9 billion.

Adding higher order terms to the saddlepoint approximation improves its accuracy, at least asymptotically. We now consider two ways of correcting (3), both discussed in Butler (2007). The first is an additive correction,

$$\tilde{f}_1(\mathbf{s}) = \hat{f}_{\mathbf{S}}(\mathbf{s})(1 + O)\,,$$

and the second is its exponential counterpart suggested by McCullagh (1987, Section 6.3)

$$\tilde{f}_2(\mathbf{s}) = \hat{f}_{\mathbf{S}}(\mathbf{s})e^{O}\,.$$

The correction term, O, is given by the formula (Butler, 2007, Section 3.2.2)

$$O = \frac{1}{8}\hat{\kappa}_4 - \frac{1}{24}(2\hat{\kappa}_{23}^2 + 3\hat{\kappa}_{13}^2) \tag{5}$$

where

$$\hat{\kappa}_4 = \sum_{i,j,k,l} \hat{K}_{ijkl}\hat{K}^{ij}\hat{K}^{kl}\,, \tag{6}$$

$$\hat{\kappa}_{13}^2 = \sum_{i,j,k,r,t,u} \hat{K}_{ijk}\hat{K}_{rtu}\hat{K}^{ij}\hat{K}^{kr}\hat{K}^{tu}\,, \quad \hat{\kappa}_{23}^2 = \sum_{i,j,k,r,t,u} \hat{K}_{ijk}\hat{K}_{rtu}\hat{K}^{ir}\hat{K}^{jt}\hat{K}^{ku}\,, \tag{7}$$

and

$$\hat{K}_{ijk} = -\frac{\partial^3 l(\hat{\mathbf{s}})}{\partial s_i \partial s_j \partial s_k}\,, \qquad \hat{K}^{ij} = (\hat{I}^{-1})_{ij}\,.$$

In the case of the double-saddlepoint approximation using the correction in the numerator and denominator leads to

$$\tilde{f}_1(\mathbf{s}_2|\mathbf{s}_1) = \hat{f}(\mathbf{s}_2|\mathbf{s}_1)(1 + O_{\mathbf{s}_1,\mathbf{s}_2} - O_{\mathbf{s}_1}) \tag{8}$$

and

$$\tilde{f}_2(\mathbf{s}_2|\mathbf{s}_1) = \hat{f}(\mathbf{s}_2|\mathbf{s}_1)\exp\{O_{\mathbf{s}_1,\mathbf{s}_2} - O_{\mathbf{s}_1}\} \tag{9}$$

respectively. The estimated number of the tables is then

$$\tilde{N}_i(\mathbf{s}_1, \mathbf{s}_2) = N(\mathbf{s}_1)\tilde{f}_i(\mathbf{s}_2|\mathbf{s}_1)\,, \tag{10}$$

for $i = 1, 2$, where we assume that $N(\mathbf{s}_1)$, the exact number of the tables with a fixed $\mathbf{s}_1$, is known.

Naive computation of the summations in (6) and (7), that are required for the correction terms, respectively involves $O(p^4)$ and $O(p^6)$ operations, a potentially time consuming task if p is large. However, the vast majority of the terms in these sums are zero. It is shown in the Appendix that this sparseness can be exploited, resulting in computational times that are almost instantaneous.

4. Multi-way Tables and Additional Constraints

It is clear, in principle, that the double saddlepoint approximation extends to multi-way tables, since the number of tables with one margin fixed is also known in this case. In addition, it is often the case that the independence assumption for a two-way (or multi-way) contingency table is unreasonable. In such cases one can attempt to describe the dependence in a parsimonious way by placing restrictions on interaction terms.

