Algebraic Algorithms for Sampling from Conditional Distributions

Persi Diaconis
Department of Mathematics
Harvard University
Cambridge, Massachusetts 02138

Bernd Sturmfels
Department of Mathematics
Cornell University
Ithaca, New York 14853

Abstract

We construct Markov chain algorithms for sampling from discrete exponential families conditional on a sufficient statistic. Examples include generating tables with fixed row and column sums and higher dimensional analogs. The algorithms involve finding bases for associated polynomial ideals and so an excursion into computational algebraic geometry.

1. Introduction. As a simple example of the problem under study consider generating a random contingency table with fixed row and column sums. Thus, fix positive integers I and J and a set of row sums $r_1, r_2, \cdots, r_I$ and column sums $c_1, c_2, \cdots, c_J$. Let $\mathcal{X} = \mathcal{X}_{\underset{\sim}{r},\underset{\sim}{c}}$ be the set of $I \times J$ arrays $X = (x_{ij})$ of non-negative integers with the given row sums and column sums. Let U be the uniform distribution on $\mathcal{X}$ and let H be the hypergeometric distribution on $\mathcal{X}$ (so $H(x) = \prod_{i,j} \binom{c_j}{x_{ij}} / \binom{N}{r_1, r_2, \cdots, r_I}$ with $N = \sum c_i = \sum r_j$). Classical tasks such as tests for independence involve approximating the distributions of a statistic S such as the chi-squared under H. Martin Löf (1974), Diaconis-Efron (1986), Good (1976) and others have asked for the distribution of S under U. In both cases, asymptotic theory is suspect and a variety of other approaches have been considered. The literature on these approaches is reviewed in Section 2.

We develop a Monte-Carlo approach along the following lines. Let X be a table which satisfies the constraints. Modify X by picking a pair of rows and a pair of columns at random. These intersect in 4 entries and X is modified as

$$\begin{matrix} + & - \\ - & + \end{matrix} \quad \text{or} \quad \begin{matrix} - & + \\ + & - \end{matrix}$$

with probability 1/2 each. This modification adds and subtracts 1 from each of the 4 entries as indicated. This doesn't change the row or column sums. If the modification forces negative entries, discard it and continue by choosing a new pair of rows and columns. This describes a Markov chain on $\mathcal{X}_{\underset{\sim}{r},\underset{\sim}{c}}$. By construction the chain is symmetric. It can be shown that the chain is connected. It follows that its stationary distribution equals U. This gives us the ability to sample (approximately) from U. A slight modification, weighting the moves above, allow us to sample from H. For 2-dimensional arrays, sampling from H can be done more easily by different methods; we know of no other way to sample from U.

As an example, Table 1 gives a 4×4 contingency table (data of Snee (1974)). The chi-square test of independence for this table is $\chi^2 = 138.29$ on 9-degrees of freedom, strongly rejecting the hypothesis of independence. Diaconis and Efron (1985) labored long and hard to determine the proportion of tables with the same row and column sums as table 1 having $\chi^2 \leq 138.29$. Their best estimate was "about 10%". Figure 1 shows a histogram from a Monte Carlo run. In the run, 16.3% of all tables had $\chi^2 \leq 138.29$. The Monte

Carlo was considerably more accurate than the best that asymptotic theory can provide. It was also trivial to run. Figure 1 is based on a variant of the algorithm described above. Section 2-B gives more details.

Eye color versus hair color for n = 592 subjects, Snee (1974)

Table 1

Eye Color	Hair Color				
	Black	Brunette	Red	Blonde	Total
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

Figure 1

Histogram of 10^6 sample values of a Monte Carlo sample of chi-square values from tables with the same row and column sums as Table 1.

Figure 1 is not available as postscriptfile.To get a paper-copy send an e-mail to bibliothek@sc.zib-berlin.de.

This papers extend these algorithms to more general settings, including three and higher dimensional arrays, for which our approach seems to be the only available route, even for H. To define this general class of problems, let $\mathcal{X}$ be a finite set and $T : \mathcal{X} \to \mathbb{Z}^d$ any function. Data with values in $\mathcal{X}$ can be summarized as a function $f : \mathcal{X} \to \mathbb{N}$, where $f(x)$ is the number of observations taking value x. This fixes the vector $t = \Sigma f(x)T(x)$ in

$\mathbb{Z}^d$. Let

$$\mathcal{X}_t \quad = \quad \big\{g : \mathcal{X} \to \mathbb{N} \quad \text{such that} \quad \Sigma g(x)T(x) = t\big\}. \tag{1.1}$$

Our algorithms give a way of sampling from the uniform and an appropriate hypergeometric distribution on $\mathcal{X}_t$. They are based on finding functions $f_{\mathbf{1}}, f_{\mathbf{2}}, \cdots, f_L : \mathcal{X} \to \mathbb{Z}$ such that

(a) $\displaystyle\sum_x f_i(x)T(x) = 0 \quad \text{for } 1 \le i \le L.$

(b) For any t, and any $g, g' \in \mathcal{X}_t$ there exists $(\epsilon_{\mathbf{1}}, f_{i_j}) \cdots (\epsilon_A, f_{i_{\boldsymbol{A}}})$ with $\epsilon_i = \pm 1$, (1.2)

$$g' = g + \sum_{j=\mathbf{1}}^{A} \epsilon_j f_{i_j} \qquad \text{and} \qquad g + \sum_{j=\mathbf{1}}^{\alpha} \epsilon_j f_{i_j} \ge 0, \quad \text{for } 1 \le \alpha \le A.$$

The set $\{f_{\mathbf{1}}, \ldots, f_L\}$ allows a Markov chain to be constructed on $\mathcal{X}_t$, for each $t \in \mathbb{Z}^d$, by choosing I at random and adding $\pm f_I$. If this gives negative entries, the chain stays fixed. Condition (a) says such moves keep the chain in $\mathcal{X}_t$. Condition (b) implies that the chain is connected. By construction the chain is symmetric and so has a uniform stationary distribution.

Section 2 lays out the stochastic underpinnings. It describes how the problems considered above arise when sampling from exponential families through T. Finding bases (as in (1.2)) can be computationally prohibitive and a computationally feasible approach called the fiber algorithm is presented.

The main new contribution is a method for finding basic moves using tools from computational algebraic geometry. Section 3 shows how finding $\{f_{\mathbf{1}}, \ldots, f_L\}$ is equivalent to finding generators for a certain polynomial ideal. We describe an effective set of techniques for computing such generators. The key word here is *Gröbner bases*.

Sections 4, 5, 6 contain detailed treatments of special cases: contingency tables, logistic regression and ranked data sets. In each case, a practical example appears along with theoretical development. These may be read now as motivation.

This paper does not address questions of running times for these Markov chains. We are in the process of developing rigorous bounds and comparing these with the results of simulation and exact computation.

2. Basic Stochastics.

A. Sample spaces. Let $\mathcal{X}$ be a finite set. Let $T : \mathcal{X} \to \mathbb{Z}^d$ be a function. A statistical problem begins with N observed values $x_{\mathbf{1}}, x_{\mathbf{2}}, \cdots, x_N$ from $\mathcal{X}$. The summary value $t = \sum_{i=\mathbf{1}}^{N} T(x_i)$ leads to consideration of the *big fiber*

$$\mathcal{Y}_t \quad = \quad \big\{(x_{\mathbf{1}}, x_{\mathbf{2}}, \cdots, x_N) \in \mathcal{X}^N \, : \, T(x_{\mathbf{1}}) + \cdots + T(x_N) = t\big\}. \tag{2.1}$$

To assure that $\mathcal{Y}_t$ is finite, we will assume the following condition throughout:

$$\text{there exists } \omega \in \mathbb{Z}^d \text{ such that } T(x) \cdot \omega > 0 \text{ for all } x \in \mathcal{X}. \tag{2.2}$$

For example, all $T(x)$ might have the same sum of coordinates or first coordinate equal to 1. One motivation for considering $\mathcal{Y}_t$ comes from exponential family theory. If the x_i are a realization of N independent and identically distributed choices from

$$P_\theta(x) = c(\theta) e^{\theta \cdot T(x)}, \quad \theta \in \mathbb{R}^d, \quad c(\theta) \quad \text{a normalizing constant,}$$

then t is a sufficient statistic for θ: the conditional distribution of the sample given t is uniform over $\mathcal{Y}_t$. This conditional uniform distribution is a basic ingredient for classical tests of the goodness of fit of this exponential family.

If N is large and $|\mathcal{X}|$ is small, it is natural to consider $f : \mathcal{X} \to \mathbb{N}$ given by $f(x) = \#\{i : x_i = x\}$. This is a sufficient statistic for any independent identically distributed data. Let the *little fiber* be

$$\mathcal{X}_t \quad = \quad \big\{g : \mathcal{X} \to \mathbb{N} \, : \, \sum_x g(x) T(x) = t\big\}. \tag{2.3}$$

There is a natural map from $\mathcal{Y}_t$ to $\mathcal{X}_t$. The image of the uniform distribution on $\mathcal{Y}_t$ will be called the *hypergeometric distribution* H_t on $\mathcal{X}_t$. Thus,

$$H_t(g) \quad = \quad \frac{N!}{|\mathcal{Y}_t|} \prod_x (g(x)!)^{-\mathbf{1}}. \tag{2.4}$$

In many problems there is no effective way to enumerate $\mathcal{Y}_t$ or sample directly from H_t.

The *uniform* distribution U_t on $\mathcal{X}_t$ is a second probability which is useful for working with H_t and also of direct interest. One motivation comes from Bayesian considerations: Let $\underset{\sim}{\theta} = \{\theta_x\}_{x \in \boldsymbol{\mathcal{X}}}$ be chosen from the uniform distribution on the $|\mathcal{X}|$ simplex. Let $X_{\mathbf{1}}, X_{\mathbf{2}}, \cdots, X_N$ be chosen from a multinomial distribution on $\mathcal{X}$ with parameter $\underset{\sim}{\theta}$. By Bayes' classical argument, this two stage process induces a Bose-Einstein distribution on

$$\{g : \mathcal{X} \to \mathbb{N} : \sum_x g(x) = N\} \quad \text{with} \quad g(x) = \#\{i : X_i = x\}.$$

Thus for any T and t the conditional distribution given t is uniform on $\mathcal{X}_t$. Good (1979) or Diaconis and Efron (1987) give modern versions of Bayes' original argument.

Diaconis and Efron (1986) motivated the uniform distribution as an easily interpretable antagonistic alternative to the hypergeometric. Testing with respect to U_t counts the number of data sets with more extreme test statistics. For example, in Table 1, a chi-square statistic of 138.29 strongly rejects the hypothesis of independence. One may ask if the underlying generating mechanism was close to independence, perhaps "blown up" by a large sample size. Our computations reject this; the table appears the same as a randomly chosen table, most likely far from independent.

