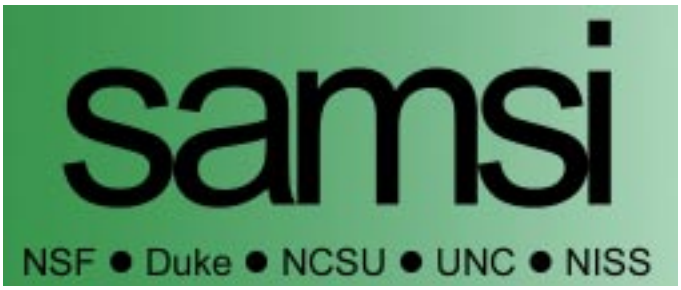


Transform methods for the hypergeometric distribution

Ian Dinwoodie, Laura Felicia Matusevich and Ed Mosteig

Technical Report #2003-15
November 4, 2003

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

Transform methods for the hypergeometric distribution

Ian H. Dinwoodie
ISDS, Duke University

Laura Felicia Matusevich
Department of Mathematics, UC Berkeley

Ed Mosteig
Department of Mathematics, Loyola Marymount University

November 4, 2003

Abstract

Two new methods for computing with hypergeometric distributions on lattice points are presented. One uses Fourier analysis, and the other uses Gröbner bases in the Weyl algebra. Both are very general and apply to log-linear models that are graphical or non-graphical.

Key Words. Contingency table, exact conditional test, exponential family, Fourier analysis, Gröbner basis, hypergeometric distribution, log-linear family, Monte Carlo method
Running Title. Transform Methods.

1. Introduction.

Some statistical questions about multidimensional contingency tables have a common mathematical formulation in terms of the distribution of a linear function on a collection of nonnegative integer lattice points. For example, one may be interested in bounds on cell entries on the collection of tables with certain linear constraints, or one may want to compute p-values of parameter significance in a conditional statistical test for a log-linear model. While the problems are combinatorial in an exact sense, practical statistical methods are based on Normal approximations, "perfect" Monte Carlo sampling, Markov Monte Carlo sampling, and, in some small cases, enumeration of lattice points. The reason that new methods are needed is that rapidly converging, simple Markov chains are only available in special cases.

Our purpose here is to develop two methods involving transforms in a broad sense. One method uses a classical discrete multidimensional Fourier transform where the "indeterminates" are orthogonal basis vectors that are indexed by the values of constraint functions, and which behave like dual variables to the original variables of integer counts. (Previous uses of the Fourier transform are in Good, Gover, and Mitchell (1987) and Baglivo, Olivier, and Pagano (1992)). This method is similar to the one in Dinwoodie (1998) where secondary indeterminates were used to keep track of linear constraints, but the method here is more general and requires far less computer memory because calculations are done numerically rather than symbolically. Second, we describe a new method from the theory of A-hypergeometric systems of linear partial differential equations. This method does Gröbner basis computations in the Weyl algebra and manipulates differential equations as much as possible rather than their generating function solutions. The computational efficiency is not clear, as the Gröbner basis calculations that replace enumeration of lattice points challenge current software. It does work well in some examples. To use this method, one needs up-to-date software implementing the algorithms of Saito, Sturmfels, and Takayama (2001), such as Macaulay 2 (see Eisenbud *et al.* 2002). This software is recent, so there is the possibility of significant computational improvements.

Let A be an $(r-1)\times c$ nonnegative integer matrix with columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_c$, let β be an $(r-1)\times 1$ column of nonnegative integers and let β_0 be a positive integer. For a $c\times 1$ vector $\mathbf{n} = (n_1, n_2, \ldots, n_c)'$ of nonnegative integers, let $\mathbf{n}_\bullet$ denote its sum $\mathbf{n}_\bullet := \sum_{i=1}^{c} n_i$.

Let $\mathbf{v} = (v_1, v_2, \ldots, v_c)'$ be a $c \times 1$ nonnegative integer vector. Let $S := \{\mathbf{n} \in Z_+^c : A\mathbf{n} = \beta, \mathbf{n}_\bullet = \beta_0\} \subset \Delta_{\beta_0} := \{\mathbf{n} \in Z_+^c : \mathbf{n}_\bullet = \beta_0\}$. One interesting computation is

$$s_k := \sum_{\substack{\mathbf{n}\in S \\ \mathbf{v}\cdot\mathbf{n}\geq k}} \frac{1}{\mathbf{n}!}$$

where $\mathbf{n}! := n_1! \times \cdots \times n_c!$ and $\mathbf{v}\cdot\mathbf{n} = \mathbf{v}'\mathbf{n}$. We are interested in this calculation because s_k/s_0 is the exact conditional p-value for a test of parameter significance in a log-linear

model with null-family sufficient statistics given by the rows of A and observed values given by β.

The quantity s_k/s_0 can be interpreted as the conditional probability $P(\mathbf{v}\cdot X \geq k \mid AX = \beta)$ when X has a multinomial $(\beta_0, \mathbf{p})$ distribution for $\mathbf{p} = (p_1, \ldots, p_c)$ any c-vector of probabilities that satisfy $\mathbf{p}^{\mathbf{a}} - \mathbf{p}^{\mathbf{b}} = 0$ for all nonnegative integer vectors $\mathbf{a}, \mathbf{b}$ such that $A\mathbf{a} = A\mathbf{b}$. This is the closure of the exponential family of multinomial probability distributions parametrized with a real r-vector θ by

$$(p_1, \ldots, p_c) = (e^{\theta A\mathbf{e}_1}, \ldots, e^{\theta A\mathbf{e}_c})/z_\theta \tag{1.1}$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_c$ are the standard basis vectors and z_θ is a normalizing constant. (The parameters will not be identifiable if A is not of rank r so other constraints may be imposed.)

It can also be interpreted as the conditional probability $P(\mathbf{v}\cdot X \geq k \mid AX = \beta, \sum_{i=1}^c X_i = \beta_0)$ when X has a Poisson(λ) distribution for $\lambda = (\lambda_1, \ldots, \lambda_c)$ any c-vector of rates that satisfy $\lambda^{\mathbf{a}} - \lambda^{\mathbf{b}} = 0$ for all nonnegative integer vectors $\mathbf{a}, \mathbf{b}$ such that $A\mathbf{a} = A\mathbf{b}$. However, neither of these interpretations leads to useful Monte Carlo methods in general, because the probability of the conditioning event is usually small. Closely related to the quantities s_k are bounds on values $\mathbf{v}\cdot\mathbf{n}$ over the set S, often specialized to the case of bounds on entries of contingency tables when $\mathbf{v}$ is a standard basis vector. We treat this case in §3 with examples that come from both decomposable and indecomposable graphical models (see Dobra and Fienberg (2000) for results on the decomposable case).

2. Fourier transforms from multivariate formulas.

To find s_k, consider the univariate Fourier transform $f(t)$ of $\mathbf{v}\cdot\mathbf{n}$ on the set of nonnegative integer lattice points $S := \{\mathbf{n} \in Z_+^c : A\mathbf{n} = \beta, \mathbf{n}_\bullet = \beta_0\} \subset \Delta_{\beta_0} := \{\mathbf{n} \in Z_+^c : \mathbf{n}_\bullet = \beta_0\}$ with hypergeometric weights:

$$f(t) := \sum_{\mathbf{n}\in S} \frac{e^{2\pi i \mathbf{v}\cdot\mathbf{n} t/b}}{\mathbf{n}!}$$

where the integer b satisfies $b > \max\{\mathbf{v}\cdot\mathbf{n} : \mathbf{n} \in \Delta_{\beta_0}\}$ (for example, $b = \beta_0 \times \max\{|v_i|\} + 1$ or the smallest power of 2 above this quantity for most efficient implementation of the fast Fourier transform). From the b values $\{f(t), t = 0, 1, \ldots, b-1\}$ one can recover the b coefficients

$$c_k = \sum_{\substack{\mathbf{n}\in S \\ \mathbf{v}\cdot\mathbf{n}=k}} \frac{1}{\mathbf{n}!},\ k = 0, 1, \ldots, b-1$$

or any function of them using basic properties of the discrete Fourier transform. Whereas $f(t)$ is defined in terms of an inverse image in Δ_{β_0}, $f(t)$ can be computed as the projection of a simple multivariate Fourier transform onto a one-dimensional subspace indexed by β. Let $b_i = \beta_0 \times \max\{A_{i,j} : j = 1, \ldots, c\} + 1,\ i = 1, \ldots, r-1$, and let $\mathbf{b} = (b_1, \ldots, b_{r-1})$. We will be working with discrete Fourier transforms of complex-valued functions on the r-dimensional domain of integer vectors

$$D := \prod_{i=1}^{r-1} \{0, 1, \ldots, b_i - 1\} \times \{0, 1, \ldots, b-1\}$$

with inner product

$$\langle f, g \rangle := \sum_{(s_1, s_2, \ldots, s_{r-1}, t) \in D} f(s_1, s_2, \ldots, s_{r-1}, t) \bar{g}(s_1, s_2, \ldots, s_{r-1}, t).$$

Each point in $(\beta, k) \in D$ indexes a basis element in the Hilbert space, so the dimension of the Hilbert space is actually $|D|$.