The most general setting is as follows. Consider the set Γ consisting of all non-negative integer vectors, $\mathbf{y}$, satisfying a set of linear constraints, $\mathbf{X}^T\mathbf{y} = \mathbf{s}$, where $\mathbf{X} \in \mathbb{Z}^{d\times f}$ and $\mathbf{s} \in \mathbb{Z}^f$; that is,

$$\Gamma := \left\{\mathbf{y} \in \mathbb{Z}^d \,:\, \mathbf{X}^T\mathbf{y} = \mathbf{s}, \mathbf{y} \geq \mathbf{0}\right\}\,.$$

We assume, without loss of generality that $\mathbf{X}$ is full column rank. For example, if $\mathbf{y}$ consists of the counts from an $r \times c$ table with fixed margins, then $d = rc$ and $f = r + c - 1$. Suppose $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ is a partition of the columns of $\mathbf{X}$, and let $(\mathbf{s}_1, \mathbf{s}_2) = (\mathbf{X}_1^T \mathbf{y}, \mathbf{X}_2^T \mathbf{y})$ be the corresponding partition of $\mathbf{s}$. Suppose that the cardinality of the set $\Gamma_1 := \left\{ \mathbf{y} \in \mathbb{Z}^d \,:\, \mathbf{X}_1^T \mathbf{y} = \mathbf{s}_1, \mathbf{y} \geq \mathbf{0} \right\}$ is known. Then the double saddlepoint approximation described in the previous section, and its corrected version (10), can be used to approximate the cardinality of Γ.

The most general loglinear association model for an $r \times c$ contingency table has a canonical linear predictor of the form

$$\theta_{ij} = \lambda + \lambda_i^R + \lambda_j^C + \lambda_{ij}^{RC} \tag{11}$$

for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. A special case is the uniform association (UA) model in which $\lambda_{ij}^{RC} = \beta ij$. This model describes the dependence between the row and column factors in terms of a single parameter, β. The model implies that all local odds-ratios (from tables formed by the intersection of two adjacent rows and two adjacent columns) are equal to e^β – hence the name "uniform association". The sufficient statistics for the UA model include the row and column margins and, in addition, the sum of products of row and column numbers weighted by the cell counts. Other examples include quasi-independence (QI), $\lambda_{ij}^{RC} = 0$ if $i \neq j$, and diagonal (D), $\lambda_{ij}^{RC} = 0$ if $i \neq j$ and $\lambda_{ii}^{RC} = \lambda$, association models.

The generalization of (11) to a three-way, $I \times J \times K$, contingency table is

$$\theta_{ijk} = \lambda + \lambda_i^R + \lambda_j^C + \lambda_k^Z + \lambda_{ij}^{RC} + \lambda_{ik}^{RZ} + \lambda_{jk}^{CZ} + \lambda_{ijk}^{RCZ} \tag{12}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$, where Z denotes the nominal-scale factor associated with the third dimension. As in the two-way case associations between the three factors are modeled by placing restrictions on the interaction terms. Some widely-used examples are given in Table 1. The models are nested in the sense that each successive model imposes a subset of the restrictions in the previous one. Since, the dimension of the statistic, **s**, increases with the model complexity, the number of possible tables with the same value of **s** decreases.

Model	Interaction constraints			
1. (R,C,Z)	$\lambda_{ijk}^{RCZ} = 0$	$\lambda_{ik}^{RZ} = 0$	$\lambda_{jk}^{CZ} = 0$	$\lambda_{ij}^{RC} = 0$
2. (RC,Z)	$\lambda_{ijk}^{RCZ} = 0$	$\lambda_{ik}^{RZ} = 0$	$\lambda_{jk}^{CZ} = 0$	
3. (RC,CZ)	$\lambda_{ijk}^{RCZ} = 0$	$\lambda_{ik}^{RZ} = 0$		
4. (RC,RZ,CZ)	$\lambda_{ijk}^{RCZ} = 0$			

Table 1: Some common loglinear association models for three-way tables (see Agresti, 1990, p.144). Model 1 implies the factors, R, C, and Z, are mutually independent; Model 2 implies Z is jointly independent of R and C; Model 3 implies R and Z are conditionally independent given C; and Model 4 implies a homogeneous pattern of conditional association between R and C across all levels of Z.