B. Markov chains on $\mathcal{X}_t$. With $\mathcal{X}_t$ as in (2.3), suppose that $\{f_{\mathbf{1}}, f_{\mathbf{2}} \cdots, f_L\}$ is a generating set as in (1.2).

LEMMA 2.1. Generate a Markov chain on $\mathcal{X}_t$ by choosing I uniformly in $\{1, 2, \ldots, L\}$ and $\epsilon = \pm 1$ with probability $1/2$ independent of I. If the chain is currently at $g \in \mathcal{X}_t$ the chain moves to $g + \epsilon f_I$, provided this is non-negative, and stays at g otherwise. This gives a connected, symmetric, aperiodic Markov chain on $\mathcal{X}_t$ with the uniform distribution as its stationary distribution.

PROOF: Write $P(g, \tilde{g})$ for the chance of going from g to $\tilde{g}$ in one step. If this is not zero and $g \neq \tilde{g}$ there is an ϵf_i such that $\tilde{g} = g + \epsilon f_i$ and $P(g, \tilde{g}) = 1/(2L)$. Then $g = \tilde{g} - \epsilon f_i$ gives the unique step taking $\tilde{g}$ to g. Thus the chain is symmetric. Condition (1.2b) says it is connected and it clearly has some holding probability (just take any $g \in \mathcal{X}_t$ and repeatedly subtract a fixed f_i until the boundary of $\mathcal{X}_t$ is reached). This implies that the chain has the uniform distribution as its unique stationary distribution. □

To generate from the hypergeometric or other distributions on $\mathcal{X}_t$ we introduce the following variant of the Metropolis algorithm.

LEMMA 2.2. Let σ be a positive function on $\mathcal{X}_t$ of (2.3). Generate a Markov chain on $\mathcal{X}_t$ by choosing I uniformly in $\{1, 2, \ldots, L\}$ and $\epsilon = \pm 1$ with probability $1/2$ independent of I. If the chain is currently at g it moves to $\tilde{g} = g + \epsilon f_I$ (provided $\tilde{g} \in \mathcal{X}_t$) with probability $\min(\sigma(\tilde{g})/\sigma(g), 1)$. In all other cases the chain stays at g. This is a connected, reversible Markov chain on $\mathcal{X}_t$ with stationary distribution proportional to $\sigma(g)$.

PROOF: It is easy to check that $\sigma(g)P(g, \tilde{g}) = \sigma(\tilde{g})P(\tilde{g}, g)$ for all $g, \tilde{g} \in \mathcal{X}_t$. Since the moves connect $\mathcal{X}_t$ and there is some holding probability the chain has a unique stationary distribution proportional to $\sigma(g)$. □

REMARKS: 1. A useful class of measures on $\mathcal{X}_t$ is specified by choosing a function $\omega_x : \mathbb{N} \to \mathbb{R}^+$ for each $x \in \mathcal{X}$. For $g \in \mathcal{X}_t$ define $\sigma(g) = \prod_x \omega_x(g(x))$. As examples, if $\omega_x(a) = \theta_x^a/a!$ with $0 < \theta_x \leq 1$, then σ becomes the multiple hypergeometric distribution which arises when carrying out power calculations or using the random walks to generate confidence regions. Taking $\theta_x = 1$ gives the hypergeometric distribution. For this class of measures the ratios $\sigma(\tilde{g})/\sigma(g)$ reduce to very few terms if g and $\tilde{g}$ only differ in a few coordinates. We have found this method and effective in the applications of Sections 4,5,6.
2. The chain of Lemma 2.2 is different from the classical Metropolis algorithm described by Hammersly and Handscomb (1964, Chapter 9). It weights the steps slightly differently. It might converge in less steps, but each step requires a considerably more detailed computation. In our experience this extra computing slows things down very much.

The algorithms in this paper are closely related to the popular Gibbs sampler. This generates a given distribution on random vectors by changing coordinates one at a time according to the conditional distribution given the complement of the coordinate. We explain the connection and offer some speed ups of Lemma 2.2. Let $Z = \prod_{i=1}^d Z_i$, with each Z_i a finite set. Let $\pi(z)$ be a probability on Z. Let $\mathcal{C}$ be a class of subsets of $\{1, 2, \cdots, d\}$. For $c \in \mathcal{C}$ and $z \in Z$, let z_c (resp. z^c) be the coordinates inside (resp. outside) of c. Let $\pi(z_c|z^c) = \pi(z)/\pi(z^c)$. Let $P_c^{z^c}(\bullet, \bullet)$ be a Markov chain on $\prod_{i \in c} Z_i$ which is reversible with respect to $\pi(z_c|z^c)$. Define a Markov chain on Z by specifying

positive weights w_c for each $c \in \mathcal{C}$ and then defining

$$P(z,y) \quad = \quad \sum_{c \in \mathcal{C}} w_c P_c^{z^c}(z_c, y_c). \tag{2.5}$$

This chain proceeds by picking a subset c and modifying those coordinates by the transition mechanism $P_c^{z^c}(\bullet, \bullet)$.

LEMMA 2.3. The chain defined by (2.5) is reversible with respect to $\pi(z)$.

PROOF: For $z \neq y, P(z,y) = P(y,z) = 0$ unless $z^c = y^c$ for some $c \in C$. Let $S = S(z,y)$ be the set of $c \in \mathcal{C}$ such that this occurs. Then

$$\begin{aligned} \pi(z)P(z,y) &= \sum_{c \in S} w_c \pi(z) P_c^{z^c}(z_c, y_c) = \sum_{c \in S} w_c \pi(z_c|z^c)\pi(z^c) P_c^{z^c}(z_c, y_c) \\ &= \sum_{c \in S} w_c \pi(y_c|y^c)\pi(y^c) P_c^{y^c}(y_c, z_c) = \pi(y)P(y,z). \end{aligned}$$ □

As an application, consider generating from the hypergeometric distribution (2.4) given a set of moves as in (1.2). Here Z may be taken as $Z = [0,N]^{|\mathcal{X}|}$ with $z \in Z$ identified with the function $\{g(x)\}_{x \in X}$. The stationary distribution π may be taken as the hypergeometric (2.4). This is supported on $\mathcal{X}_t \subseteq Z$. Each move f_i in (1.2) defines a support set $c_i = \{x : f_i(x) \neq 0\}$. Take uniform weights on i. If the chain is currently at g, it moves to $\tilde{g} = g + jf_i$. Here j varies in an interval of values $-N \leq j \leq N$ such that $g + jf_i \in \mathcal{X}_t$. The weights for choosing j are taken proportional to the stationary probability of $g + jf_i$. Using the notation above,

$$P_i^{g^i}(g_i, (g + jf_i)_i) \quad \propto \quad \prod_{x \in c_i} [(g(x) + jf_i(x))!]^{-1}. \tag{2.6}$$

To use this, given g and c, one would have to run through an interval of j values, keeping track of both which j values are possible and the relative weight (2.6). Of course, if this last step can be done in closed form, things will go faster.

For two-way tables this algorithm is simple: pick a pair of rows and columns at random. This fixes a 2×2 table. Replace it by a second 2×2 table with the same margins, chosen from the hypergeometric distribution. This procedure, adapted for uniform generation, produced Figure 1. Here the 2×2 array was replaced by a uniformly chosen 2×2 array

with the same row and column sums. Figure 1 is based on 10^6 trials with 500 steps between each trial. It agrees in every regard with results in Gangolli (1991) who ran an extensive Monte Carlo trial on the same 4×4 table using the original "move one" algorithm.

C. Fiber Walks. This section uses previous ideas on small parts of a larger problem. Let $f_1, f_2, \cdots, f_L$ be a generating set as in (1.2). Write $f_i^+ = \max\{f_i, 0\}$, $f_i^- = \max\{-f_i, 0\}$ so $f_i = f_i^+ - f_i^-$. Let $\deg f_i = \max\{\sum_x f_i^+(x), \sum_x f_i^-(x)\}$. Let $D = \max\{\deg f_i : i = 1, \ldots, L\}$. For contingency tables the basic $\pm, \mp$ moves have degree 2. In generating a *fiber walk* we pick a multi-set h^* of degree D from some sampling distribution, often just by sampling at random from $\mathcal{X}$ with replacement. We calculate $t^* = \sum_x h^*(x)T(x)$, thus fixing the *subfiber* $\mathcal{X}_{t^*} = \{h : \sum h(x)T(x) = t^*\}$. The next step is to choose an element of $\mathcal{X}_{t^*}$ at random (from the uniform distribution on $\mathcal{X}_{t^*}$). For D small this may be done by enumerating $\mathcal{X}_{t^*}$. If enumeration is not feasible it may be done by running a Markov chain with respect to h^*. This random choice is then swapped for h^* within the original data set. This gives a connected Markov chain. In what follows we explain how to use this construction to sample from the hypergeometric and uniform distributions.

Lemma 2.4 (Hypergeometric Generation). Suppose we are given a generating set as in (1.2) having degree $\leq D$. Generate a Markov chain on $\mathcal{X}_t$ of (2.3) as follows. If the chain is currently at $g \in \mathcal{X}_t$, then choose h^* of degree D with probability

$$\prod_x \binom{g(x)}{h^*(x)} \Big/ \binom{N}{D}. \tag{2.7}$$

Choose an element h from the uniform distribution on the subfiber $\mathcal{X}_{t^*}$. The chain moves to $g(x) - h^*(x) + h(x)$. This generates a connected symmetric Markov chain with the hypergeometric (2.4) as its stationary distribution.

Proof: The argument is best pictured as a process on the big fiber $\mathcal{Y}_t$ of (2.1). Let $x_1, x_2, \cdots, x_N$ be N elements of $\mathcal{X}$ consistent with g. The chain on $\mathcal{Y}_t$ proceeds by choosing D elements from $x_1, \cdots, x_N$ without replacement. Let $h^*(x)$ be the number of chosen elements equal to x. Then h^* has the distribution (2.7). Choose h from the uniform distribution on the little subfiber $\mathcal{X}_{t^*}$. Replace the subset corresponding to h^* by elements corresponding to h in random order. This generates a symmetric chain on $\mathcal{Y}_t$: a move is

specified by a subset and a unique point in its subfiber. This gives a 1-1 correspondence between moves forward and backward. This chain is connected by our assumptions on the moves $f_1, \ldots, f_L$. Since the order in the big fiber $\mathcal{Y}_t$ does not enter the considerations (one could always add gratuitous random permutations of size N at each stage), this chain has the uniform distribution as its stationary distribution on $\mathcal{Y}_t$, and so induces a hypergeometric stationary distribution on $\mathcal{X}_t$. □

REMARKS: Examples of these fiber walks are given in Section 4. Of course, it is possible to iterate, using the same idea for the problem of choosing a point in the subfiber $\mathcal{X}_{t^*}$. A crucial ingredient for fiber walks is a bound on the degree D. These can sometimes be quite sharp (see e.g. Theorem 6.1). A discussion of general bounds for D is given in Section 3D.