For simplicity below we will use the notation $\mathbf{a}/\mathbf{b} := (a_1/b_1, \ldots, a_{r-1}/b_{r-1})$ for vectors $\mathbf{a}$ and $\mathbf{b}$. For variables $(\mathbf{s}, t) = (s_1, s_2, \ldots, s_{r-1}, t) \in D$, define the r-variate function

$$\begin{aligned} g(\mathbf{s}, t) &:= \sum_{\mathbf{y} \geq 0} \sum_{\substack{A\mathbf{n} = \mathbf{y} \\ \mathbf{n}_\bullet = \beta_0}} \frac{e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_1/\mathbf{b}) n_1} \times \ldots \times e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_c/\mathbf{b}) n_c} e^{2\pi i t \mathbf{v} \cdot \mathbf{n}/b}}{\mathbf{n}!} \\ &= \frac{1}{\beta_0!} \big(e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_1/\mathbf{b}) + 2\pi i t v_1/b} + \ldots + e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_c/\mathbf{b}) + 2\pi i t v_c/b} \big)^{\beta_0}. \end{aligned} \tag{2.1}$$

The equality above comes from the multinomial expansion. That is, if one expands the final expression, the terms can be grouped by their value $\mathbf{y}$ of $A\mathbf{n}$, and the coefficient on each term is $\beta_0!/\mathbf{n}!$ before cancelling the $\beta_0!$ The basic result is that the univariate Fourier transform $f(t)$ that we seek is a component of $g(\mathbf{s}, t)$ at certain frequencies of the $\mathbf{s}$ variables, which can be found by summation over a product space rather than over the inverse image S. Define the functions $h_k, k = 0, 1, \ldots, b-1$ on D by

$$h_k(\mathbf{s}, t) = e^{2\pi i (\mathbf{s} \cdot (\beta/\mathbf{b}) + tk/b)}.$$

Theorem 2.1. The univariate Fourier transform $f(t)$ can be computed from g with the identity $f(t) = \frac{1}{|D|} \sum_{k=0}^{b-1} \langle g, h_k \rangle h_k(\mathbf{0}, t)$.

Remark. The probability generating function $f_p(z)$ of $\mathbf{v} \cdot \mathbf{n}$ is then given by

$$f_p(z) = \frac{1}{s_0 \, |D|} \sum_{k=0}^{b-1} \langle g, h_k \rangle z^k.$$

Proof. Let B be the $(r-1)\times c$ matrix given by

$$B = \begin{pmatrix} 1/b_1 & 1/b_2 & \dots & 1/b_{r-1} \\ 1/b_1 & 1/b_2 & \dots & 1/b_{r-1} \\ \dots & \dots & \dots & \dots \\ 1/b_1 & 1/b_2 & \dots & 1/b_{r-1} \end{pmatrix}.$$

$$\begin{aligned}
\langle g, h_k\rangle &= \sum_{\mathbf{s},t}\sum_{\mathbf{y}\geq\mathbf{0}}\sum_{\substack{A\mathbf{n}=\mathbf{y}\\ \mathbf{n}_\bullet=\beta_0}} \frac{e^{2\pi i(\mathbf{s}\cdot(\mathbf{a}_1/\mathbf{b})n_1+\ldots+\mathbf{s}\cdot(\mathbf{a}_c/\mathbf{b})n_c+t\mathbf{v}\cdot\mathbf{n}/\mathbf{b})}}{\mathbf{n}!} e^{-2\pi i(\mathbf{s}\cdot(\beta/\mathbf{b})+tk/b)} \\
&= \sum_{\mathbf{y}\geq 0}\sum_{\mathbf{s},t}\sum_{\substack{A\mathbf{n}=\mathbf{y}\\ \mathbf{n}_\bullet=\beta_0}} \frac{e^{2\pi i(\mathbf{s}\cdot BA\cdot\mathbf{n}+t\mathbf{v}\cdot\mathbf{n}/b)}}{\mathbf{n}!}\ e^{-2\pi i(\mathbf{s}\cdot(\beta/\mathbf{b})+tk/b)} \\
&= \sum_{\mathbf{y}\geq 0}\sum_{\mathbf{s}} e^{2\pi i\mathbf{s}\cdot B\mathbf{y}} e^{-2\pi i\mathbf{s}\cdot\beta/\mathbf{b}} \sum_{t=0}^{b-1}\sum_{\substack{A\mathbf{n}=\mathbf{y}\\ \mathbf{n}_\bullet=\beta_0}} \frac{e^{2\pi it\mathbf{v}\cdot\mathbf{n}/b}}{\mathbf{n}!} e^{-2\pi itk/b} \\
&= b_1b_2\ldots b_{r-1}\sum_{t=0}^{b-1}\sum_{\substack{A\mathbf{n}=\beta\\ \mathbf{n}_\bullet=\beta_0}} \frac{e^{2\pi it\mathbf{v}\cdot\mathbf{n}/b}}{\mathbf{n}!}\ e^{-2\pi itk/b} \\
&= b_1b_2\ldots b_{r-1}b\sum_{\substack{A\mathbf{n}=\beta\\ \mathbf{n}_\bullet=\beta_0\\ \mathbf{v}\cdot\mathbf{n}=k}} \frac{1}{\mathbf{n}!} = |D|\, c_k.
\end{aligned}$$

Finally, $h_k(\mathbf{0},t) = e^{2\pi itk/b}$, which proves the identity.

Example 2.1. Consider the set S of 2×2 tables with row sums of 2 and column sums of 2 and cells labelled left to right from the top row down. There are four cells, so the number of columns in the constraint matrix is four. For the statistical model of independence of row and column factors, the row and column sums are the two constraints from sufficient statistics that give the two rows of the constraint matrix A. The total count is $\beta_0 = 4$, and $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}$, $\beta = (\,2,2\,)'$. To get the marginal distribution of entry $(1,1)$ use $\mathbf{v} = (1,0,0,0)$. There are three tables consistent with row sums of 2 and column sums of 2, with probability 1/6 for $\begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$ where $\mathbf{v}\cdot\mathbf{n}$ takes the value 0, probability 4/6 for $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ where $\mathbf{v}\cdot\mathbf{n}$ takes the value 1, and probability 1/6 for the table on which $\mathbf{v}\cdot\mathbf{n} = 2$.

For Fourier analysis, take $b = 5$ and then

$$\begin{aligned}
f(t) &= \frac{1}{4} + e^{2\pi it/5} + \frac{e^{2\pi it2/5}}{4} \\
g(s_1,s_2,t) &= \frac{1}{4!}\big(e^{2\pi i((s_1+s_2)/5+t/5)} + e^{2\pi is_1/5} + e^{2\pi is_2/5} + 1\big)^4.
\end{aligned}$$

The coefficients on $h_k, k = 0, 1, 2, 3, 4$ are

$$\langle g, h_k \rangle = \frac{1}{5^3} \sum_{s_1=0}^{4} \sum_{s_2=0}^{4} \sum_{t=0}^{4} g(s_1, s_2, t) e^{-2\pi i((2s_1+2s_2)/5+tk/5)}$$

which yield the values $1/4, 1, 1/4$.

An extension of Theorem 2.1 is a computational advantage in most examples. Suppose the coordinates are partitioned into m subsets $R_1 = \{1, 2, \ldots, \alpha_1\}, R_2 = \{\alpha_1 + 1, \ldots, \alpha_2\}, \ldots, R_m = \{\alpha_{m-1} + 1, \ldots, \alpha_m = c\}$, which may be thought of as grouping the table cells into m rows, and let

$$S_\alpha := \{\mathbf{n} \in Z_+^c : A\mathbf{n} = \beta, \sum_{i=1}^{\alpha_1} n_i = \beta_{0,1}, \ldots, \sum_{i=\alpha_{m-1}}^{c} n_i = \beta_{0,m}\}. \tag{2.2}$$

for nonnegative integers $\beta_{0,i}$, and define $\beta_0 = \beta_{0,1} + \ldots + \beta_{0,m}$. We are interested in

$$f(t) := \sum_{\mathbf{n} \in S_\alpha} \frac{e^{2\pi i \mathbf{v} \cdot \mathbf{n} t/b}}{\mathbf{n}!}.$$

Let $\mathbf{b} = (b_1, \ldots, b_{r-1})$ satisfy $A\mathbf{n} < \mathbf{b}$ (strict inequality in each coordinate) on $\{\mathbf{n} \in Z_+^c : \sum_{i=1}^{\alpha_1} n_i = \beta_{0,1}, \ldots, \sum_{i=\alpha_{m-1}}^{c} n_i = \beta_{0,m}\}$, and let $\mathbf{v} \cdot \mathbf{n} < b$ on the same set. Note that the positive integers $\mathbf{b}$ and b depend now on $\alpha_1, \ldots, \alpha_m$ as they are defined over the subset of Δ_{β_0} that depends on these quantities. With $D_\alpha := \prod_{i=1}^{r-1} \{0, 1, \ldots, b_i - 1\} \times \{0, 1, \ldots, b-1\}$, define

$$\begin{aligned} g_\alpha(\mathbf{s}, t) :=& \frac{1}{\beta_{0,1}!} \left(e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_1/\mathbf{b}) + 2\pi i t v_1/b} + \ldots + e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_{\alpha_1}/\mathbf{b}) + 2\pi i t v_c/b}\right)^{\beta_{0,1}} \times \ldots \\ & \ldots \times \frac{1}{\beta_{0,m}!} \left(e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_{\alpha_{m-1}+1}/\mathbf{b}) + 2\pi i t v_c/b} + \ldots + e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_c/\mathbf{b}) + 2\pi i t v_c/b}\right)^{\beta_{0,m}}. \end{aligned} \tag{2.3}$$

Theorem 2.2. $f(t) = \frac{1}{|D_\alpha|} \sum_{k=0}^{b-1} \langle g_\alpha, h_k \rangle h_k(\mathbf{0}, t)$.