5. Examples

To approximate the number of $r \times c$ tables with fixed marginal totals and additional constraints we can apply the formula (4) with **s** equal to the full vector of sufficient statistics and $\mathbf{s}_1$ equal to the sub-vector of row (or column) marginal totals. This results in an approximation to the conditional probability that the column marginal totals and the additional sufficient statistic take their observed values conditional on the row margins. This, in turn, can be multiplied by the known number of tables with the same row margins to get an approximation to the number of tables meeting all sufficiency constraints. The same approach can be applied in multi-way tables with $\mathbf{s}_1$ equal to any one of the table margins.

Tables 2, 3 and 4 summarize the accuracy of the double saddlepoint approximation, and the additive and exponential corrections, for approximating the numbers of contingency tables with different linear constraints on the counts. Table 2 concerns 5×5 tables constrained to have the same sufficient statistics as the pathologists' ratings data assuming I, UA, QI and D loglinear association models. Table 3 gives the analogous results for 4×4 tables using the sexual fun data reported in Hout et al. (1987) (see also Agresti, 1990, p.32). Table 4 gives results for $2 \times 2 \times 8$ tables with sufficiency constraints equal to those from a smoking and lung cancer study in eight Chinese cities (Agresti, 1996, p.60).

The value tabulated for each approximation is its accuracy, defined as (signed) percentage relative error

$$\mathrm{PRE}(\hat{N}) = 100 \times \frac{\hat{N} - N}{N}$$

where N is the exact number of the tables satisfying the relevant set of constraints. In every case the exponentially corrected approximation is the most accurate, and in many cases the improvement over the uncorrected double saddlepoint is substantial. Also, the exponentially corrected approximation does not appear to be affected much by which margin is conditioned upon. In general the accuracy of the approximation decreases with the dimension of the statistic, $\mathbf{s}$, relative to the number of cells in the table. This is to be expected because, the larger the dimension of $\mathbf{s}$, the fewer counts being summed over. In Table 4, the dimension of the statistic $\mathbf{s}$ is equal to the number of parameters in each association model. The least accurate approximation was for the homogeneous association model, (RC,RZ,CZ), which, for a $2 \times 2 \times 8$ table, has 25 parameters.

Model	Margin	$\hat{N}$	$\tilde{N}_1$	$\tilde{N}_2$	N
I	Row	$+4.18$	-3.36	-3.10	1.933×10^{11}
	Column	$+5.92$	-3.51	-3.11	
UA	Row	$+16.17$	-8.78	-6.28	34,670
	Column	$+14.26$	-9.02	-6.80	
QI	Row	$+31.12$	-9.19	-3.58	435
	Column	$+45.11$	-14.29	-3.63	
D	Row	$+23.08$	-7.59	-4.07	1,132,576
	Column	$+21.07$	-7.09	-4.06	

Table 2: Percentage relative errors of the double saddlepoint approximation, and higher-order corrections, for the numbers of 5 × 5 tables meeting the same set of linear constraints as the pathologists' ratings data from Agresti (1990, p.368)

Model	Margin	$\hat{N}$	$\tilde{N}_1$	$\tilde{N}_2$	N
I	Row	-12.61	-1.95	-1.27	$947,766,430$
	Column	-12.67	-1.95	-1.27	
UA	Row	-12.08	-2.32	-1.76	$8,137,492$
	Column	-12.15	-2.33	-1.76	
QI	Row	$+27.64$	-12.00	-6.43	15,708
	Column	$+27.55$	-11.97	-6.43	
D	Row	-13.79	-3.83	-3.24	27,209,031
	Column	-13.85	-3.84	-3.24	

Table 3: Percentage relative errors of the double saddlepoint approximation, and higher-order corrections, for the numbers of 4×4 tables meeting the same set of linear constraints as the sexual fun data from Agresti (1990, p.32)