D. Literature Review for Conditional and Exact Analysis. The work presented here has numerous links to inferential and algorithmic problems. In this section we give pointers to the most closely related literature.

As with so many topics of inferential interest, conditional testing was first studied by R.A. Fisher. He systematically used the conditional distribution of the data given a sufficient statistic as a base for tests of a model in statistical methods for research workers (Fisher (1925)). He suggested and defended the use of conditional tests in regression, contingency tables, and elsewhere. Savage (1976) contains an overview and Yates (1984) gives a careful history of the controversy over conditional testing for 2×2 tables. Cox (1958), Kiefer (1977), Efron and Hinkly (1978) and Brown (1990) have been influential papers which have extensive literature reviews. Lehmann (1986, Chapter 10) gives a splendid overview of the issues. Barndorff-Nielsen (1978) gives fresh perspectives and subtleties.

One side of the argument is a feeling that the margins of a table (or other ancillaries) contain "no information" about independence. Plackett (1977) presents this view forcefully. Our own view is that there are clear differences between the conclusions that can be stated following a conditional versus an unconditional test. Once this is understood, the appeal of conditional tests is largely one of convenience. In the contingency table setting,

conditioning gets us out of having to fuss with the margins. One has a clean, exact statement. This also holds for regression and Mantel-Henzel type methods of combining many test statistics.

On the computational side, there has become a growing awareness that the usual asymptotic approximations of mathematical statistics can be poor for moderate sample sizes. Clear examples in a contingency table setting one given by Yarnold (1970), Odoroff (1970), Larntz (1978), and many later writers. This has led recent investigators to pursue an intensive program of exact computation or better approximation. The Monte Carlo approach described here seems to be "in the air" currently. Versions for 2-way tables are explicitly described by Aldous (1987), Gangolli (1991), and Glonek (1987). It is easy to generate an $I \times J$ table with fixed margins from the hypergeometric distribution: generate a random permutation of n items. Look in the first $r_{\mathbf{1}}$ places; the number of entries between $c_{\mathbf{1}} + \cdots + c_{j-\mathbf{1}} + 1$ and $c_{\mathbf{1}} + \cdots + c_j$ is $n_{\mathbf{1}j}$, $1 \le j \le J$. The number of such entries in the next $r_{\mathbf{2}}$ places is $n_{\mathbf{2}j}$, and so on.

Marcello Pagano, working with a variety of co-authors, has suggested methods for exact computations using the fast Fourier transform. Papers by Baglivio, Olivier, and Pagano (1988, 1992, 1993) contain refined versions of these ideas and pointers to earlier literature. Exact computational procedures are given for contingency tables, logistic regression, and a variety of standard discrete data problems.

Mehta and Patel (1983) proposed a novel network approach which achieves exact enumeration by using dynamic programming ideas. This has been refined and extended into the program STATEXACT (Mehta and Patel (1991)) which carries out tests for contingency tables and other problems.

A third approach uses the representation of the hypergeometric distribution as the conditional distribution for an exponential family (2.3) given t. Choosing an appropriate value of θ (e.g., $\hat{\theta}$ the maximum likelihood estimator) Edgeworth approximations to the probability $P_{\hat{\theta}}(t)$ and $P_{\hat{\theta}}(x, t)$ are computed. Their ratio gives an approximation to H_t. These seem quite accurate for a variety of applications with moderate sample sizes. Levin (1983, 1992) sets out the general theme which is developed in Levin and Kong (1993) and in Kong (1993). McCullogh (1985, 1986), Diaconis and Freedman (1987), Jensen (1991)

and Skovgaard (1987) give further relevant results for such conditional approximations. Kolassa and Tanner (1993) is a recent contribution in this direction. Agresti (1992) gives a survey of recent work on exact conditional inference for contingency tables with discussion by the major contributors to the field.

The material in this paper is closely related to a classical problem of combinatorics: *enumerating* the number of arrays with given row and column sum. See Good (1977) for a review. Results of Jerrum, Valiant and Vazerani (1986), see also Sinclair (1993), show that there is a provably close connection between enumeration and random generation. While no one has proved that table enumeration is intractable (#-p complete), it seems likely to be so. Indeed, if structural zeros are prescribed, then intractability can be shown as follows: Take $\mathcal{X}$ to be the edges of a bipartite graph $\mathcal{G}$ on $2n$ vertices. Let $T : \mathcal{X} \to \mathbb{N}^{\mathbf{2}n}$ be the indicator function of the vertices in x. Let $t = \underset{\sim}{1}$ be the vector of all ones in $\mathbb{Z}^{\mathbf{2}n}$. The little fiber $\mathcal{X}_t$ consists precisely of the perfect matchings in $\mathcal{G}$. It is known that counting $\mathcal{X}_t$ for general $\mathcal{G}$ is #-p complete. Jerrum and Sinclair (1986) have given random walk algorithms for generating elements of $\mathcal{X}_t$ and proved that they are rapidly mixing. See also Diaconis and Stroock (1991) and Sinclair (1993).

Of course, examples may still be computed: David des Jardins has computed that there are 1, 225, 914, 276, 768, 514 tables with the same row and column sums as Table 1.

3. Algebraic basics.

A. Toric Ideals and Their Generators. In this section we set up the relation between generating sets as in (1.2) and a certain class of polynomial ideals. Throughout, $\mathcal{X}$ is a finite set and $T : \mathcal{X} \to \mathbb{Z}^d$ is given. For each $x \in \mathcal{X}$ we introduce an indeterminate also denoted x. Consider the ring $K[\mathcal{X}]$ of polynomials in these indeterminates, where K is any field. For computational purposes it is convenient to use $K = GF(2)$, the field of two elements. A function $g : \mathcal{X} \to \mathbb{N}$ will be represented by a monomial $\prod_{x \in \boldsymbol{\mathcal{X}}} x^{g(x)}$. We denote this monomial as $\mathcal{X}^g$. Likewise, we identify lattice points $(i_{\mathbf{1}}, i_{\mathbf{2}}, \cdots, i_d)$ in $\mathbb{Z}^d$ with monomials $t_{\mathbf{1}}^{i_{\mathbf{1}}} t_{\mathbf{2}}^{i_{\mathbf{2}}} \cdots t_d^{i_{\boldsymbol{d}}}$ in the ring of Laurent polynomials $K[t_{\mathbf{1}}, \cdots, t_d, t_{\mathbf{1}}^{-\mathbf{1}}, \cdots, t_d^{-\mathbf{1}}]$. The function $T : \mathcal{X} \to \mathbb{Z}^d$ is represented by the K-algebra homomorphism

$$\begin{aligned} \varphi_T \,:\, K[\mathcal{X}] \quad &\to \quad K[\, t_{\mathbf{1}}, \cdots, t_d, t_{\mathbf{1}}^{-\mathbf{1}} \cdots t_d^{-\mathbf{1}}] \\ x \quad &\mapsto \quad t_{\mathbf{1}}^{T(x)_{\mathbf{1}}} t_{\mathbf{2}}^{T(x)_{\mathbf{2}}} \cdots t_d^{T(x)_{\boldsymbol{d}}}. \end{aligned}$$

Here $T(x)_i$ denotes the i-th coordinate of $T(x) \in \mathbb{Z}^d$ and the map φ_T is K-linear and multiplicative on products. Writing $T(g) = \sum_{x \in \boldsymbol{\mathcal{X}}} g(x)T(x)$, we thus have $\varphi_T(\mathcal{X}^g) = t_{\mathbf{1}}^{T(g)_{\mathbf{1}}} t_{\mathbf{2}}^{T(g)_{\mathbf{2}}} \cdots t_d^{T(g)_{\boldsymbol{d}}}$. Our basic object of study is $\mathcal{I}_T$, the kernel of φ_T. The prime ideal $\mathcal{I}_T$ is called the *toric ideal* associated with T. This terminology stems from the fact that the zero set of $\mathcal{I}_T$ is an *affine toric variety* (see e.g. (Fulton 1993)).

The fundamental relationship to the problems of Section 1 is a correspondence between generating sets as in (1.2) with generating set for the toric ideal $\mathcal{I}_T$. This will be established in Theorem 3.2. To state this correspondence and resulting algorithms, we need the following notation. Any function $f : \mathcal{X} \to \mathbb{Z}$ can be written as the difference of two functions $f^+, f^- : \mathcal{X} \to \mathbb{N}$ having disjoint support; $f^+(x) := \max(f(x), 0)$ and $f^-(x) := \max(-f(x), 0)$. A function f satisfies $\sum_x f(x)T(x) = 0$ if and only if the monomial difference $\mathcal{X}^{f^+} - \mathcal{X}^{f^-}$ is in $\mathcal{I}_T$.

LEMMA 3.1. The ideal $\mathcal{I}_T$ is generated by the monomial differences $\mathcal{X}^{f^+} - \mathcal{X}^{f^-}$, where f runs over all functions $f : \mathcal{X} \to \mathbb{Z}$ with $\sum_x f(x)T(x) = 0$.

PROOF: Let $\mathcal{I}'$ be the ideal generated by all monomial differences $\mathcal{X}^{f^+} - \mathcal{X}^{f^-}$ with $\Sigma f(x)T(x) = 0$. Clearly $\mathcal{I}' \subseteq \mathcal{I}_T$. We fix a total order on the set of all monomials as

follows. First linearly order the variables $\mathcal{X}$. One monomial is larger than a second if either the degree of the first is larger, or the degrees are equal and on the first variable where they disagree, the first has a higher power.