The proof is similar to Theorem 2.1.

Example 2.2. Consider a 2×2 table with row sums of 2 and column sums of 2 and cells labelled left to right from the top row down. Then $R_1 = \{1, 2\}, R_2 = \{3, 4\}, \beta_{0,1} = 2, \beta_{0,2} = 2, \mathbf{b} = (3 \quad 3)', b = 3, \beta_0 = 4, A = (1 \quad 0 \quad 1 \quad 0), \beta = (2, 2)'$. To get the marginal distribution of entry $(1, 1)$ use $\mathbf{v} = (1, 0, 0, 0)$. Then

$$g_\alpha(s, t) = (e^{2\pi i s/3 + t/3} + 1)^2 (e^{2\pi i s/3} + 1)^2$$

and the inversion Theorem 2.2 gives the coefficients $1/4, 1, 1/4$.

Computing the coefficients $\frac{1}{|D|}\langle g, h_k\rangle$ can be done many ways but some are better than others. The first obvious problem is one of scaling from the factorial terms in the denominator. The computation is a delicate numerical challenge for problems of interesting size. Second, it is a good idea to remove redundant constraints in A, because they increase the dimension of the variable $\mathbf{s}$. The dimension of $\mathbf{s}$ is the number of rows of A. In Example 2.1 this was two, but the example could have been presented with four rows corresponding to the two row sums and the two column sums of the original 2×2 table. This would have been valid but inefficient. Third, one can effectively reduce the dimension by one by using the fast fourier transform in the t variable. Rewrite the inner product

$$\begin{aligned}\frac{1}{|D|}\langle g, h_k\rangle &= \frac{1}{b}\,\frac{1}{\prod_{i=1}^{r-1} b_i}\sum_{\mathbf{s}} e^{-2\pi i\beta\cdot(\mathbf{s}/\mathbf{b})}\sum_{t=0}^{b-1} e^{-2\pi i t\,k/b} g(\mathbf{s},t)\\ &= \frac{1}{b}E_{\mathbf{s}}\big(e^{-2\pi i\beta\cdot(\mathbf{s}/\mathbf{b})}\hat{g}_{\mathbf{s}}(k)\big)\end{aligned}$$

where $\hat{g}_{\mathbf{s}}(k)$ is the discrete Fourier transform of $g(\mathbf{s}, t)$ (inner product with $(e^{2\pi i t\,k/b}, k = 0, \ldots, b-1)$) and the expectation is with respect to the uniform distribution on the product space $\Omega := \prod_{i=1}^{r-1}\{0, 1, \ldots, b_i - 1\}$. If b is taken to be a power of 2, then the expectation above can be computed with a Monte Carlo method by sampling $\mathbf{s}$ uniformly from Ω, and then using the fast Fourier transform to get the full sequence of coefficients $(\hat{g}_{\mathbf{s}}(k), k = 0, \ldots, b-1)$ for this $\mathbf{s}$.

In applications it is often impossible to calculate exhaustively through all points in D. Going through the t variable deterministically with the fast Fourier transform but randomly sampling the $\mathbf{s}$ variable uniformly as described above is an obvious compromise. However, there are some curious numerical challenges with this Monte Carlo scheme. The "Gibbs phenomenon" in the approximation can be significant and can lead to bad approximations. For example, tail probabilities s_k/s_0 when k is located in a "downward overshoot" are underestimated. Examples show a lot of variability in the size of the sample as a fraction of $|D|$ that is necessary for an accurate approximation.

Example 2.3. Consider data from Haberman (1978, p. 312) on attitude toward "women staying at home" for Male and Female respondents, with covariate the number of years of education which we restrict to the six levels 0 years through 5 years. Now, we are interested in the significance of an odds ratio parameter γ which quantifies the logarithm of the ratio of the probabilities of Female to Male for the "yes" outcome, so the roles of the sexes are reversed for the odds ratio. The data indicates that the ratio could be positive.

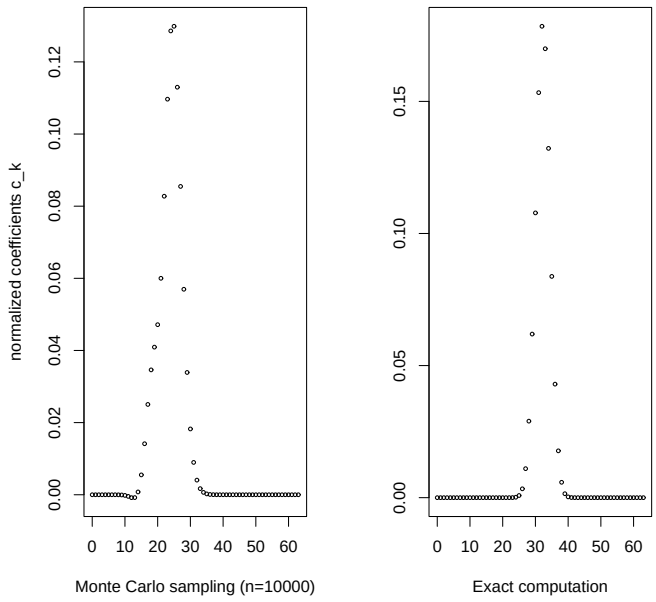
The $c = 2 \times 6 \times 2 = 24$-dimensional data is summarized in the table below. The "years" covariate is labelled "0 1 2 3 4 5" across the top:

$$\begin{array}{r} \texttt{Years:} \\ \texttt{Male yes} \\ \texttt{Male no} \\ \texttt{Female yes} \\ \texttt{Female no} \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 4 & 2 & 4 & 6 & 5 & 13 \\ 2 & 0 & 0 & 3 & 5 & 7 \\ 4 & 1 & 0 & 6 & 10 & 14 \\ 2 & 0 & 0 & 1 & 0 & 7 \end{pmatrix} \end{array}.$$

The set S consists of tables (elements of Z_+^{24}) with a total count $\beta_0 = 96$ and with the same number of Females as the given table at each covariate level, the same number of Males as the given table at each covariate level, the same total number of "yes" responses over both Male and Female, and the same regression statistic $0\times$ (#yes outcomes at year 0) + $1\times$ (#yes at year 1) + $\ldots$ + $5\times$ (#yes at year 5). These constraints give the matrix A of size 14×24. We keep track of the vector $\mathbf{v}$ which counts the number of 1's in the Female dimensions:

$$\begin{array}{r} \texttt{Years:} \\ \texttt{Male yes} \\ \texttt{Male no} \\ \texttt{Female yes} \\ \texttt{Female no} \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

In the original column vector notation this would be twelve 0's, six 1's, six 0's. The generalized linear model fitted in the language R gives a p-value of 0.103 on the standard Normal scale for the estimate of the odds ratio parameter comparing the Female "yes" rate to the Male "yes" rate with a common slope parameter. The exact Fourier analysis method of §2 gives the value $s_{34}/s_0 = .284323$ reported in Dinwoodie (1998) as the p-value. However, the p-value should be $s_{35}/s_0 = .152108$, so the fraction in Dinwoodie (1998) wrongly gave the tail probability from 34 rather than 35. The figure below compares the results from Monte Carlo sampling of 10,000 uniformly chosen points $\mathbf{s}$, versus exhaustive computation over all 33,465 points. The notorious Gibbs overshoot/undershoot is evident and can be significant with smaller sample sizes.