Model	Margin	$\hat{N}$	$\tilde{N}_1$	$\tilde{N}_2$	N
(R,C,Z)	$I \times JK$	$+3.93$	-0.12	-0.04	3.918×10^{54}
	$J \times IK$	$+3.93$	-0.12	-0.04	
	$K \times IJ$	-11.06	-0.64	-0.01	
(RC,Z)	$I \times JK$	$+4.85$	-1.22	-1.05	2.530×10^{51}
	$J \times IK$	$+4.85$	-1.22	-1.05	
	$K \times IJ$	-10.28	-1.46	-1.02	
(RC,CZ)	$I \times JK$	$+63.83$	-24.37	-4.37	3.425×10^{33}
	$J \times IK$	$+63.83$	-24.37	-4.37	
	$K \times IJ$	$+40.18$	-13.38	-4.34	
(RC,RZ,CZ)	$I \times JK$	$+135.35$	-85.14	-7.77	2.262×10^{15}
	$J \times IK$	$+135.35$	-85.14	-7.77	
	$K \times IJ$	$+101.37$	-55.82	-7.74	

Table 4: Percentage relative errors of the double saddlepoint approximation, and higher order corrections, for the numbers of 2 × 2 × 8 tables with the same sufficient statistics as the Chinese smoking and lung cancer data from Agresti (1996, p.60) under various loglinear association models.

6. Counting Tables of Zeros and Ones

The number of tables with only 0-1 entries meeting linear constraints can be also be approximated based on a GLM formulation. This case requires a logistic model for binary observations instead of the geometric model discussed above. Specifically, suppose that $\{Y_{ij}\}$ is an $r \times c$ table of independent binary counts with associated success probabilities $\{\pi_{ij}\}$. Consider a model of the form (2), where θ is now the logit of π. Once again the marginal totals are the sufficient statistics, and hence the conditional distribution given the margins is independent of the parameters. In particular, if $\lambda_1^C = \cdots = \lambda_c^C = 0$, the conditional distribution of the counts in each row, given the row margins, is uniform over the set of all possible assignments of the zeros and ones; that is, every possible assignment in row i has probability $\binom{c}{y_{i\cdot}}^{-1}$

Margin	$\hat{N}$	$\tilde{N}_1$	$\tilde{N}_2$	N
Row	302.9	-258.1	0.131	6.715×10^{16}
Column	238.3	-174.0	0.030	

Table 5: Percentage relative errors of the double saddlepoint approximation, and higher-order corrections, for the number of 13 × 17 tables of zeros and ones with the same margins as Darwin's finch data.

For an illustration, consider Darwin's data concerning the presence or absence of 13 species of finch in 17 Galápagos islands (see Liu, 2001, p.93). The exact number of tables with the same margins as this dataset is given in Chen et al. (2005, Section 6.1). To four significant figures it is 6.715×10^{16}. Percentage relative errors of the double saddlepoint approximation, and the additive and exponential corrections are given in Table 5. In this case the uncorrected double saddlepoint is off by over 200%, and the additive correction over-corrects, resulting in a negative estimate. However, the exponentially corrected double saddlepoint is almost exact, with a relative error significantly less than 1%.

7. Other Counting Methods

7·1. *Exact Algebraic Computation*

The problem of counting the number of contingency tables meeting certain linear constraints is equivalent to counting the set of integral points of a rational convex polytope of the form

$$P := \left\{ \mathbf{y} \in \mathbb{R}^d \, : \, \mathbf{X}^T \mathbf{y} = \mathbf{s}, \mathbf{y} \geq \mathbf{0} \right\} ,$$

where $\mathbf{X} \in \mathbb{Z}^{d \times f}$ and $\mathbf{s} \in \mathbb{Z}^f$. If we now define the generating function,

$$f(P; \mathbf{y}) = \sum_{\boldsymbol{\alpha} \in P \cap \mathbb{Z}^d} \mathbf{y}^{\boldsymbol{\alpha}} ,$$

then $|P \cap \mathbb{Z}^d| = f(P; \mathbf{1})$.