Suppose the inclusion $\mathcal{I}' \subset \mathcal{I}_T$ is strict. Let $p \in \mathcal{I}_T \setminus \mathcal{I}'$ have its largest monomial, say $\mathcal{X}^\alpha$, a minimum. Since $\varphi_T(p) = 0$ there must be a second monomial $\mathcal{X}^\beta$ in p such that $\varphi_T(\mathcal{X}^\beta) = \varphi_T(\mathcal{X}^\alpha)$. Factor out common variables, writing $\mathcal{X}^\alpha - \mathcal{X}^\beta = \mathcal{X}^\gamma(\mathcal{X}^{\alpha'} - \mathcal{X}^{\beta'})$ with α' and β' having disjoint support. Clearly $\varphi_T(\mathcal{X}^{\alpha'}) - \varphi_T(\mathcal{X}^{\beta'}) = 0$. Setting $h(x) := \alpha'_x - \beta'_x$, we have $\Sigma h(x)T(x) = 0$ and hence $\mathcal{X}^\gamma(\mathcal{X}^{h^+} - \mathcal{X}^{h^-}) \in \mathcal{I}'$. Subtracting this expression from p, we get a polynomial in $\mathcal{I}_T \setminus \mathcal{I}$ whose leading monomial is smaller than p. □

We remark that, by Hilbert's Basis Theorem, there exists a finite subset of functions f which generate the toric ideal $\mathcal{I}_T$ in the sense of Lemma 3.1. The question of how to find such a finite set will be addressed in Section 3.B. We next first show that any such set solves the problem stated in Section 1. Given $\mathcal{X}$ and $T : \mathcal{X} \to \mathbb{Z}^d$, our problem was to find functions $f_{\mathbf{1}}, f_{\mathbf{2}}, \cdots, f_L : \mathcal{X} \to \mathbb{Z}$ satisfying (1.2), repeated here for ease of reference:

(3.1)
(a) $\sum_x f_i(x)T(x) = 0, \qquad 1 \le i \le L$

(b) For $g, g' : \mathcal{X} \to \mathbb{N}$ with $\sum(g(x) - g'(x))T(x) = 0$, there exist $\epsilon_j, f_{i_j},\ 1 \le j \le A$ with $\epsilon_j = \pm 1,\ g + \sum_{j=\mathbf{1}}^{A} \epsilon_j f_{ij} = g'$ and $g + \Sigma_{j=\mathbf{1}}^{\alpha} \epsilon_j f_{i_j}$ non-negative, for $1 \le \alpha \le A$.

As was demonstrated in Section 2, such functions give rise to random walks for sampling from $\mathcal{X}_t$. For readers interested in optimization we mention the related *integer programming* problem of minimizing a linear functional over $\mathcal{X}_t$. Algebraic techniques for this problem along the same lines were developed by Conti and Traverso (1991) and Thomas (1993).

THEOREM 3.2. A collection of functions $f_{\mathbf{1}}, f_{\mathbf{2}}, \cdots, f_L : \mathcal{X} \to \mathbb{Z}$ satisfies (3.1a,b) if and only if the set $\left\{\mathcal{X}^{f_i^+} - \mathcal{X}^{f_i^-} \ : \ 1 \le i \le L\right\}$ generates the toric ideal $\mathcal{I}_T$.

PROOF: Let $\mathcal{F} = \{\mathcal{X}^{f_i^+} - \mathcal{X}^{f_i^-} : i = 1, \ldots, L\}$. Property (a) is equivalent to $\mathcal{F} \subset \mathcal{I}_T$. Thus we must show (b) holds if and only if $\mathcal{F}$ generates $\mathcal{I}_T$. Assume (b) holds. By Lemma 3.1 it is enough to show that, for any $f : \mathcal{X} \to \mathbb{Z}$ with $\sum_x f(x)T(x) = 0$, the monomial difference $\mathcal{X}^{f^+} - \mathcal{X}^{f^-}$ is in the ideal generated by $\mathcal{F}$. Apply (b) to $g = f^+$ and $g' = f^-$. If $A = 1$ and say $\epsilon_{\mathbf{1}} = 1$, then $f^- = f^+ + f_{i_1}$ or $f^- - f^+ = f_{i_1}^+ - f_{i_1}^-$. This implies $f^- = f_{i_1}^+, f^+ = f_{i_1}^-$ so $\mathcal{X}^{f^+} - \mathcal{X}^{f^-} = -(\mathcal{X}^{f_i^+} - \mathcal{X}^{f_i^-}) \in \mathcal{I}_T$. A similar argument works if $A = 1$ and $\epsilon_{\mathbf{1}} = -1$.

In the general case $A > 1$. By induction on A, the monomial differences $\mathcal{X}^{f^+} - \mathcal{X}^{f^+ + \epsilon_1 f_{i_1}}$ and $\mathcal{X}^{f + \epsilon_1 f^+_{i_1}} - \mathcal{X}^{f^-}$ lie in the ideal generated by $\mathcal{F}$. So does their sum.

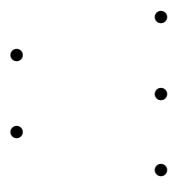
In the other direction, suppose that $\mathcal{F}$ generates $\mathcal{I}_T$. For $g, g' : \mathcal{X} \to \mathbb{N}$ such that $\sum_x (g(x) - g'(x))T(x) = 0$, there exists a representation

$$\mathcal{X}^g - \mathcal{X}^{g'} \quad = \quad \sum_{j=1}^{A} \mathcal{X}^{h_r} (\mathcal{X}^{f^+_{i_r}} - \mathcal{X}^{f^-_{i_r}}). \tag{3.2}$$

Here $h_r : \mathcal{X} \to \mathbb{N}$ and the polynomial on the right has no coefficients, since the proof of Lemma 3.1 works over the integers (hence is field independent). If $A = 1$, the identity (3.2) translates directly into property (b). For $A > 1$, we proceed by induction. From (3.2), $\mathcal{X}^g = \mathcal{X}^{h_r} \mathcal{X}^{f^{\pm}_{i_r}}$ for some r, say $\mathcal{X}^g = \mathcal{X}^{h_r} \mathcal{X}^{f^-_{i_r}}$. Then $g - f^-_{i_r}$ is non-negative and so $g + f_{i_r}$ is non-negative. Subtracting $\mathcal{X}^{h_r}(\mathcal{X}^{f^+_{i_r}} - \mathcal{X}^{f^-_{i_r}})$ from both sides of (3.2) and using $h_r + f^+_{ir} = g + f_{i_r}$ we get an expression for $\mathcal{X}^{g+f_{i_r}} - \mathcal{X}^{g'}$ having length $A-1$. By induction, $g + f_{i_r}$ can be connected to g' by allowable steps, and so (b) holds for all g, g'. □

EXAMPLE: We illustrate Theorem 3.2 for contingency tables. Here $K[\mathcal{X}]$ is the ring of polynomial functions on a generic I by J matrix $(\mathcal{X}_{ij})$. The toric ideal $\mathcal{I}_T$ is the ideal generated by the 2×2 minors $x_{ij}x_{k\ell} - x_{i\ell}x_{kj}$. The local changes induced by these binomials are precisely the $\pm\mp$ or $\mp\pm$ moves described in Section 1. These determinantal ideals have been the object of intense study by algebraists and geometers. Sturmfels (1991, 1992) gives further discussion and references.

Two other sets of moves are worth mentioning for this example. Let $K_{I\,J}$ be the complete bipartite graph on I and J nodes, so $K_{\mathbf{23}}$ appears as



Any cycle in $K_{I\,J}$ gives a possible move for the contingency table problem in an obvious way by adding and subtracting alternately along the cell entries determined by the edges in the cycle. Longer cycles move things further, so there might be a possibility for a speed up. As will emerge, these longer cycles have a natural algebraic interpretation; they are a

universal Gröbner basis for $\mathcal{I}_T$. In integer programming (cf. (Thomas 1993)) these moves constitute a *universal test set* for the *transportation problem.*

A second set of moves consists of using only pairwise adjacent squares as $x_{ij}x_{i+\mathbf{1},j+\mathbf{1}} - x_{i,j+\mathbf{1}}x_{i+\mathbf{1},j}$. These moves fail to connect the set $\mathcal{X}_{\underset{\sim}{r},\underset{\sim}{c}}$ in general. For example, the following pair of 2×3 tables have the same row and column sums but cannot be connected:

$$\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 1 \end{array} \qquad\qquad \begin{array}{ccc} 0 & 0 & 1 \\ 1 & 0 & 0 \end{array}$$

It can be proved that adjacent switches connect $\mathcal{X}_{\underset{\sim}{r},\underset{\sim}{c}}$ provided $\min\{r_i, c_j\} \geq 2$.

B. Gröbner Bases.

Any present day discussion of computing generating sets for a polynomial ideal is steeped in the language of Gröbner bases. This subsection presents the definitions, tailored to the present applications, and it explains how the ideas of Section 3.A can be put in this framework (so small problems can be routinely solved using computer algebra packages such as MAPLE or MACAULAY). The literature on these topics is vast. Fortunately, the undergraduate text book by Cox, Little, and O'Shea (1992) has just appeared.

B.1 TERM ORDERS: A *term order* is a linear order $\prec$ on $\mathbb{N}^n$ which satisfies $0 \preceq \alpha$ and $\alpha \preceq \beta$ implies $(\gamma + \alpha) \preceq (\gamma + \beta)$. A familiar example is lexicographic order (lex). This requires choosing an ordering of the variables. Graded lexicographic order (grlex) declares $\alpha \prec \beta$ if either $\Sigma\beta_i > \Sigma\alpha_i$ or $\Sigma\beta_i = \Sigma\alpha_i$ and the first non vanishing difference has $\beta_i - \alpha_i > 0$. (This was used in our proof of Lemma 3.1). An important variant is graded reverse lexicographic order (grevlex) which has $\alpha \prec \beta$ if either $\Sigma\beta_i > \Sigma\alpha_i$ or $\Sigma\beta_i = \Sigma\alpha_i$ and the first non-vanishing difference, working from the right, has $\beta_i - \alpha_i < 0$. Thus $(0, 2, 0) \prec (1, 0, 1)$ in grlex but $(1, 0, 1) \prec (0, 2, 0)$ in grevlex.

A large class of partial orders is given by choosing a weight vector $u \in \mathbb{N}^n$ and declaring $\alpha \prec \beta$ if $\alpha \cdot u < \beta \cdot u$. This can be made into a linear order by breaking ties using a second linear order or by a second (and third $\cdots$) weight order. It is known that any term order is the intersection of at most n weight orders (cf. (Weispfenning 1987)).

If $\mathcal{X}$ is a finite set, say $\mathcal{X} = \{x_{\mathbf{1}}, x_{\mathbf{2}}, \ldots . x_n\}$, then the functions $f : \mathcal{X} \to \mathbb{N}$ can be linearly ordered using a term order on the vectors $(f(x_{\mathbf{1}}), f(x_{\mathbf{2}}), \cdots, f(x_n))$. We write

$f \prec g$ for this order. This gives a linear order for the monomials in $K[\mathcal{X}]$ via $\mathcal{X}^f \prec \mathcal{X}^g$: $\Longleftrightarrow \quad f \prec g$. Every non-zero polynomial $p \in K[\mathcal{X}]$ contains a unique highest monomial with respect to $\prec$. This is called the *initial monomial* and denoted $in(p)$.