Example 2.4. Consider the three-chain $\bullet - - \bullet - - \bullet$ decomposable graphical model with levels 2, 6, and 2 at the three variables. We will use this to model data on the relationship between Salk vaccine, Age, and Paralysis from Agresti (p. 256, 1990) in that order, meaning a model of independence of Salk Vaccine and Paralysis, given Age. The data is a $c = 2 \times 6 \times 2$ cube with sample size $\beta_0 = 174$ given below with 6 levels 0-4, 5-9, 10-14, 15-19, 20-39, 40+ of the Age variable from left to right:

$$\begin{array}{c} \\ \texttt{Salk Yes} \\ \texttt{Salk No} \end{array} \begin{array}{c} \begin{array}{cc|cc|cc|cc|cc|cc} \texttt{No} & \texttt{Yes} & \texttt{No} & \texttt{Yes} & \texttt{No} & \texttt{Yes} & \texttt{No} & \texttt{Yes} & \texttt{No} & \texttt{Yes} & \texttt{No} & \texttt{Yes} \end{array} \\ \left(\begin{array}{cc|cc|cc|cc|cc|cc} 20 & 14 & 15 & 12 & 3 & 2 & 7 & 4 & 12 & 3 & 1 & 0 \\ 10 & 24 & 3 & 15 & 3 & 2 & 1 & 6 & 7 & 5 & 3 & 2 \end{array}\right) \end{array}$$

We are interested in the distribution of the sum of the number of counts across the Age variable at the levels of Salk "No", and Paralysis "Yes". The observed value of this statistic is $54 = 24 + 15 + 2 + 6 + 5 + 2$, the entrywise scalar product of the observed table with the vector $\mathbf{v} := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$, whereas its expected value based on the fitted independence model is 40.9. We should compute the conditional probability of 54 or greater of the statistic $\mathbf{v} \cdot \mathbf{n}$, fixing the sufficient statistics. One can see that the distribution is a convolution of six hypergeometric distributions, so exact computations are quite easy for comparison purposes. The p-value is 0.00163 and the Monte Carlo Fourier analysis confirms this.

Let us mention two simulation methods for this example which would be useful for goodness-of-fit tests and which also put in perspective the advantages of the Fourier analysis. Consider the collection of sequences $a_1, \ldots, a_{174}$ where each $a_i \in \{1, 2, \ldots, 24\}$ codes

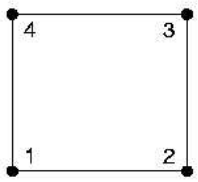
a box in the table. The uniform distribution on such sequences, constrained so that their summary tables satisfy the constraints on the sufficient statistics, maps to the hypergeometric distribution on the constrained tables. Therefore one may attempt to use the Gibbs sampler on the 174 coordinates (see Liu (2001)). Since the joint distribution under the constraints is not positive on all sequences in the product space $\{1, 2, \ldots, 24\}^{174}$, the Gibbs sampler is not irreducible, and this method will not work. We are lead to the "big-fiber walk" of Diaconis and Sturmfels (1998) that uses a degree bound on a Gröbner basis in order to get irreducibility. Basically, one has to update several coordinates simultaneously, say d, and the value of d is related to the degree of polynomials in a Gröbner basis of a toric ideal.

First, an irreducible, symmetric Markov chain on the constrained tables (those with fixed sufficient statistics $\mathbf{n}_{\{1,2\}}, \mathbf{n}_{\{2,3\}}$) is described by Dobra (2002) and has $\binom{2}{2} \times 6 \times \binom{2}{2} = 6$ moves, one for each level of the separating variable Age. Each of these moves adds 1 to two cells entries, and subtracts 1 from two others. So computations can be done with the uniform distribution on constrained tables with this Markov chain. Now the extension of this result in Geiger, Meek, and Sturmfels (2002) shows that a Gröbner basis for the toric ideal (see Kreuzer and Robbiano 2000 for definitions) of the decomposable model is quadratic, so one can do the "big-fiber walk" with $d = 2$ as follows. Make an initial sequence of length 174 composed of 24 box labels as above, with 20 "1"s, 14 "2"s, 10 "3"s, ..., 2 "24"s, which corresponds to a particular outcome in a multinomial sample space consistent with the observed table. Then choose an unordered pair of coordinates uniformly at random, each with probability $1/\binom{174}{2}$. If the box labels in the two chosen coordinates are in different Age levels, swap the two labels. If the box labels in the two chosen coordinates are in the same Age level and in the same row or column of that Age level, swap the two labels. If the box labels in the two chosen coordinates are in the same Age level and in a diagonal configuration in that Age level, then replace the labels with a random choice of four new label assignments: the original assignment, swap the original assignment, or replace the original labels with those of the other two box labels in that Age level in any order. This results in an irreducible and symmetric Markov chain on constrained sequences, so the corresponding summary tables have the hypergeometric distribution. Of course, the easiest way to simulate this example is to independently fill in the six 2×2 tables with the hypergeometric distribution fixing the 2×2 row and column sums, but the big-fiber walk works with $d = 2$ for any decomposable graphical model.

3. Marginal Conditional Distributions.

Now we are interested in the distribution of a single table entry under the hypergeometric distribution with constraints. Such distributions also give bounds on entries under the given constraints (see Dobra and Fienberg (2000) for recent results on decomposable models). This is a special case of the previous problem with $\mathbf{v}$ set equal to a standard basis vector (a single nonzero entry of 1). The specialization allows a new approach which is the use of A-hypergeometric systems. First we do an example of the Fourier analysis method for an indecomposable graphical model.

Example 3.1. Here is an example of computations for a graphical indecomposable model, the square. We model data from Agresti (p. 308, 1990) on alligator counts, with four factors: Lake (variable 1 at levels Hancock and Oklawaha), Gender (variable 2 at two levels), Size (variable 3 at two levels, ≤ 2.3 meters and > 2.3 meters), and Primary Food (variable 4 at five levels: Fish, Invertebrate, Reptile, Bird, Other), with an indecomposable graphical model:

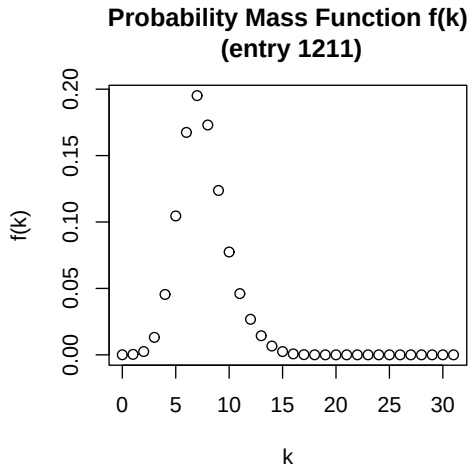


implying sufficient statistics $\mathbf{n}_{\{1,2\}}, \mathbf{n}_{\{2,3\}}, \mathbf{n}_{\{3,4\}}, \mathbf{n}_{\{4,1\}}$.

The four-way data can be summarized in the following table:

$$\begin{array}{c} \\ \mathbf{n}_{111l} \\ \mathbf{n}_{112l} \\ \mathbf{n}_{121l} \\ \mathbf{n}_{122l} \\ \mathbf{n}_{211l} \\ \mathbf{n}_{212l} \\ \mathbf{n}_{221l} \\ \mathbf{n}_{222l} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{n}_{ijk1} & \mathbf{n}_{ijk2} & \mathbf{n}_{ijk3} & \mathbf{n}_{ijk4} & \mathbf{n}_{ijk5} \end{array} \\ \left(\begin{array}{ccccc} 7 & 1 & 0 & 0 & 5 \\ 4 & 0 & 0 & 1 & 2 \\ 16 & 3 & 2 & 2 & 3 \\ 3 & 0 & 1 & 2 & 3 \\ 2 & 2 & 0 & 0 & 1 \\ 13 & 7 & 6 & 0 & 0 \\ 3 & 9 & 1 & 0 & 2 \\ 0 & 1 & 0 & 1 & 0 \end{array} \right) \end{array}$$

Our question is "What is the probability of 16 or more in box 1211, fixing the sufficient statistics?". The method of Theorem 2.2 based on sample sizes of 5,000,000 gives an answer of 0.001.



We now introduce an algebraic method based on A-hypergeometric systems and D-modules for studying the distribution of a single table entry.

D-modules are modules in the algebra of differential and multiplication operators on polynomials in several variables (Coutinho 2000). The connection between D-modules, hypergeometric functions, and solutions to certain systems of linear partial differential equations with polynomial coefficients is developed systematically in Saito, Sturmfels, and Takayama (2001) with recent results in Matusevich (2002). The part of this theory that is needed for our application is small.

Recall the set $S = \{\mathbf{n} \in Z_+^c : A\mathbf{n} = \beta, \mathbf{n}_\bullet = \beta_0\}$, which is a finite set since $\mathbf{n}$ is a bounded nonnegative integer vector. Consider the exponential generating function

$$g_e := \sum_{\mathbf{n}\in S} \frac{\mathbf{x}^{\mathbf{n}}}{\mathbf{n}!}$$

in the ring $Q[x_1, \ldots, x_c]$, where $\mathbf{x}^{\mathbf{n}} = x_1^{n_1} \ldots x_c^{n_c}$. Although computing g_e is equivalent to enumerating S, computation of s_k, defined in §1 as certain sums over S, requires only one-dimensional information from g_e and may be done using Gröbner bases in the Weyl algebra.