As an example, suppose that P is the one-dimensional polytope $[0, N]$. Then, $f(P; x) = 1 + x + x^2 + \cdots + x^N$, $f(P; x)$ can be represented by the rational function $\frac{1-x^{N+1}}{1-x}$, and $f(P; 1) = N + 1$, the number of integer points in P. Note that substituting $x = 1$ yields a denominator equal to zero in the rational function, so some analytic technique must be used to evaluate $f(P; 1)$. In this particular case, we could take the limit as x approaches 1 and apply l'Hospital's rule. In general, we must use more complicated residue calculus as described in Barvinok (1994). The exact answers for the examples in Tables 2 and 3 each took less than 30 seconds using a C++ implementation of Barvinok's algorithm available at `http://www.math.ucdavis.edu/~latte` (DeLoera et al., 2004). For the examples in Table 4 the computations took 4450, 1209, and 85 seconds for Models 1-3 respectively, and less than 1 second for Model 4. Other analytical methods have been developed which are faster in special cases such as counting two-way tables with fixed margins (Beck, 2000). For a recent review, see Yoshida (2004). The computing time for algebraic methods can be prohibitive for larger tables. However, this is precisely the situation in which the saddlepoint approximation is likely to be most accurate because the dimension of $\mathbf{s}$ relative to the number of cells in the table decreases as the dimensions of the table grow.

7·2. *Importance Sampling*

Let $q : \Gamma \to \mathbb{R}$ be a probability mass function which assigns positive probability to all vectors, $\mathbf{y}$, in the finite set Γ. Then the cardinality of Γ can be expressed as

$$|\Gamma| = \sum_{\mathbf{y} \in \Gamma} 1 = \sum_{\mathbf{y} \in \Gamma} \frac{1}{q(\mathbf{y})} q(\mathbf{y}) = E_q \left\{ \frac{1}{q(\mathbf{y})} \right\} .$$

Hence, if it is possible to simulate an i.i.d. sequence, $\mathbf{y}_1 \ldots, \mathbf{y}_N$, from q, and to evaluate, $q(\mathbf{y}_i)$, $i = 1, \ldots, N$, then a Monte Carlo approximation to $|\Gamma|$ is given by

$$\widehat{|\Gamma|} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{q(\mathbf{y}_i)} \,.$$

For example, Chen et al. (2005) construct a probability mass function for counting two-way tables with fixed margins of the form

$$q(\mathbf{y}) = q_1(y_1) q_2(y_2|y_1) \cdots q_d(y_d|y_{-d}).$$

Since the approximation is the mean of an i.i.d. sample, standard errors for $\widehat{|\Gamma|}$, and hence confidence intervals for $|\Gamma|$, can also be constructed. Chen et al. (2005) also develop an importance sampling method for counting two-way zero-one tables with fixed margins.

8. Discussion

We have proposed a new way of approximating the number of contingency tables, and tables of zeros and ones, that satisfy certain linear constraints. The approximations involve fitting generalized linear models which can be accomplished almost instantaneously. The approximations are much more accurate than analytical approximations that have been proposed previously, and can be applied in a wider range of problems. In addition, they can be applied in problems for which exact algebraic methods are not yet computationally feasible. An alternative approach is to use Monte Carlo methods such as those developed recently by Chen et al. (2005) for two-way tables with fixed margins. It is not clear to what extent this approach can be modified to deal with additional constraints such as those imposed by sufficiency in association models. In any case, the approximations developed in this paper provide a computational feasible method for counting tables in a wide variety of settings.

Acknowledgment

Professor Booth was partially supported by the National Science Foundation award, DMS-0405543.

References

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley and Sons.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley and Sons.

Barvinok, A. (1994). Polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. *Math of Operations Research* 19 769–779.

Beck, M. (2000). Counting lattice points by means of the residue theorem. *Ramanujan Journal* 4 299–310.

Butler, R. W. (2007). *Saddlepoint Approximations with Applications*. No. 22 in Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press.

Chen, Y., Diaconis, P., Holmes, S. & Liu, J. S. (2005). Sequential monte carlo methods for statistical analysis of tables. *Journal of the American Statistical Association* 100 109–120.

DeLoera, J., Hemmecke, R., Tauzer, J. & Yoshida, R. (2004). Effective lattice point counting in rational convex polytopes. *Journal of Symbolic Computation* 38 1273–1302.