B.2 GRÖBNER BASES: Let $\mathcal{X}$ be a finite set. Linearly order the elements of $\mathcal{X}$ and choose a term order $\prec$ on functions from $\mathcal{X}$ to $\mathbb{N}$. Let $T : \mathcal{X} \to \mathbb{Z}^d$ be given. A *Gröbner basis* for T with respect to $\prec$ is a finite collection of functions $f_i : \mathcal{X} \to \mathbb{Z}$, $1 \le i \le L$ such that

$$\text{(3.3)} \qquad \begin{array}{ll} (a) & \displaystyle\sum_x f_i(x)T(x) = 0 \\ (b) & \text{For any two functions } f, g : \mathcal{X} \to \mathbb{N} \quad \text{with} \quad \sum f(x)T(x) = \sum g(x)T(x) \end{array}$$

there exist two sequences f_{i_r}, $1 \le r \le A$, f_{j_μ}, $1 \le \mu \le B$ such that

$$f + \sum_{r=1}^{\alpha+1} f_{i_r} \preceq f + \sum_{r=1}^{\alpha} f_{i_r}, \quad g + \sum_{\mu=1}^{\beta+1} f_{i_r} \preceq f + \sum_{\mu=1}^{\beta} f_{j_\mu} \quad \text{for } 0 \le \alpha \le A-1, 0 \le \beta, \le B-1$$

$$\text{and} \qquad f + \sum_{r=1}^{A} f_{i_r} \quad = \quad g + \sum_{\mu=1}^{B} f_{j_\mu},$$

with all of the above functions non-negative $\mathcal{X} \to \mathbb{N}$.

REMARKS: 1. Clearly, a Gröbner basis satisfies (1.2). Indeed, we get from f and g to a common function by strictly decreasing moves. It follows that algorithms for finding Gröbner bases will solve our problem in Section 1.

2. The definition above has been tailored to the present application. Gröbner bases are usually defined for ideals $\mathcal{I} \subset K[\mathcal{X}]$ as a finite set of polynomials $\{p_1, \ldots, p_L\} \subset \mathcal{I}$ such that the ideal $\langle in(p_1), \ldots, in(P_L)\rangle$ generated by its initial terms equals $\langle in(p) \ : p \in \mathcal{I}\rangle$. For the toric ideal $\mathcal{I}_T$ defined in Section 3.A, it not hard to show that a collection of functions $f_i : \mathcal{X} \to \mathbb{Z}$ forms a Gröbner basis in the sense of (3.3) if and only if the monomial differences $\mathcal{X}^{f_i^+} - \mathcal{X}^{f_i^-}$ form a Gröbner basis for $\mathcal{I}_T$. See (Sturmfels 1991), (Conti & Traverso 1991) or (Thomas 1993) for details.

A collection of functions $\{f_i : \mathcal{X} \to \mathbb{Z}$ is called a *universal Gröbner basis* if it is a Gröbner basis for all term orders simultaneously. The fact that a finite universal Gröbner basis exists is not obvious; a combinatorial proof can be found in (Thomas 1993).

Given a fixed term order, then a Gröbner basis is called *reduced* if it is minimal with respect to inclusion and all paths in (3.3b) have minimal length. Cox, Little, and O'Shea (1992, p. 91) show that a reduced Gröbner basis exists and is unique.

EXAMPLE: Take $\mathcal{X} = \{x, y, z\}$ and $T(x) = T(y) = T(z) = 1$. The little fiber $\mathcal{X}_t$ is just the set of compositions of t into three non-negative parts. Take $x > y > z$ and lex order. The set $\{x - y, y - z\}$ is a Gröbner basis for $\mathcal{I}_T$. It is not reduced because the resulting path from x to z has length two, which is not minimal. The set $\{x - z, y - z\}$ is the reduced Gröbner basis. The set $\{x - y, x - z, y - z\}$ is a universal Gröbner basis.

EXAMPLE (2-WAY TABLES): The basic $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ moves in all possible row and column positions form a reduced Gröbner basis for the following order: linear order pairs (i, j) row wise $(1, 1) \succ (1, 2) \succ \cdots \succ (1, J) \succ (2, 1) \succ \cdots \succ (I, J)$. Then use lexicographic order on functions. Sturmfels (1991, p. 260) gives an example of a term order for which these basic moves do not form a Gröbner basis (so they are not universal). It is known that the circuits described at the end of 3.A form a universal Gröbner basis for the 2-way tables.

One important and well-known fact about Gröbner bases is the elimination property of lexicographic term orders. Let $\mathcal{X}'$ be subset of $\mathcal{X}$, and let $K[\mathcal{X}']$ be the corresponding polynomial subring of $K[\mathcal{X}]$. We write T' for the restriction of the map T to $\mathcal{X}'$. Algebraically, this corresponds to forming the *elimination ideal*

$$\mathcal{I}_{T'} \quad = \quad \mathcal{I}_T \ \cap \ K[\mathcal{X}']. \tag{3.4}$$

The following is a direct translation of Theorem 2 in (Cox, Little, O'Shea, 1992, p. 114):

PROPOSITION 3.4. Order the set $\mathcal{X}$ such that each element of $\mathcal{X}'$ comes before each element of $\mathcal{X} \setminus \mathcal{X}'$, and let $\mathcal{G}$ be a Gröbner basis for T with respect to the resulting lexicographic term order. Then $\{g \in \mathcal{G} \ : \ supp(g) \subseteq \mathcal{X}'\}$ is a Gröbner basis for T'.

The following corollary will be useful for constructing our Markov chains in those cases where structural zeros may be present (see Section 4.E)

COROLLARY 3.5. If $\mathcal{U}$ is a universal Gröbner basis for T, then, for every subset $\mathcal{X}'$ of $\mathcal{X}$, the restriction of $\mathcal{U}$ to $\mathcal{X}'$ is a Gröbner basis for $T' = T|_{\mathcal{X}'}$.

C. How to compute a Gröbner Basis for a Toric Ideal. In his 1965 dissertation Bruno Buchberger found an algorithm for computing the reduced Gröbner basis of an ideal from any generating set. This algorithm is of striking elegance and simplicity. We refer to (Cox, Little, O'Shea, 1992) for a general introduction and (Thomas 1993) for a self-contained combinatorial treatment of the toric case. For our purposes the Buchberger algorithm can be treated as a black box. Implementations are readily available in computer algebra systems such as AXIOM, MAPLE, MACSYMA, MATHEMATICA. All serious examples in this paper were computed using the program MACAULAY of Bayer and Stillman (1989). MACAULAY is fast, and available at no cost (via anonymous ftp from `zariski.harvard.edu`), but its interface requires a certain expertise and patience.

Let $\mathcal{X}$ be a finite set and $T : \mathcal{X} \to \mathbb{Z}^d$. The toric ideal $\mathcal{I}_T$ was defined in Section 3A, and in Lemma 3.1 we presented an infinite generating set. The following proposition shows how to compute a reduced Gröbner basis for $\mathcal{I}_T$.

PROPOSITION 3.6. Let $\mathcal{Y} = \{y_1, \cdots, y_d\}, \mathcal{Y}^- = \{y_1^-, \cdots, y_d^-\}$ be indeterminates. Given a term order for $\mathcal{X}$, extend it to a term order on $\mathcal{X} \cup \mathcal{Y} \cup \mathcal{Y}^-$ such that $z \succ x$ for all $x \in \mathcal{X}$, $z \in \mathcal{Y} \cup \mathcal{Y}^-$. In $K[\mathcal{X}, \mathcal{Y}, \mathcal{Y}^-]$ define $\mathcal{J}_T = \langle x - \mathcal{Y}^{T(x)}, x \in \mathcal{X}\,;\ y_i y_i^- - 1,\ 1 \leq i \leq d \rangle$. Then $\mathcal{I}_T = \mathcal{J}_T \cap K[\mathcal{X}]$, and the reduced Gröbner basis for $\mathcal{I}_T$ can be found by computing a reduced Gröbner basis for $\mathcal{J}_T$ and taking those output polynomials which only involve $\mathcal{X}$.

Here $\mathcal{Y}^{T(x)}$ abbreviates the monomial $\prod\{y_i^{T(x)_i} : T(x)_i > 0\} \cdot \prod\{(y_j^-)^{-T(x)_j} : T(x)_j < 0\}$. The variables y_i^- are needed to cope with the possibility of negative exponents. In all our stochastic applications, the image of the map T lies in $\mathbb{N}^d$, in which case it suffices to work with $\mathcal{J}_T = \langle x - \mathcal{Y}^{T(x)}, x \in \mathcal{X} \rangle \subset K[\mathcal{X}, \mathcal{Y}$. The proof of Proposition 3.6 is straightforward from the Elimination Theorem (Proposition 2.4). The method we propose is a special case of the Implicitization Algorithm given in (Cox, Little, and O'Shea, 1992, p. 128)

To illustrate Proposition 3.6 in a simple example, let us pretend we wish to find a Gröbner basis for the case of 3×3 contingency tables. Using the computer algebra system MAPLE, the following explicit sequence of commands will do the job:

```
> with(grobner);

> ideal := [ x11-y1*z1, x12-y1*z2, x13-y1*z3, x21-y2*z1, x22-y2*z2,
```

```
x23-y2*z3, x31-y3*z1, x32-y3*z2, x33-y3*z3 ];
> varlist := [y1,y2,y3,z1,z2,z3,x11,x12,x13,x21,x22,x23,x31,x32,x33];
> G := gbasis(ideal, varlist,plex);
```

After about one minute we see the output of 36 monomial differences on the screen. Deleting all expressions which contain `y1,y2,y3,z1,z2` or `z3`, we are left with our nine basic moves of type $\pm\mp$. We remark that MACAULAY does the same computation in less than one second. (But it takes somewhat longer to get used to MACAULAY's interface). We recommend trying the same MAPLE computation using the variable ordering

```
> varlist := [y1,y2,y3,z1,z2,z3,x11,x22,x33,x12,x13,x21,x23,x31,x32];
```

Among the output polynomials we see `x12 x23 x31 - x13 x21 x32`, witnessing the fact that the nine basic $\pm\mp$ moves are not a universal Gröbner basis: they fail to connect if structural zeros for `x11,x22,x33` are prescribed.

D. Degree bounds. An general problem in computational commutative algebra is to find bounds for the degrees of the polynomials that appear in Gröbner bases and minimal generating sets of ideals. Such bounds are important also for our applications: for instance, the fiber walk of Section 2D is based on the knowledge of an explicit degree bound. It is known that the degree can depend critically on the term order. For homogeneous ideals in generic coordinates, Bayer and Stillman (1987) have shown that grevlex order produces a Gröbner basis of smallest degree. Our toric ideals, however, are not in generic coordinates, so the conclusion of the Bayer-Stillman Theorem is sometimes false (but it's a good rule of thumb nevertheless, cf. Theorem 6.1).

Mayr and Meyer (1982) produced an ideal generated by monomial differences in n variables of degree at most d such that each Gröbner basis contains polyomials of degree $d^{\mathbf{2^n}}$. This lower bound matches the known upper bounds, which are also doubly-exponential in n. The Mayr-Meyer example is not a prime ideal, so it is not toric. For Gröbner bases of toric ideals $\mathcal{I}_T$ we will see below that the degree bounds are not doubly-exponential but singly-exponential in n.