The essential fact is that g_e is the unique polynomial solution (up to complex scalar coefficients) to the A-hypergeometric system $H_A(\beta)$ of linear partial differential equations:

$$(H_A(\beta)) \qquad \begin{aligned} &(\partial^{\mathbf{a}} - \partial^{\mathbf{b}})g_e = 0, \quad A\mathbf{a} = A\mathbf{b}, a_1 + \ldots + a_c = b_1 + \ldots + b_c, \\ &\sum_{j=1}^{c}(x_j\partial_j - \beta_0)g_e = 0, \\ &\sum_{j=1}^{c}(a_{ij}x_j\partial_j - \beta_i)g_e = 0, \quad i = 1, \ldots, r-1. \end{aligned}$$

This is Proposition 3.4.11 of Saito, Sturmfels, Takayama (p. 132, 2000). To clarify notation, $\partial^{\mathbf{a}}$ with $\mathbf{a} = (a_1, \ldots, a_c)$ is the differential operator on polynomials defined by $\partial^{\mathbf{a}}(p) = \partial_1^{a_1} \circ \cdots \circ \partial_c^{a_c}\, p$. The fact that g_e is a solution is easy to verify, but the uniqueness is deeper and is essential. The total number of independent solutions $\mathrm{rank}(H_A(\beta))$ to the system $H_A(\beta)$ is related to the geometry of A and is typically greater than 1. The rules for manipulating the indeterminates $x_1, \ldots, x_c, \partial_1, \ldots, \partial_c$ as operators on polynomial functions come from elementary calculus, and the noncommutative ring in these indeterminates with rational coefficients is denoted D_c. The quick summary is that $\partial_2 x_1 = x_1 \partial_2$, but $\partial_1 x_1 = 1 + x_1 \partial_1$ by the product rule.

We will only consider the case where the vector $\mathbf{v}$, whose values $\mathbf{v} \cdot \mathbf{n}$ over S are our main interest, is a basis vector, say $\mathbf{v} = (1, 0, 0, \ldots, 0)$. We want to find $g_1(x_1) := g_e(x_1, 1, 1, \ldots, 1)$. Let $I_{A,\beta}$ denote the left ideal in D_c generated by the equations of $H_A(\beta)$, which roughly speaking is the entire collection of equations that are implied by the given system. Then the set $I_{A,\beta} \cap \langle x_1, x_2, \ldots, x_c, \partial_1 \rangle$ contains the equations that involve only the derivative ∂_1.

Lemma 3.1. If $P \in I_{A,\beta} \cap \langle x_1, x_2, \ldots, x_c, \partial_1 \rangle$, define $P_1(x_1, \partial_1) = P/_{x_2=1, x_3=1, \ldots, x_c=1}$. Then $P_1(g_1) = 0$.

Proof. The elements of D_c that annihilate g_e form a left ideal, so $P(g_e) = 0$, the vanishing polynomial. Then a weaker statement is $P(g_e)(x_1, 1, \ldots, 1) = 0$, and since the operator P only takes derivatives in x_1, it follows that $0 = P(g_e)(x_1, 1, \ldots, 1) = P_1(g_e(x_1, 1, \ldots, 1))$.

The above result means that g_1 solves any differential equation in the ∂_1 variable that one can derive from $H_A(\beta)$. Finding such an equation is noncommutative elimination theory, explained in Saito, Sturmfels, and Takayama (2001) and implemented in Macaulay 2. The property that the system $H_A(\beta)$ is holonomic implies that one can rewrite the system $H_A(\beta)$ in the Weyl algebra D_c to eliminate variables $\partial_2, \ldots, \partial_c$. The triangular system has one final equation in the variables $x_1, \ldots, x_c, \partial_1$. One can set $x_2 = 1, \ldots, x_c = 1$, and solve it as an ordinary differential equation (ode). For example, consider the following Macaulay 2 code to get the exponential generating function $x_1^2/4 + x_1 + 1/4$ of the variable x_1 for the upper left box in a 2×2 table with row sums and column sums of 2:

```
load "D-modules.m2"
I=gkz(matrix{{1,1,1,1},{1,1,0,0},{1,0,1,0}},{4,2,2})
GB=gens gbw(I,{0,0,0,0,0,1,1,1})
p=last last entries GB
p1=substitute(p,{x_2=>1,x_3=>1,x_4=>1})
R2=QQ[x_1,D_1,WeylAlgebra=>{x_1=>D_1}]
```

```
q=substitute(p1,R2)
use R2
a=PolySols(ideal(q))
```

The ode `q` is $-x_1^2\partial_1^2 + x_1\partial_1^2 + 3x_1\partial_1 + \partial_1 - 4 = 0$ which is solved as $x_1^2 + 4x_1 + 1$. From this, one can do Fisher's exact test.

There are two difficulties, one theoretical and one practical. The theoretical problem is that the final differential equation in ∂_1 may have more than one polynomial solution (only one of which extends up to a real solution for the entire system, which is the one we want). The practical problem is that the Gröbner basis calculation in D_c to eliminate differential variables can be slow with many variables.

Example 3.2. Consider the following death penalty data from Agresti (p. 136, 1990), where `yes` means there was a death penalty and `no` means there was not.
White Defendent:

$$\begin{array}{r} \\ \texttt{white victim} \\ \texttt{black victim} \end{array} \begin{array}{c} \begin{array}{cc} \texttt{yes} & \texttt{no} \end{array} \\ \begin{pmatrix} 19 & 132 \\ 0 & 9 \end{pmatrix} \end{array}$$

Black Defendent:

$$\begin{array}{r} \\ \texttt{white victim} \\ \texttt{black victim} \end{array} \begin{array}{c} \begin{array}{cc} \texttt{yes} & \texttt{no} \end{array} \\ \begin{pmatrix} 11 & 52 \\ 6 & 97 \end{pmatrix} \end{array}$$

This is an eight-cell table, so the calculations would be done in the D-module D_8. Our question is to find the probability of 11 or higher in box 5 (numbered left to right row-by-row), fixing the total number of White defendents, the total number of Black defendents, the total number of `yes` outcomes, the total number of `white/white`, `white/black`, `black/white`, and `black/black` events. The exact answer is 0.0612. If the factors are ordered Defendent, Victim, Death Penalty, then the constraints correspond to fixing statistics $\mathbf{n}_{\{1,2\}}, \mathbf{n}_{\{3\}}$ in common marginal notation.

To do the calculation, consider the A-hypergeometric system $\texttt{I} = \texttt{gkz}(\texttt{A}, \beta)$ with

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\beta = \begin{pmatrix} 326 & 160 & 166 & 36 & 151 & 9 & 63 & 103 \end{pmatrix}'$$

Get a Gröbner basis with a weighted order `gbw(I, {0,0,0,0,0,0,0,0,1,1,1,1,0,1,1,1})` to eliminate ∂_5. The element of the Gröbner basis in the ∂_5 variable is $p = x_2\,x_4\,x_5^3\,x_6\,x_7\,\partial_5^4 - x_2\,x_3\,x_5^2\,x_6^2\,x_7\,\partial_5^4 - x_1\,x_4\,x_5^2\,x_6^2\,x_7\,\partial_5^4 + x_1\,x_3\,x_5\,x_6^3\,x_7\,\partial_5^4 - x_2\,x_4\,x_5^4\,x_8\,\partial_5^4 + x_2\,x_3\,x_5^3\,x_6\,x_8\,\partial_5^4 + x_1\,x_4\,x_5^3\,x_6\,x_8\,\partial_5^4 - x_1\,x_3\,x_5^2\,x_6^2\,x_8\,\partial_5^4 - 52\,x_2\,x_4\,x_5^2\,x_6\,x_7\,\partial_5^3 - 18\,x_2\,x_3\,x_5\,x_6^2\,x_7\,\partial_5^3 - 160\,x_1\,x_4\,x_5\,x_6^2\,x_7\,\partial_5^3 + 230\,x_1\,x_3\,x_6^3\,x_7\,\partial_5^3 + 216\,x_2\,x_4\,x_5^3\,x_8\,\partial_5^3 - 146\,x_2\,x_3\,x_5^2\,x_6\,x_8\,\partial_5^3 - 4\,x_1\,x_4\,x_5^2\,x_6\,x_8\,\partial_5^3 - 66\,x_1\,x_3\,x_5\,x_6^2\,x_8\,\partial_5^3 - 4636\,x_2\,x_4\,x_5\,x_6\,x_7\,\partial_5^2 + 4758\,x_2\,x_3\,x_6^2\,x_7\,\partial_5^2 + 13420\,x_1\,x_4\,x_6^2\,x_7\,\partial_5^2 - 17568\,x_2\,x_4\,x_5^2\,x_8\,\partial_5^2 + 6832\,x_2\,x_3\,x_5\,x_6\,x_8\,\partial_5^2 - 10492\,x_1\,x_4\,x_5\,x_6\,x_8\,\partial_5^2 + 7686\,x_1\,x_3\,x_6^2\,x_8\,\partial_5^2 + 257176\,x_2\,x_4\,x_6\,x_7\,\partial_5 + 635376\,x_2\,x_4\,x_5\,x_8\,\partial_5 - 98332\,x_2\,x_3\,x_6\,x_8\,\partial_5 + 438712\,x_1\,x_4\,x_6\,x_8\,\partial_5 - 8577576\,x_2\,x_4\,x_8$.