Diaconis, P. & Efron, B. (1985). Testing for independence in a two-way table: new interpretations of the chi-square statistic. *The Annals of Statistics* 13 845–874.

Gail, M. & Mantel, N. (1977). Counting the number of contingency tables with fixed margins. *Journal of the American Statistical Association* 72 859–862.

Holmes, R. B. & Jones, L. K. (1996). On uniform generation of two-way tables with fixed margins and the conditional volume test of Diaconis and Efron. *The Annals of Statistics* 24 64–68.

Holmquist, N. S., McMahon, C. A. & Williams, O. D. (1967). Variability in classification of carcinoma in situ of the uterine cervix. *Arch. Pathol.* 84 334–345.

Hout, M., Duncan, O. D. & Sobel, M. E. (1987). Association and heterogeneity: Structural models of similarities and differences. *Sociological Methodology* 17 145–184.

Liu, J. S. (2001). *Monte Carlo Stategies in Scientific Computing*. Springer.

McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall.

Yoshida, R. (2004). *Barvinok's Rational Functions: Algorithms and Applications to Optimization, Statistics, and Algebra*. Ph.D. thesis, University of California at Davis. http://www.ms.uky.edu/~ruriko/.

Appendix

We show how to calculate the correction term O in a computationally efficient way by using sparse structure of matrix $\mathbf{X}$. From equations (5), (6), and (7) it is clear that the complexity of calculating O is dominated by the number of operations required to get $\hat{\kappa}_{13}^2$ and $\hat{\kappa}_{23}^2$ which is $O(p^6)$. Note that, for the Finch data example, $p = 1 + (I-1) + (J-1) = 28$. Therefore, it would need 4×10^9 operations to calculate $\hat{\kappa}_{13}^2$ and $\hat{\kappa}_{23}^2$. Calculating $\hat{\kappa}_4$ is much less costly, involving $O(p^4)$ operations.

However, due to the sparsity of matrix $\mathbf{X}$, most of the terms in all three sums of (6), and (7) turn out to be zeros and may be ignored, substantially decreasing the complexity. We now describe an efficient algorithm for finding the non-zero terms.

Let

$$T^3 = \{(t_1, t_2, t_3) \in \mathbb{P}^3 : \hat{K}_{t_1 t_2 t_3} \neq 0\}$$

and

$$T^4 = \{(t_1, t_2, t_3, t_4) \in \mathbb{P}^4 : \hat{K}_{t_1 t_2 t_3 t_4} \neq 0\}$$

where $\mathbb{P} = \{1, 2, \ldots, p\}$. Then clearly

$$\hat{\kappa}_{13}^2 = \sum_{(t_1,t_2,t_3)\in T^3} \sum_{(t_4,t_5,t_6)\in T^3} \hat{K}_{t_1t_2t_3}\hat{K}_{t_4t_5t_6}\hat{K}^{t_1t_2}\hat{K}^{t_3t_4}\hat{K}^{t_5t_6}. \tag{13}$$

$$\hat{\kappa}_{23}^2 = \sum_{(t_1,t_2,t_3)\in T^3} \sum_{(t_4,t_5,t_6)\in T^3} \hat{K}_{t_1t_2t_3}\hat{K}_{t_4t_5t_6}\hat{K}^{t_1t_4}\hat{K}^{t_2t_5}\hat{K}^{t_3t_6}. \tag{14}$$

$$\hat{\kappa}_4 = \sum_{(t_1,t_2,t_3,t_4)\in T^4} \hat{K}_{t_1t_2t_3t_4}\hat{K}^{t_1t_2}\hat{K}^{t_3t_4}. \tag{15}$$

The third and the fourth derivatives of the log-likelihood function are given by

$$l^{(3)}(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} b^{(3)}(\mathbf{x}_{ijk}^T\hat{\boldsymbol{\theta}})\left[\mathbf{x}_{ijk} \otimes \mathbf{x}_{ijk}\mathbf{x}_{ijk}^T\right]$$

and

$$l^{(4)}(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} b^{(4)}(\mathbf{x}_{ijk}^T\hat{\boldsymbol{\theta}})\left[\mathbf{x}_{ijk}\mathbf{x}_{ijk}^T \otimes \mathbf{x}_{ijk}\mathbf{x}_{ijk}^T\right] .$$