Many naturally occurring examples have degrees which are quite small: for 2-way tables of size $I \times J$ the degree is 2 for generators and $max\{I, J\}$ for the universal Gröbner basis. In Theorem 6.1 we show that for permutation data on S_n there are Gröbner bases

of degree n. Sharp bounds are also available for binary logistic regression (cf. Section 5).

We now summarize the best known general degree bounds for Gröbner bases of toric ideals. In a subsequent paper (Diaconis & Sturmfels 1993) we show that the same problem for minimal generators is equivalent. Therefore we may here restrict ourselves to Gröbner bases. Let $\mathcal{X}$ be a finite set with $|\mathcal{X}| = n$. Let $T : \mathcal{X} \to \mathbb{Z}^d$ be given and suppose that the $d \times n$-matrix $\big(T(x) : x \in X\big)$ has rank r. Let $D(T)$ be the maximum absolute value of any $r \times r$-subdeterminant of this matrix.

THEOREM 3.7.

(a) The total degree of any polynomial in a reduced Gröbner basis for $\mathcal{I}_T$ is bounded above by $n \cdot (n-d) \cdot D(T)$.

(b) For fixed d, this degree is bounded by $\gamma_d \cdot D(T)$, for a constant γ_d depending on d.

The bound in (a) is polynomial in the coordinates of $T(x)$ and singly-exponential in n. It is proved in (Sturmfels 1991). The bound in (b) does not depend in any explicit fashion on n, the number of variables. It is proved in (Diaconis, Graham & Sturmfels 1993).

Concluding Section 3 we remark that our presentation does not do justice to the richness and utility of Gröbner bases. Yet, it does suffice for what we need; for more we refer to (Cox, Little, and O'Shea 1992) and the references given there.

4. Contingency Tables.

Two-way contingency tables have been a running example in the previous sections. In this section we treat three-way and higher tables. Section 4A develops Monte Carlo algorithms for exact tests of classical models of independence and conditional independence. Section 4B develops tests for the model of "no three-way interaction". Theory for sampling from a three-dimensional table with given line sums is developed in Section 4C. Finally, Section 4D discusses general hierarchical, graphical, and decomposable models.

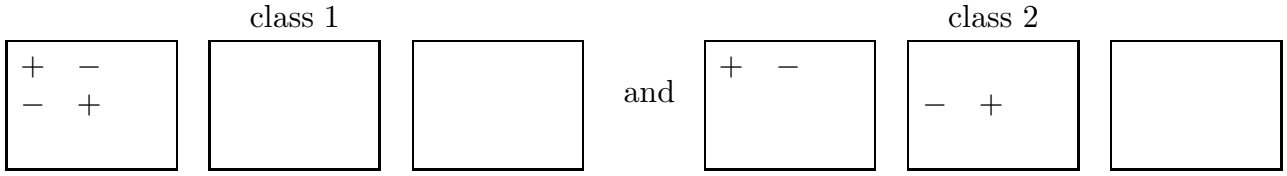
There is a vast modern literature on contingency tables. Agresti (1990), Bishop, Fineberg, and Holland (1975), Christensen (1990), and Haberman (1978) all treat the topics presented here and give surveys of the literature.

A. Three-way tables. A population of N people classified into three categories of I, J, K levels, respectively, leads to data n_{ijk}, $1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$. Under the multinomial sampling model the chance of falling into cell (i, j, k) is p_{ijk}. A variety of models for p_{ijk} are in wide use.

model	description	sufficient statistic
$p_{ijk} = p_{i..}p_{.j.}p_{..k}$	complete independence	$n_{i..}, n_{.j.}, n_{..k}$
$p_{ijk} = p_{i..}p_{.jk}$	independence of one variable	$n_{i..}, n_{.jk}$
$p_{ijk} = p_{i.k}p_{.jk}/p_{..k}$	conditional independence	$n_{i.k}, n_{.jk}$

The final widely used model, no three-way interaction, is discussed in Section 4B below. In each of the cases above, a test of adequacy of the model can be based on a chi-square or likelihood ratio test statistic S. Under the model the distribution of the data given the sufficient statistic is an appropriate hypergeometric. All of these cases fall into the framework of Sections 2 and 3. These cases *also* fall into the class of decomposable models discussed in section D below. It is known that decomposable models admit a simple sampling algorithm for the hypergeometric distribution. This is described in Lauritzen (1993) and implemented in the program CoCo. We give the basic moves for a Markov chain as a means of generating from the relevant uniform distribution.

Example. Complete independence: In this model we fix all face sums $n_{i..}, n_{.j.}, n_{..k}$, $1 \le i \le I$, $1 \le i \le J$, $1 \le k \le K$. There are two classes of moves which are depicted as



The moves are described algebraically, up to permutation of indices, as

$$x_{111}x_{122} - x_{112}x_{121} \qquad \text{and} \qquad x_{111}x_{222} - x_{112}x_{221}.$$

These generate an irreducible Markov chain. The ring map φ_T of Section 3 takes x_{ijk} to $u_i v_j w_k$. The associated ideal $\mathcal{I}_T$ is studied in algebraic geometry as the Segre embedding of the product of three projective spaces of dimension $I-1, J-1, K-1$. See Harris (1992).

Example. Independence of one variable: There are three choices here. For definiteness say that the variable i is independent of (j,k). Then we fix $n_{i\cdot\cdot}$ and $n_{\cdot jk}$, $1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$. An easy to implement Markov chain identifies the pairs (j,k) with a new variable ℓ, for $1 \le \ell \le L = JK$. Now consider the table as an I by L array and use the 2-dimensional moves as in the introduction.

Example. Conditional independence: Again there are three choices. For definiteness, say variables i and j are conditionally independent given k. Then, we fix $n_{i\cdot k}$ and $n_{\cdot jk}$. Here, for each fixed value of k, one has a two-dimensional face with k fixed. The walk proceeds independently as k walks in each of these tables.

These three walks were straightforward variations of the two-dimensional case. The following two subsections treat a model which is more difficult.

B. A $3 \times 3 \times 3$ example. Let N objects be classified into three categories with I, J, K levels respectively. The chance of an object falling into category (i,j,k) is p_{ijk}. A classical statistical hypothesis, called *no 3-way interaction*, gives rise to a test which can be carried out by generating a table from the hypergeometric distribution conditional on all two-dimensional faces. If the table entries are denoted n_{ijk}, the faces may be denoted $n_{\cdot jk}, n_{i\cdot k}, n_{ij\cdot}$, where $n_{\cdot jk} = \sum_i n_{ijk}$. Tests for no 3-way interaction are described by Birch (1963) or Bishop, Fineberg, and Holland (1975). We first treat the case $I = J = K = 3$.

It is natural to consider basic $2 \times 2 \times 2$ moves like

$$(4.1) \qquad \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} \qquad \begin{matrix} 0 & 0 & 0 \\ 0 & + & - \\ 0 & - & + \end{matrix} \qquad \begin{matrix} 0 & 0 & 0 \\ 0 & - & + \\ 0 & + & - \end{matrix}$$

There are 27 such moves; alas the chain they generate is not connected. Using the program MACAULAY, we ran the basic algorithm of Proposition 3.6. This involved computations in a polynomial ring with 54 variables (27 variables x_{ijk} for the table entries and 27 variables $y_{ij}^{(1)}, y_{ik}^{(2)}, y_{jk}^{(3)}$ for the three 3×3 margins). We found that a minimal set of generators consists of the 27 moves as in (4.1) and 54 moves of degree 6 like

$$(4.2) \qquad \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} \qquad \begin{matrix} 0 & - & + \\ + & 0 & - \\ - & + & 0 \end{matrix} \qquad \begin{matrix} 0 & + & - \\ - & 0 & + \\ + & - & 0 \end{matrix}.$$

The pattern in the last two layers can be permuted in 6 ways and the two layers can be placed in 9 ways. This gives rise to 54 moves.

In carrying out this computation, we ordered the cells (i, j, k) lexicographically and used the induced degrevlex term order on the monomials in the y_{ijk}. The reduced Gröbner basis for this order contains 110 basic moves: the 27 + 54 minimal generators above plus

$$(4.3) \qquad \text{28 relations of degree 7 like} \qquad \begin{matrix} 0 & 0 & 0 \\ 0 & - & + \\ 0 & + & - \end{matrix} \qquad \begin{matrix} + & 0 & - \\ - & + & 0 \\ 0 & - & + \end{matrix} \qquad \begin{matrix} - & 0 & + \\ + & 0 & - \\ 0 & 0 & 0 \end{matrix}.$$

$$(4.4) \qquad \text{one relation of degree 9 like} \qquad \begin{matrix} -2 & + & + \\ + & 0 & - \\ + & - & 0 \end{matrix} \qquad \begin{matrix} + & 0 & - \\ 0 & 0 & 0 \\ - & 0 & + \end{matrix} \qquad \begin{matrix} + & - & 0 \\ - & 0 & + \\ 0 & + & - \end{matrix}.$$

This computation required 52 hours and used 23 megabytes of memory on a SPARC 1. We emphasize that we used off-the-shelf software with no attempt to write fast code.

Example: Haberman (1978) reports data drawn from the 1972 national opinion research center on attitudes toward abortions among white Christian subjects. The part of the data to be analyzed here is a $3 \times 3 \times 3$ table. The first variable is type of Christian (Northern Protestant, Southern Protestant, Catholic). The second variable is education (low (less than nine years), medium (nine through twelve years), high (more than 12 years)). The third variable is attitude to nontherapeutic abortion (positive, mixed, negative).

The data appear as

	P	M	N						
L	9	16	41	8	8	46	11	14	38
M	85	52	105	35	29	54	47	35	115
H	77	30	38	37	15	22	25	21	42
	Northern Protestant			Southern Protestant			Catholic		

The rows index level of education, the columns index attitude. The data are here treated as a simple random sample of size 1,055 from the U.S. population in 1972. Let the chance of falling into cell (i, j, k) be p_{ijk}. We carry out a test of the "no three factor interaction" model. This specifies constant log odds:

$$\frac{p_{111}p_{ij1}}{p_{i11}p_{1j1}} = \frac{p_{11k}p_{ijk}}{p_{i1k}p_{1jk}} \quad \text{for all } 2 \le i \le I,\ 2 \le j \le J,\ 2 \le k \le K.$$

The maximum likelihood estimates of the cell entries under the model are found by iterative proportional fitting to be

12.01	14.43	39.58	9.436	12.25	40.27	6.552	11.32	45.13
85.75	52.51	103.8	36.55	24.17	57.27	44.68	39.32	113.0
73.24	31.06	40.66	34.01	15.58	24.45	31.77	19.36	36.87

The chi-square statistic for goodness of fit is 10.37. The usual asymptotics refer this to a chi-square distribution with $(I-1)(J-1)(K-1) = 8$ degrees of freedom. To calibrate the asymptotics, we ran the random walk described in Lemma 2.2 using the 110 moves described above to get the hypergeometric distribution. Every 500 steps a chi-square value was computed. One million chi-square values were accumulated. The resulting data are in remarkably good agreement with the chi-square distribution. Figure 2 shows a $p-p$ plot of the million chi-square values versus the chi-square (8) distribution. We conclude that the algorithm seems to work well, that the chi-square approximation seems very good (there is a small, systematic bias upward in Figure 2), and that the no-three factor model fits this data. Haberman (1978, Section 4.2) presents a serious analysis of this data along with data from subsequent years.