Setting all variables but x_5, ∂_5 equal to 1 and solving in a new ring gives three polynomial solutions, say p_{36}, p_{62}, p_{63} of degrees 36, 62, and 63 respectively. Our polynomial generating function must be a linear combination of these three polynomials. But since its degree is at most 36, it must be a multiple of p_{36}. The coefficients on $x^{36}, x^{35}, \ldots, x^0$ start like `1 2367/7 1550385/29 5271309` ... The probability we seek is the sum of coefficients on $x^{36}, \ldots, x^{11}$ divided by the sum of all of them, which works out to 0.0612.

This method works very well when the number of cells in the table are no more than a dozen, even when some cell entries are large. The hardest part is searching through the Gröbner basis output from Macaulay 2. We have seen problems where "monomial overflow" prevents the completion of the calculation of the polynomial solutions. Valuable research could be done on understanding the possible polynomial solutions to the final equation in the Gröbner basis and classifying the difficulties one may encounter. In all the examples we have done, it has been possible to find the right solution by examination. In some examples, this method has the upper hand in terms of speed, simplicity and accuracy over alternatives like Monte Carlo methods, particularly for the problem of bounds on entries where the tail probabilities in the hypergeometric distribution are small. Another advantage of this method is that many computations can be done parametrically. Consider, for instance, a 2×3 contingency table with row sums a and b, and column sums c, d, and $e = a+b-c-d$; that is, we consider the marginals as parameters. To compute g_1, consider the following matrix:

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix} .$$

The corresponding A-hypergeometric system with parameter $(a, b, c, d)'$ is

$$\begin{aligned}
(\partial_1\partial_5 - \partial_4\partial_2)g_e &= 0\ ,\\
(\partial_2\partial_6 - \partial_5\partial_3)g_e &= 0\ ,\\
(\partial_1\partial_6 - \partial_4\partial_3)g_e &= 0\ ,\\
(x_1\partial_1 + x_2\partial_2 + x_3\partial_3 - a)g_e &= 0\ ,\\
(x_4\partial_4 + x_5\partial_5 + x_6\partial_6 - b)g_e &= 0\ ,\\
(x_1\partial_1 + x_4\partial_4 - c)g_e &= 0\ ,\\
(x_2\partial_2 + x_5\partial_5 - d)g_e &= 0\ .
\end{aligned}$$

The following Maple script computes an elimination Gröbner basis of this system without specializing the parameters, selects the element which involves only ∂_1, converts this element into a differential equation. After this is done, we can specialize the parameters, and look for the rational function solutions of the resulting ordinary differential equation. These computations can also be done in Macaulay 2, but we present the Maple commands as alternates. The main advantage of Maple is that the element of the elimination Gröbner basis that we need is placed automatically in the first place of the output, so that we do not need to search for it.

```
> with(Groebner):
> with(Ore_algebra):
> with(DEtools):
> contingency:=d1*d5-d4*d2,d2*d6-d5*d3,d1*d6-d4*d3,
x1*d1+x2*d2+x3*d3-a,x4*d4+x5*d5+x6*d6-b,x1*d1+x4*d4-c,x2*d2+x5*d5-d:
> R:=diff_algebra([d1,x1],[d2,x2],[d3,x3],[d4,x4],[d5,x5],[d6,x6],
[da,a],[db,b],[dc,c],[dd,d]):
> G:=gbasis(contingency,termorder(R,lexdeg([d2,d3,d4,d5,d6],[d1]))):
> eqn1:=Ore_to_diff(G[1],f,R);
> eqn:=subs(f(a,b,x6,x5,x4,x3,x1,x2,c,d)=f(t),x1=t,x2=1,x3=1,x4=1,
x5=1,x6=1,eqn1);
> ex:=subs(a=30,b=30,c=20,d=20,eqn);
> L:=ratsols(ex,f(t)); sort(L[1]); sort(L[2]);
```

The ordinary differential equation `eqn` is:

$$\begin{aligned}
&(t^3 - 2t^2 + t)\partial_t^3 + [(4 - a - 2c)t^2 + (-6 + a - b + 3c)t + 2 + b - c]\partial_t^2 +\\
&[(2ac + c^2 - 2a - 3c + 2)t - ac + bc - c^2 + a - b + 3c - 2]\partial_t + (-ac^2 + ac)
\end{aligned}$$

If we use, as an example, $a = 30$, $b = 30$, $c = 20$, $d = 20$, and look for rational solutions, we find that our equation has two polynomial solutions, one of degree 20, and the other

of degree 19. It is clear that in this case, g_1 has degree 20, so that g_1 is a nontrivial linear combination of the two solutions of our ordinary differential equation. The coefficients of this linear combination cannot be found easily. It is necessary to compute two coefficients of g_1 by a different method, and then solve a simple system of linear equations. This is a drawback of the Gröbner bases method to compute g_1.

On the other hand, the fact that Gröbner computations can be done without specializing the parameters has two interesting consequences. First, there is a distinct advantage when we are interested in varying parameters. The parametric Gröbner computation is done only once, and the substitution is done at the level of ordinary differential equations. Finding polynomial solutions of these can be done efficiently, even if the coefficients are large.

The other consequence we should mention is that, for a given A, the complexity of the Gröbner computation is bounded independently of the size of the parameters, namely, the complexity of the parametric Gröbner basis is a bound. This is important since, in more naive methods for computing g_1 (such as simple enumeration), the complexity grows with the size of the parameters.

In the examples above, the distribution of the linear statistics $\mathbf{v} \cdot \mathbf{n}$ could be well approximated by a Normal density with a ".5" continuity correction, if one knew the exact mean and variance of $\mathbf{v} \cdot \mathbf{n}$ on the constraint set. Here is another example where this is true and where another transform technique can be used to get a formula. Fraser and Hunter (1975) published the following table of pairs of siblings with different types of congenital heart malformations. Exact calculations were done in Dinwoodie and MacGibbon (2001) using a simulation method for triangular tables with fixed row and column sums that goes back to Karl Pearson (see Stigler (1992)) and is described clearly in Diaconis, Graham, and Holmes (1999). Another simulation method is in McDonald and Smith (1995). The question for this example is "What is the probability of 19 or greater in entry (1,3)?"

Table 1. Distribution of pairs of siblings with unlike cardiac malformations-major lesion approach from Fraser and Hunter (1975)

	ToF	VSD	PS	TGV	PDA	AS	ASD	Tru	TA	CoA	Dex	Ptr	A − V	Total
ToF	−	13	19	10	4	1	1	0	1	0	1	2	0	52
VSD		−	3	5	3	3	6	1	0	0	2	1	0	24
PS			−	2	0	1	1	3	1	0	0	0	0	8
TGV				−	4	1	2	1	0	1	0	0	0	9
PDA					−	2	0	1	2	0	0	0	1	6
AS						−	2	0	1	3	2	0	0	8
ASD							−	0	1	1	0	0	1	3
Tru								−	0	0	0	1	0	1
TA									−	0	0	0	0	0
CoA										−	0	0	0	0
Dex											−	0	0	0
Ptr												−	0	0
A − V													−	−
Total	0	13	22	17	11	8	12	6	6	5	5	4	2	111

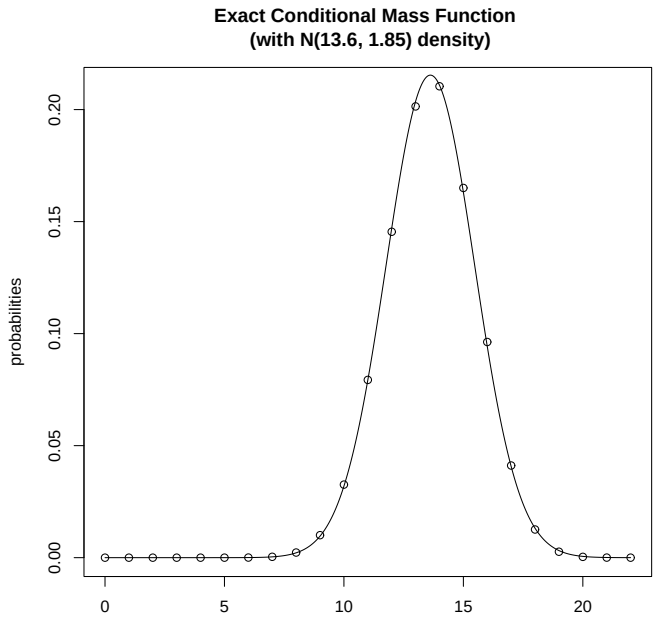
The Monte Carlo simulation found a value of .003. However, an exact formula is possible. Trim the table to an 11×11 triangle by removing the unnecessary first and last boxes, then use the following generating function:

$$g = (xy_1 + y_2 + \ldots + y_{11})^{r_1}(y_1 + \ldots + y_{11})^{r_2}(y_2 + \ldots + y_{11})^{r_3} \ldots (y_{10} + y_{11})^{r_{11}}$$

where $(r_1, \ldots, r_{11}) = (39, 24, 8, 9, 6, 8, 3, 1, 0, 0, 0)$. The y's code for column location. We want the coefficients a_i on $x^i y_1^{22} y_2^{17} \ldots y_{11}^2$. The triangular structure leads to the formula below for the coefficient on $x^{i_1} y_1^{i_1+i_2} y_2^{i_3} y_3^{i_4} \ldots y_{10}^{i_{11}} y_{11}^{98-i_1-i_2-\ldots-i_{10}-i_{11}}$:

$$a_i = \binom{r_1}{i_1}\binom{r_2}{c_1 - i_1}\binom{r_1 + r_2 + r_3 - i_1 - i_2}{c_2}\binom{r_1 + r_2 + r_3 + r_4 - i_1 - i_2 - i_3}{c_3} \ldots \ldots \binom{r_1 + \ldots + r_{11} - i_1 - \ldots - i_{10}}{c_{10}}$$

where $c = (22, 17, 11, \ldots, 2)$ are the column totals. Then the probability mass function is the collection of normalized a_i, pictured below. The probability of 19 or greater is in fact .003035790. With the normal density fitted with the same mean and variance as the exact pmf, the probability of 18.5 or greater is .0042, and the graph indicates that a Normal approximation should work well.



4. χ^2 Goodness of Fit.

Consider the problem of testing whether the exponential family (1.1) includes the multinomial probability distribution that was in effect in an experiment where β_0 trials were made independently, drawing from objects $1, 2, \ldots, c$ and with summary counts $\mathbf{n} := (n_1, \ldots, n_c)'$. Such a test can be done by measuring the distance from the normalized data $\mathbf{n}/\beta_0 \in \Delta_1$ to the model $V_A \cap \Delta_1$, where Δ_1 is the standard (nonnegative) probability simplex and $V_A := \{\mathbf{p} \in R^c : \mathbf{p^a} - \mathbf{p^b} = 0, A\mathbf{a} = A\mathbf{b}\}$. The strictly positive elements of $V_A \cap \Delta_1$ are parametrized with the exponential family at (1.1). $V_A \cap \Delta_1$ has dimension that we define in a complementary sense to be $c - 1 - d_f$ in terms of a nonnegative integer parameter d_f. Then $c - 1 - d_f$ is no greater than the rank of A (A is $(r-1) \times c$).

Let $\hat{n}_i = \beta_0 \hat{p}_i, i = 1, \ldots, c$ where $(\hat{p}_i) \in V_A \cap \Delta_1$ satisfies

$$(\hat{p}_i) := \texttt{arg max}\{\frac{\mathbf{p^n}}{\mathbf{n!}} : \mathbf{p} \in V_A \cap \Delta_1\},$$

so that $\hat{\mathbf{n}} := (\hat{n}_i)$ are the fitted cell counts (which sum to β_0 just like the components of $\mathbf{n}$). Iterative proportional fitting can be used to get maximum likelihood $\hat{n}_i$ without an identifiable parametrization in θ, but asymptotics require at least the number of free parameters (the dimension of $V_A \cap \Delta_1$, which is either the rank of A or one less).

The χ^2 statistic measures the distance between $\mathbf{n}/\beta_0$ and $\hat{p} \in V_A \cap \Delta_1$ as

$$\begin{aligned}\chi^2(\mathbf{n}) &:= \sum_{i=1}^{c} (n_i - \hat{n}_i)^2/\hat{n}_i \\ &= \sum_{i=1}^{c} n_i^2/\hat{n}_i \ - \beta_0 \\ &= \|\mathbf{n}\|^2_{\hat{N}^{-1}} - \beta_0\end{aligned}$$

where $\|\mathbf{n}\|^2_{\hat{N}^{-1}} := \mathbf{n}'\hat{N}^{-1}\mathbf{n}$ and $\hat{N} := \begin{pmatrix} \hat{n}_1 & 0 & 0 & \dots \\ 0 & \hat{n}_2 & 0 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$.

Under multinomial sampling of c objects $1, 2, \dots, c$ from a positive distribution $\mathbf{p}$ within the family (1.1), it can be proved that as the number of trials increases ($\beta_0 \to \infty$),

$$P_{\mathbf{p}}\{\mathbf{n} : \chi^2(\mathbf{n}) \geq x\} \to 1 - F_{d_f}(x)$$

where F_{d_f} is the cdf for the χ^2 distribution with d_f degrees of freedom. The exact conditional computation for the p-value is

$$P_{\mathbf{p}}\{\mathbf{n} : \chi^2(\mathbf{n}) \geq x \ | A\mathbf{n} = \beta\} = P_h\{\|\mathbf{n}\|_{\hat{N}^{-1}} \geq x + \beta_0\} \tag{4.1}$$

where P_h refers to the parameter-free hypergeometric distribution on $S := \{A\mathbf{n} = \beta, \mathbf{n}_\bullet = \beta_0$. The quantity (4.1) is an exact conditional p-value, which can be used for nonasymptotic tests that are uniform in size over the model $V_A \cap \Delta_1$.

To use the discrete Fourier transform it is convenient to approximate the matrix $\hat{N}$ with a rational matrix. If the statistical model is for a decomposable graphical model, then $\hat{\mathbf{n}}$ is already rational. Let

$$Q := \begin{pmatrix} m_1 & 0 & 0 & \dots \\ 0 & m_2 & 0 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix},$$

where $m_1, m_2, \dots, m_c$ are positive integers, satisfy $\hat{N}^{-1} \approx Q/m$ for a positive integer m. The basic quantities for computation are

$$S_x := \sum_{\substack{\mathbf{n} \in S \\ \|\mathbf{n}\|^2_Q \geq m(x+\beta_0)}} \frac{1}{\mathbf{n}!}. \tag{4.2}$$

Then the approximation to the p-value defined at (4.1) would be S_x/S_0, whose computation we describe below.

Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_\kappa$ be a basis of integer (column) vectors for $\texttt{ker} \begin{pmatrix} \mathbf{1} \\ A \end{pmatrix}$, and let A_0 be the integer matrix given by $A_0' = (\, \mathbf{w}_1 \quad \mathbf{w}_2 \quad \ldots \quad \mathbf{w}_\kappa\,)$. Define the inner product $\langle \mathbf{v}, \mathbf{w} \rangle := \mathbf{v}' Q \mathbf{w}$ on R^c. Then the rows of A_0 span the space orthogonal to the row space of AQ^{-1} with respect to the inner product with Q, because $(AQ^{-1})QA_0' = AA_0' = 0$.

Now $\|\mathbf{n}\|_Q^2 = \|P_{AQ^{-1}}(\mathbf{n})\|_Q^2 + \|P_{A_0}(\mathbf{n})\|_Q^2$ where $P_{AQ^{-1}}, P_{A_0}$ are the projection operators onto the rowspaces of AQ^{-1} and A_0 respectively. Elementary linear algebra shows that the norm squared $\|P_B(\mathbf{n})\|_Q^2$ of the projection onto the row space of a matrix B is $(BQ\mathbf{n})'(BQB')^{-1}BQ\mathbf{n}$, so $\|P_{AQ^{-1}}(\mathbf{n})\|_Q^2$ and $\|P_{A_0}(\mathbf{n})\|_Q^2$ are given by

$$\begin{aligned} \|P_{AQ^{-1}}(\mathbf{n})\|_Q^2 &= \beta'(AQA')^{-1}\beta \\ \|P_{A_0}(\mathbf{n})\|_Q^2 &= (A_0Q\mathbf{n})'(A_0QA_0')^{-1}A_0Q\mathbf{n}. \end{aligned}$$

If P_h is the hypergeometric distribution on S, write

$$\begin{aligned} \frac{S_x}{S_0} &= P_h\{\|\mathbf{n}\|_Q^2 \geq mx + m\beta_0\} \\ &= P_h\{(A_0Q\mathbf{n})'(A_0QA_0')^{-1}A_0Q\mathbf{n} \geq mx + m\beta_0 - \beta'(AQA')^{-1}\beta\} \\ &= P_h\{Y'Q_0Y \geq y_x\} \end{aligned}$$

where $Y := A_0Q\mathbf{n}, Q_0 := A_0QA_0'$ and $y_x = mx + m\beta_0 - \beta'(AQA')^{-1}\beta$.