Notice that the only difference between the geometric and binomial models is function $b(\cdot)$. This difference is immaterial to the arguments that follow. Now note that

$$\hat{K}_{t_1t_2t_3} = l^{(3)}(\hat{\boldsymbol{\theta}})_{t_1t_2t_3} = -\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} b^{(3)}(\mathbf{x}_{ijk}^T\hat{\boldsymbol{\theta}})x_{ijk}^{t_1}x_{ijk}^{t_2}x_{ijk}^{t_3}$$

and

$$\hat{K}_{t_1t_2t_3t_4} = l^{(4)}(\hat{\boldsymbol{\theta}})_{t_1t_2t_3t_4} = -\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} b^{(4)}(\mathbf{x}_{ijk}^T\hat{\boldsymbol{\theta}})x_{ijk}^{t_1}x_{ijk}^{t_2}x_{ijk}^{t_3}x_{ijk}^{t_4}$$

are symmetric with respect to permutation of the subindices (t_1, t_2, t_3) and (t_1, t_2, t_3, t_4). Hence, the sets of indexes T^3 and T^4 may be represented as

$$T^3 = \bigsqcup_{\mathbf{t}\in T^3_{\leq}} \mathbf{Or}^3(\mathbf{t}) \tag{16}$$

and

$$T^4 = \bigsqcup_{\mathbf{t}\in T^4_{\leq}} \mathbf{Or}^4(\mathbf{t})\,,$$

where

$$T^3_{\leq} = \{(t_1, t_2, t_3) \in T^3 : t_1 \leq t_2 \leq t_3\}$$

and

$$T^4_{\leq} = \{(t_1, t_2, t_3, t_4) \in T^4 : t_1 \leq t_2 \leq t_3 \leq t_4\}\,,$$

and where $\mathbf{Or}^3(\mathbf{t})$ denotes all the triples in T^3 that can be obtained by permuting $\mathbf{t}$ from $T^3_{\leq}$. Similarly, $\mathbf{Or}^4(\mathbf{t})$ contains all four-tuples which are permutations of $\mathbf{t}$ from $T^4_{\leq}$. For instance, if $\mathbf{t} = (2, 4, 4, 9)$ belongs to $T^4_{\leq}$, then $\mathbf{Or}^4(\mathbf{t})$, in addition to the original four-tuple $(2, 4, 4, 9)$, includes also its permutations $(2,9,4,4)$, $(9,2,4,4)$, $(9,4,4,2)$, $(4,4,2,9)$, and $(4,4,9,2)$. Therefore, it suffices to determine the sets $T^3_{\leq}$ and $T^4_{\leq}$ to calculate the sums in (13) - (15).

We illustrate the algorithm using the independence model in a three-way table. For each count, y_{ijk}, the corresponding vector, $\mathbf{x}_{ijk} \in \{0,1\}^p$, with $p = 1 + (I-1) + (J-1) + (K-1)$, has at most four non-zero coordinates. So it is convenient to view $\mathbf{x}_{ijk}$ as the union of four subcomponents,

$$\mathbf{x}_{ijk} = \left(\boxed{Z_1} \boxed{\quad Z_2 \quad} \boxed{\quad Z_3 \quad} \boxed{\quad Z_4 \quad} \right),$$

with Z_1 of size 1, Z_2 of size $I-1$, Z_3 of size $J-1$, and Z_4 of size $K-1$. Let

$$\mathbf{e}_l^{L-1} = \begin{cases} \mathbf{e}_l \in \mathbb{R}^{L-1}, & l = 1, \ldots, L-1 \\ \mathbf{0} \in \mathbb{R}^{L-1}, & l = L \end{cases}$$

where $\mathbf{e}_l$ is a component of the standard basis for $\mathbb{R}^{L-1}$. Then we can write the vector, $\mathbf{x}_{ijk}$, as