Figure 2

C. $a \times b \times c$ **tables.** For higher way tables there are an ever growing collection of models. We present a general discussion in Section 4D. Here we treat three way tables of format $a \times b \times c$, fixing all line sums $n_{\cdot jk}, n_{i\cdot k}, n_{ij\cdot}$. No neat description of the moves for general a, b, c is known to us. What we can give is a neat description for $2 \times a \times b$ tables. We begin with this and then show that things get very complicated if a, b, c are large.

For a $2 \times n \times n$-table consider the following move:

$$(4.5) \qquad \begin{array}{cccccc} + & - & 0 & 0 & \cdots & 0 \\ 0 & + & - & 0 & \cdots & 0 \\ 0 & 0 & + & - & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & + & - \\ - & 0 & 0 & \cdots & 0 & + \end{array} \qquad \begin{array}{cccccc} + & + & 0 & 0 & \cdots & 0 \\ 0 & - & + & 0 & \cdots & 0 \\ 0 & 0 & - & + & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & - & + \\ + & 0 & 0 & \cdots & 0 & - \end{array}$$

The product of symmetric groups $S_n \times S_n$ acts on the rows and columns. This gives $(n-1)!n!/2$ distinct permutations of (4.5). We call them the *basic moves of degree* $2n$. For $n \leq b \leq c$, any of these basic moves can be placed in a $2 \times b \times c$ array. There are $\binom{b}{n}\binom{c}{n}$ distinct ways to do this, so altogether we get $\sum_{n=2}^{b} \frac{(n-1)!n!}{2}\binom{b}{n}\binom{c}{n}$ basic moves for the $2 \times b \times c$ array. In (Diaconis and Sturmfels, 1993) we show that these basic moves form a minimal generating set which is at the same time a universal Gröbner basis.

The discussion above gives satisfactory results for $2 \times b \times c$ tables. We next present the little we know about $a \times b \times c$ tables. We write $\mathcal{I}_{a\,b\,c}$ for the toric ideal, which is the kernel of

the ring map $x_{ijk} \mapsto y_{ij}^{(1)} y_{ik}^{(2)} y_{jk}^{(3)}$. Let T_{abc} denote the $(ab+ac+bc) \times abc$-matrix with entries in $\{0,1\}$ which represents this map. The columns of T_{abc} are indexed by the variables x_{ijk}; each column has precisely three entries 1, namely, in the rows indexed by $y_{ij}^{(1)}$, $y_{ik}^{(2)}$ and $y_{jk}^{(3)}$. In other words, T_{abc} represents the linear map $\mathbb{Z}^{a \times b \times c} \to \mathbb{Z}^{a \times b} \oplus \mathbb{Z}^{a \times c} \oplus \mathbb{Z}^{b \times c}$, which takes 3-way tables to their 2-way margins. It is easy to see that the kernel of T_{abc} has rank $(a-1)(b-1)(c-1)$. Therefore rank of the matrix T_{abc} equals

$$r \quad = \quad rank(T_{abc}) \quad = \quad ab + ac + bc - a - b - c - 1 \tag{4.6}$$

Let $D(a,b,c)$ denote the largest absolute value of any $r \times r$-minor of the matrix T_{abc}.

THEOREM 4.1. Let a, b, c be positive integers with $3 \leq a \leq b \leq c$.

(a) A universal Gröbner basis for $\mathcal{I}_{abc}$ is given by all binomials $\mathcal{X}^{m+} - \mathcal{X}^{m-}$, $m \in \ker(T_{abc})$, of degree at most $a(a-1)b(b-1)c(c-1) \cdot D(a,b,c)$.

(b) We have the inequalities $\min(a,b,c) - 1 \leq D(a,b,c) \leq 3^{r/\mathbf{2}}$.

PROOF: Part (a) follows from Theorem 3.7 (a). To prove the upper bound in (b), we note that the integer $D(a,b,c)$ is the determinant of an $r \times r$-matrix which has at most three ones and otherwise zeros in each column. Hadamard's inequality implies that such a determinant has absolute value at most $3^{r/\mathbf{2}}$. For the lower bound in part (b) we use the fact that $D(a,b,c)$ is an upper bound for the degree of any variable in a circuit of I_{abc}. A *circuit* of I_{abc} is a binomial whose support is minimal with respect to inclusion (cf. (Sturmfels 1991, 1992)). It can be shown that the following binomial is a circuit for the $a \times a \times a$-table:

$$\begin{aligned} & \prod_{i=\mathbf{1}}^{a} x_{\mathbf{1},i,i} \cdot \prod_{j=\mathbf{2}}^{a-\mathbf{1}} (x_{j,\mathbf{1},\mathbf{1}}\, x_{j,j,j+\mathbf{1}}) \cdot \prod_{k=\mathbf{2}}^{a} (x_{a,\mathbf{1},k}\, x_{a,k,\mathbf{1}}) \\ - \quad & x_{a,\mathbf{1},\mathbf{1}}^{a-\mathbf{1}} \cdot x_{\mathbf{1},a,\mathbf{1}} \cdot \prod_{i=\mathbf{1}}^{a-\mathbf{1}} x_{\mathbf{1},i,i+\mathbf{1}} \cdot \prod_{j=\mathbf{2}}^{a-\mathbf{1}} (x_{j,\mathbf{1},j+\mathbf{1}}\, x_{j,j,\mathbf{1}}) \cdot \prod_{k=\mathbf{2}}^{a} x_{a,k,k} \end{aligned} \tag{4.7}$$

The variable $x_{a,\mathbf{1},\mathbf{1}}$ appears with degree $a-1$ in the circuit (4.7), and so we are done. □

Based on the special cases $2 \times b \times c$ and $3 \times 3 \times 3$ we wishfully conjecture:

CONJECTURE 4.2. $D(a,b,c) = \min(a,b,c) - 1$.

A relation in I_{abc} is called *critical* if it cannot be written as a polynomial linear combination of relations of lower degree. Thus the critical relations are needed for a generating set of I_{abc}. The *type* of a critical relation is the size of the smallest three-way-table which supports it. We now present two non-trivial examples of critical relations. These show that basic moves for $2 \times b \times c$-tables do not generate $\mathcal{I}_{abc}$ for large a, b, c.

(4.8) A critical relation of type $4 \times 4 \times 6$ is

$$\begin{aligned} & x_{\mathbf{131}}x_{\mathbf{241}}x_{\mathbf{142}}x_{\mathbf{322}}x_{\mathbf{123}}x_{\mathbf{433}}x_{\mathbf{214}}x_{\mathbf{344}}x_{\mathbf{235}}x_{\mathbf{415}}x_{\mathbf{316}}x_{\mathbf{426}} \\ -& x_{\mathbf{141}}x_{\mathbf{231}}x_{\mathbf{122}}x_{\mathbf{342}}x_{\mathbf{133}}x_{\mathbf{423}}x_{\mathbf{244}}x_{\mathbf{314}}x_{\mathbf{215}}x_{\mathbf{435}}x_{\mathbf{416}}x_{\mathbf{326}}. \end{aligned}$$

(4.9) A critical relation of type $3 \times 6 \times 9$ is

$$\begin{aligned} & x_{\mathbf{111}}x_{\mathbf{361}}x_{\mathbf{132}}x_{\mathbf{342}}x_{\mathbf{153}}x_{\mathbf{323}}x_{\mathbf{124}}x_{\mathbf{214}}x_{\mathbf{225}}x_{\mathbf{335}}x_{\mathbf{356}}x_{\mathbf{266}}x_{\mathbf{147}}x_{\mathbf{257}}x_{\mathbf{318}}x_{\mathbf{248}}x_{\mathbf{169}}x_{\mathbf{239}} \\ -& x_{\mathbf{161}}x_{\mathbf{311}}x_{\mathbf{142}}x_{\mathbf{332}}x_{\mathbf{123}}x_{\mathbf{353}}x_{\mathbf{114}}x_{\mathbf{224}}x_{\mathbf{325}}x_{\mathbf{235}}x_{\mathbf{256}}x_{\mathbf{366}}x_{\mathbf{157}}x_{\mathbf{247}}x_{\mathbf{218}}x_{\mathbf{348}}x_{\mathbf{139}}x_{\mathbf{269}}. \end{aligned}$$

We briefly explain the derivation of (4.8), (4.9). First note that we get zero after deleting the third subscript. This amounts to a non-trivial identity among six (resp. nine) carefully chosen 2×2 minors of a 4×4 matrix (resp. 3×6 matrix). Identities of this type are called *biquadratic final polynomials* in oriented matroid theory; see e.g. (Björner et.al., 1993, Section 8.5). They encode projective incidence theorems or non-realizability proofs of oriented matroids. The relation (4.8) encodes the biquadratic final polynomial for the Vamos matroid (Bokowski and Richter (1990)). The relation (4.7) encodes the biquadratic final polynomial for the Non-Pappus matroid (Bokowski and Richter-Gebert (1991)).

It was shown by Bokowski and Richter-Gebert (1991, Remark 3.6) that there exist arbitrarily large biquadratic final polynomials. ¿From this we can infer the following result.

PROPOSITION 4.3. Given any triple of interges $c \geq b \geq a - geq2$, there exists a critical relation of type $a' \times b' \times c'$ for some integers $a' \geq a, b' \geq b, c' \geq c$.

None of this says that is impossible to find some "nice" set of generators for I_{abc}; it only says that the simple moves we found so far do not suffice.

D. Higher way tables. Let Γ be a finite set. For each $\gamma \in \Gamma$, let I_γ be a finite set. Take $\mathcal{X} = \prod_{\gamma \in \mathbf{\Gamma}} I_\gamma$. This is the base space for data classified into $|\Gamma|$ categories with $|I_\gamma|$ levels of the $\gamma^{\mathbf{th}}$ category. Let $p(x)$ be the probability of falling into cell $x \in \mathcal{X}$. A log linear model can be specified by assuming

$$\log p(x) \quad = \quad \sum_{a \subseteq \mathbf{\Gamma}} \varphi_a(x).$$

Here the sum ranges over subsets $a \subseteq \Gamma$ and the notation $\varphi_a(x)$ means that the function φ_a only depends on x through coordinates in a. Thus $\varphi_{\emptyset}$ is a constant and $\varphi_{\mathbf{\Gamma}}$ is a completely general function. Specifying $\varphi_a \equiv 0$ for various classes of sets a determines various models.