Now Y is a bounded integer vector. Let $M_i = \beta_0 \times \max\{|(A_0Q)_{ij}| : j = 1, \ldots, c\} + 1$ for $i = 1, \ldots \kappa$, so the range of Y is contained in the rectangle

$$\mathcal{R} := \prod_{i=1}^{c} \{-M_i, -M_i + 1, \ldots, M_i - 2, M_i - 1\}.$$

Let $\mu_h(\mathbf{y})$ be the probability mass function of Y in the rectangle $\mathcal{R}$ induced by the hypergeometric distribution on S, and let E_u denote the expectation of a random variable defined on $\mathcal{R}$ with respect to the uniform distribution. Then one can write

$$\begin{aligned} \frac{S_x}{S_0} &= P_h\{Y'Q_0Y \geq y_x\} \\ &= \sum_{\mathbf{y} \in R} \mu_h(\mathbf{y}) I_{\{\mathbf{y}'Q_0\mathbf{y} \geq y_x\}} \\ &= |\mathcal{R}|\ E_u\big(\mu_h(\mathbf{Y}) I_{\{\mathbf{Y}'Q_0\mathbf{Y} \geq y_x\}}\big), \end{aligned}$$

where $\mathbf{Y}$ has the uniform distribution on $\mathcal{R}$. This is useful because the law $\mu_h(\mathbf{y})$ can be computed using Fourier analysis as in §2, and then one can use a simple iid sampling scheme from the rectangle $\mathcal{R}$ to compute the p-value for goodness-of-fit.

For $\mathbf{y} \in \mathcal{R}$, let $c_{\mathbf{y}}$ be the coefficient

$$c_{\mathbf{y}} = \sum_{\substack{\mathbf{n} \in S \\ A_0 Q \mathbf{n} = \mathbf{y}}} \frac{1}{\mathbf{n}!}.$$

Then $\mu_h(\mathbf{y}) = c_{\mathbf{y}} / \sum_{\mathbf{y} \in \mathcal{R}} c_{\mathbf{y}}$. To compute $c_{\mathbf{y}}$, let $\mathbf{M} := (M_1, M_2, \ldots, M_c)$, and let $A_0 Q$ have columns $\mathbf{a}_1^0, \mathbf{a}_2^0, \ldots, \mathbf{a}_c^0$. Consider the Fourier transform

$$g_0(\mathbf{s}, \mathbf{t}) := \frac{1}{\beta_0!} \big(e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_1 / \mathbf{b}) + 2\pi i \mathbf{t} \cdot (\mathbf{a}_1^0 / 2\mathbf{M})} + \ldots + e^{2\pi i \mathbf{s} \cdot (\mathbf{a}_c / \mathbf{b}) + 2\pi i \mathbf{t} \cdot (\mathbf{a}_c^0 / 2\mathbf{M})} \big)^{\beta_0}.$$

Let $D_0 := \prod_{i=1}^{r-1} \{0, 1, \ldots, b_i - 1\} \times \mathcal{R}$. Define basis elements $h_{\mathbf{k},\mathbf{y}}$ for $(\mathbf{k}, \mathbf{y}) \in D_0$ by

$$h_{\mathbf{k},\mathbf{y}}(\mathbf{s}, \mathbf{t}) = e^{2\pi i \mathbf{s} \cdot (\mathbf{k} / \mathbf{b}) + 2\pi i \mathbf{t} \cdot (\mathbf{y} / 2\mathbf{M})}.$$

Then the proof of the formula below is like the proof of Theorem 2.1.

Theorem 4.1. For each $\mathbf{y} \in \mathcal{R}$, $c_{\mathbf{y}} = \langle g_0, h_{\beta,\mathbf{y}} \rangle / |D_0|$.

Example 4.1. Consider a 2×2 table with row sums of 2 and column sums of 2 and cells labelled left to right from the top row down, as in Example 2.1. Then $\beta_0 = 4, A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}, \beta = (2, 2)'$. Now $\mathbf{w}_1 = (1, -1, -1, 1)'$, and $\hat{N} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, hence Q can be taken to be equal to the identity matrix as well. The vector $\mathbf{b} = (5, 5, 5)$ as before, and $M_1 = 5$. Then the mass function of $A_0 Q \mathbf{n} = (n_1 - n_2 - n_3 + n_4)$ on S is 1/4 at -4, 1/2 at 0, and 1/4 at 4. With data of $(2, 0, 0, 2)$, the χ^2 value x is 4, and the conditional p-value of the χ^2 statistic is 1/2.

The transform $g_0(\mathbf{s}, t)$ is

$$g(s_1, s_2, t) = \frac{1}{4!} \big(e^{2\pi i ((s_1 + s_2)/5 + t/10)} + e^{2\pi i (s_1/5 - t/10)} + e^{2\pi i (s_2/5 - t/10)} + e^{2\pi i t/10} \big)^4.$$

which yield the values $1/4, 1, 1/4$ at values $-4, 0, 4$ of $\mathbf{w}_1 \cdot \mathbf{n}$.

5. Conclusions.

We have presented two very general methods for computing the distributions of linear statistics on multidimensional lattice points. The Fourier analysis method has its foundations in signal processing, and the algebraic method comes from operator theory for solving hypergeometric systems of differential equations. We have applied the methods to examples of tables under both graphical and nongraphical models, and we have shown by comparison with other solution methods that the two new methods can be very accurate and efficient. Further research problems are 1) to get "extension" theorems for the A-hypergeometric systems in order to identify the desired solution and justify theoretically

the proposed elimination method, 2) to estimate sample sizes for the Monte Carlo variation on the Fourier analysis computations. While the use of differential algebra is new and promising, its use in statistics would benefit from fast, convenient software that builds on the implementation in Macaulay 2.

Acknowledgments. We have used the software R from lib.stat.cmu.edu, Macaulay 2, and Maple 7 for computations. The second author was supported by a Sarah M. Hallam Fellowship at UC Berkeley, and by the Clay Mathematics Institute. The first and third authors were supported by NSF grant DMS-0200888, and a Board of Regents grant from the State of Louisiana. Part of the work was done while the first author was visiting SAMSI supported with DMS-0112069. We would like to thank Bernd Sturmfels and Michael Singer for essential help and discussions.

References

Agresti A. 1990. *Categorical Data Analysis.* Wiley, New York.

Baglivo J., Olivier D., and Pagano M. 1992. Methods for exact goodness-of-fit tests. *JASA*, **87**, 464-469.

Coutinho S. C. 1995. *A Primer of Algebraic D-Modules.* Cambridge University Press, Cambridge.

Diaconis P., and Sturmfels B. 1998. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics* **26**, 363-397.

Diaconis P., Graham R., and Holmes S. P. 1999. Statistical problems involving permutations with restricted positions. In *State of the Art in Probability and Statistics: Festschrift for Willem R. van Zwet*, eds. M. de Gunst, C. Klaassen, A. Van der Vart, Institute of Mathematical Statistics, Beachwood Ohio, 195-222.

Dinwoodie I. H. 1998. Generating functions for exact p-values of odds ratios in logistic regression. *Journal of the Italian Statistical Society* **3**, 221-232.

Dinwoodie I. H., and MacGibbon B. 2001. Exact analysis of a paired sibling study. GERAD-École des HEC Technical Report G-2001-35.

Dobra A., and Fienberg S. E. 2000. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *PNAS* **97**, 11885-11892.

Dobra A. 2002. Markov bases for decomposable graphical models. *Bernoulli*, to appear.

Eisenbud D., Grayson D. R., Stillman M., and Sturmfels B. 2002. *Computations in*

Algebraic Geometry with Macaulay 2. Springer, New York.

Fraser F. C., and Hunter A. D. W. 1975. Etiologic relations among categories of congenital heart malformations. *The American Journal of Cardiology* **36**, 793-796.

Geiger D. , Meek C., and Sturmfels B. 2002. On the toric algebra of graphical models. Manuscript available at research.microsoft.com/research/pubs/.

Good I. J., Gover T. N. , and Mitchell G. J. 1970. Exact distributions for χ^2 and for the likelihood ratio statistic for the equiprobable multinomial model. *JASA* **65**, 267-283.

Kreuzer M., and Robbiano L. 2000. *Computational Commutative Algebra I.* Springer, New York.

Liu J. S. 2001 . *Monte Carlo Strategies in Scientific Computing.* Springer, New York.

MacGibbon B. 1983. A log-linear model of a paired sibling study. In *Proceedings of Statistics '81 Canada Conference,* eds. Y. Chaubey, T. D. Dwivedi, 193-197.

Matusevich, L. F. 2002. Combinatorial aspects of hypergeometric functions. Ph. D. Thesis, University of California at Berkeley.

McDonald J. W., and Smith P. W. F. 1995. Exact conditional tests of quasi-independence for triangular contingency tables: estimating significance levels. *Applied Statistics* **44**, 143-151.

Saito M., Sturmfels B., and Takayama N. 2001. *Gröbner Deformations of Hypergeometric Differential Equations.* Springer, New York.

Stigler S. 1992. Studies in the history of probability and statistics XLIII. Karl Pearson and quasi-independence. *Biometrika* **79**, 563-575.

ISDS
Old Chemistry Building
Duke University
Durham, NC, 27708
ihd@stat.duke.edu