$$(1, \mathbf{e}_i^{I-1}, \mathbf{e}_j^{J-1}, \mathbf{e}_k^{K-1}). \tag{17}$$

To compute (13) - (14) we need to find all triples, $(t_1 \leq t_2 \leq t_3)$, for which there exists at least one combination (i,j,k) with $x_{ijk}^{t_1} x_{ijk}^{t_2} x_{ijk}^{t_3} \neq 0$. The set of such triples obviously includes $T^3_{\leq}$, and is exactly equal to $T^3_{\leq}$ in all examples considered in this paper. Table 6 contains all possible combinations for triples

t_1	Z_1	Z_1	Z_1	Z_1	Z_1	Z_1	Z_1	Z_1	Z_1	Z_1
t_2	Z_1	Z_1	Z_1	Z_1	Z_2	Z_2	Z_2	Z_3	Z_3	Z_4
t_3	Z_1	Z_2	Z_3	Z_4	Z_2	Z_3	Z_4	Z_3	Z_4	Z_4

t_1	Z_2	Z_2	Z_2	Z_2	Z_2	Z_2	Z_3	Z_3	Z_3	Z_4
t_2	Z_2	Z_2	Z_2	Z_3	Z_3	Z_4	Z_3	Z_3	Z_4	Z_4
t_3	Z_2	Z_3	Z_4	Z_3	Z_4	Z_4	Z_3	Z_4	Z_4	Z_4

Table 6: All combinations for indexes $t_1 \leq t_2 \leq t_3$.

$t_1 \leq t_2 \leq t_3$. For each $(Z_{l_1}, Z_{l_2}, Z_{l_3})$ in Table 6 all the relevant triples can be calculated by using ”inversion” functions $h_{Z_l} : \mathbb{I} \times \mathbb{J} \times \mathbb{K} \to Z_l$ where

$$h_{Z_l}(i,j,k) = \begin{cases} 1, & l = 1 \\ 1 + i, & l = 2 \text{ and } i \leq I - 1 \\ \emptyset, & l = 2 \text{ and } i = I \\ 1 + (I-1) + j, & l = 3 \text{ and } j \leq J - 1 \\ \emptyset, & l = 3 \text{ and } j = J \\ 1 + (I-1) + (J-1) + k, & l = 4 \text{ and } k \leq K - 1 \\ \emptyset, & l = 4 \text{ and } k = K \end{cases}$$

It follows from (17) that for any given (i,j,k) the corresponding value $x_{ijk}^{t_1} x_{ijk}^{t_2} x_{ijk}^{t_3}$ is non-zero if and only if $t_1 = h_{Z_{l_1}}(i,j,k)$, $t_2 = h_{Z_{l_2}}(i,j,k)$, and $t_3 = h_{Z_{l_3}}(i,j,k)$. Hence, if we define

$$g_i(Z_{l_1}, Z_{l_2}, Z_{l_3}) = \begin{cases} 1, & l_1 = 2 \,\text{or}\, l_2 = 2 \,\text{or}\, l_3 = 2 \\ I - 1, & \text{otherwise} \end{cases} ,$$

and the similar functions for the indices, j and k, then the number of all relevant triples from $(Z_{l_1}, Z_{l_2}, Z_{l_3})$ is given by

$$g_i(Z_{l_1}, Z_{l_2}, Z_{l_3}) g_j(Z_{l_1}, Z_{l_2}, Z_{l_3}) g_k(Z_{l_1}, Z_{l_2}, Z_{l_3}).$$

The cardinality of $T_{\leq}^3$ is obtained by summing up these values for all zone combinations $(Z_{l_1}, Z_{l_2}, Z_{l_3})$ from Table 6. Using (16) one gets the cardinality of T^3. For the Finch data, this cardinality equals 2441. Hence, the complexity of calculating $\hat{\kappa}_{13}^2$ and $\hat{\kappa}_{23}^2$ is reduced to 29,792,405 operations. A similar argument may be applied to $T_{\leq}^4$, and all the other models we consider in this paper.