Goodman's hierarchical models (Goodman (1970), Haberman (1978), Darroch, Lauritzen and Speed (1980)) begin with a class $\mathcal{C}$ of subsets $c_i \subset \Gamma$ with the assumption that no c_i contains another c_j. A *hierarchical model* is defined by specifying $\varphi_a \equiv 0$ unless $a \subseteq c$ for some $c \in \mathcal{C}$. For example, with $\Gamma = \{\alpha, \beta, \gamma\}$, the class $\mathcal{C} = \{\{\alpha, \beta\}, \{\alpha, \gamma\}, \{\beta, \gamma\}\}$ defines the no-three way interaction model of Sections 4B and 4C.

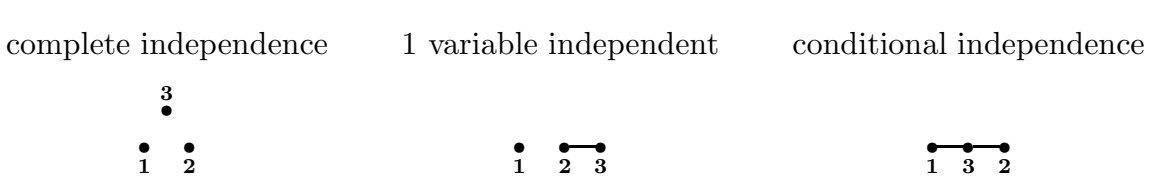
The sufficient statistics for a hierarchical model are $\{n(i_c)\}$, where c ranges over $\mathcal{C}$, $i_c \in \prod_{\gamma \in \mathcal{C}} I_\gamma$ and $n(i_c)$ is the sum over all x that agree with i_c in the coordinates determined by c. This can be expressed in the general form required in Section 2.

Hierarchical models have unique maximum likelihood estimates, which can be computed efficiently using Newton-Raphson or the iterated proportional fitting method. These lead to estimates $\hat{p}_c(x)$. If $\mathcal{C} \subset \mathcal{D}$ are two generating classes, an exact test for adequacy of model $\mathcal{C}$ within $\mathcal{D}$ may be based on the conditional distribution (under $\mathcal{C}$) of the chi-square statistic

$$\sum_x \frac{(N\hat{p}_{\mathcal{C}}(x) - N\hat{p}_{\mathcal{D}}(x))^{\mathbf{2}}}{N\hat{p}_{\mathcal{C}}}.$$

Graphical models are a subclass of hierarchical models obtained from a graph with vertex set Γ and edge set E. One specifies that the cliques of the graph (maximal complete subgraphs) are the generating class. These models can be characterized by conditional independence properties: for $a, b, c \subset \Gamma$, variables a and b are conditionally independent given c if and only if any path in the graph from a point in a to a point in b must pass

through c. The models of Section 4A are graphical:



The no three way interaction model is the simplest hierarchical model that is not graphical.

A particularly nice subclass of graphical models are the *decomposable models*. They arise from those graphs for which any cycle of length 4 or more contains a chord. Decomposable models allow closed form maximum likelihood estimates and simple algorithms for generating from the hypergeometric distribution. Glonek (1987) has conjectured that the moves determined by the generalized cross product ratio of Darroch and Speed (1983) lead to an irreducible Markov chain. Our results in Section 3 translate this conjecture into a question of combinatorial commutative algebra.

E. Structural Zeros and Incomplete Tables. Contingency tables sometimes have forced zero entries. For example, one of the categories may be pregnant males, or counts along the diagonal of a square table may be forced to be zero. See Bishop, Fineberg, and Holland (1975) or Haberman (1978, Chapter 7) for discussion and examples. It is straightforward to adopt our algebraic approach to deal with structural zeros.

The problem of structural zeros has the following formulation in the general setting of Sections 1-3. Let $\mathcal{X}'$ be a subset of $\mathcal{X}$ such that all observable functions $f : \mathcal{X} \to \mathbb{N}$ satisfy $f(x) = 0$ for $x \in \mathcal{X} \setminus \mathcal{X}'$. The question is to how to maintain this property during the random walks in Section 2. In other words, how to find a connecting set $f_1, \ldots, f_L$ in (1.2) which remains connecting when restricted to functions $f : \mathcal{X}' \to \mathbb{N}$. Corollary 3.5 tells us the answer:

COROLLARY 4.2. Suppose that the set of moves $f_1, \ldots, f_L$ is a universal Gröbner basis. Then the Markov chains in Section 2 remain irreducible if structural zeros are prescribed.

A general algorithm for computing universal Gröbner bases of toric ideals is presented in (Diaconis & Sturmfels 1993). To illustrate Corollary 4.2 in a simple example, we consider two-way tables with structural zeros. For an $I \times J$-table this is modelled by a bipartite graph on I and J points. There is an edge from i to j if and only if the (i, j) entry of the table is allowed to be non-zero. There is a simple description of a basic set of moves:

PROPOSITION 4.3. The circuits in a bipartite graph form a universal Gröbner basis.

The same result holds for arbitrary graphs and, more generally, for unimodular matroids. This result was proved in (Sturmfels 1992, Section 5).

REMARK: An amusing consequence of Proposition 4.3 is this: if there are no circuits in the bipartite graph, then any table is uniquely determined by its margins.

5. Logistic Regression.

Logistic regression is a standard technique for dealing with discrete data regression problems. Christensen (1990) or Haberman (1978) give background and details. We begin with binary data, then treat an example with equally spaced regressors. In Section 5C we treat multiple response models. A detailed algebraic study of the class of toric ideals arising from logistic regression is carried out in our subsequent papers (Diaconis & Sturmfels 1993), (Diaconis, Graham & Sturmfels 1993). These ideals have the remarkable property that each minimal generating set is automatically a universal Gröbner bases.

A. Binary Data. For each of N subjects a binary indicator Y and a vector of covariates z is observed. We assume that the covariates z are taken from a fixed finite subset $\mathcal{A}$ of $\mathbb{Z}^d$. A logistic model specifies a log-linear relation of form

$$P(\,Y = 1 \,|\, z\,) \quad = \quad e^{z\cdot\beta}/(1 + e^{z\cdot\beta})$$

where the parameter vector $\beta \in \mathbb{R}^d$ is to be estimated. With N subjects the likelihood function is

$$\prod_{i=1}^{N} e^{Y_i(z_i\cdot\beta)}/(1 + e^{z_i\cdot\beta}).$$

Let $n(z)$ be the number of indices $i \in \{1, \ldots, N\}$ with $z_i = z$, and let $n_{\mathbf{1}}(z)$ be the number of $i \in \{1, \ldots, N\}$ with $z_i = z$ and $Y_i = 1$. The collection $\{n(z)\}_{z\in\mathcal{A}}$ and the sum $\sum_z z n_{\mathbf{1}}(z)$ together are sufficient statistics (they determine the likelihood function). Our objective is to give random walk algorithms for generating data sets with these sufficient statistics.

To put the problem into the notation of the previous sections, let $\mathcal{X} = \{(0, z), (1, z), z \in \mathcal{A}\}$, and let $T : \mathcal{X} \to \mathbb{Z}^{d+|\mathcal{A}|}$ be defined by

$$\begin{aligned} T(0, z) \quad &= \quad (\,0\,;\, 0, \cdots, 0, 1, 0, \ldots, 0) \\ T(1, z) \quad &= \quad (\,z\,;\, 0, \cdots, 0, 1, 0, \ldots, 0) \end{aligned} \tag{5.1}$$

where there is a single 1 in the last $|Z|$ coordinates at the $z^{\mathbf{th}}$ position. Then for a given data $f : \mathcal{X} \to \mathbb{N}$, the sum $t = \sum_{x\in\mathcal{X}} f(x)T(x)$ fixes the sufficient statistics.

This general problem can be solved using the techniques of Sections 2-4. In what follows we restrict ourselves to developing tools for widely applicable special cases.

B. Equally spaced covariates. Haberman (1978, Chapter 7) gives data from the 1974 social science survey on men's response to the question "Women should run their homes and leave men to run the country." Let $Y = 1$ if the respondent "approves" and $Y = 0$ otherwise. For each respondent the number i of years in school is reported, $1 \le i \le 12$. The data are

Table 2. Men's response to "Women should run their homes and leave men to run the country" (1974/75). With years of education i.

i	0	1	2	3	4	5	6	7	8	9	10	11	12
$n_{\mathbf{1}}(i)$	4	2	4	6	5	13	25	27	75	29	32	36	115
$n(i)$	6	2	4	9	10	20	34	42	124	58	77	95	360
$p(i)$	.66	1	1	.66	.5	.65	.74	.64	.60	.50	.42	.38	.32

Here $n_{\mathbf{1}}(i)$ is the number "approving" and $n(i)$ is the total number in the sample with i years of education. Also shown are $p(i) = n_{\mathbf{1}}(i)/n(i)$, the proportion approving. These proportions seem to decrease with years of education. It is natural to fit a logistic model of form

$$P(\,Y = 1 \,|\, i\,) = e^{\alpha+i\beta}/(1 + e^{\alpha+i\beta}). \tag{5.2}$$

This falls into the framework above with $d = 2$, $\mathcal{A} = \{(1,1),(1,2),\cdots,(1,12)\}$. The sufficient statistics to be preserved are

$$\{n(i)\}_{i=\mathbf{0}}^{\mathbf{12}}, \quad \sum_{i=\mathbf{1}}^{\mathbf{12}} n_{\mathbf{1}}(i), \quad \sum_{i=\mathbf{1}}^{\mathbf{12}} n_{\mathbf{1}}(i)\cdot i. \tag{5.3}$$

A randomization test with these statistics fixed would be appropriate in testing the linear logistic model (5.2) against the non-parametric alternative $P(Y = 1 \,|\, i) = \theta_i$.

For the data of Table 2, the maximum likelihood estimates of α and β in the model (5.2) are $\hat\alpha = 2.1959, \hat\beta = -.2440271$. The chi-squared statistic for goodness of fit is $\sum_{i=\mathbf{1}}^{\mathbf{12}}(n\hat p(i) - n_{\mathbf{1}}(i))^{\mathbf{2}}/n\hat p(i) = 8.91$. The classical asymptotics for this problem calibrate this value with the chi-square (10) distribution. The uneven nature of the counts with some counts small gives cause for worry about the classical approximation. We ran the

basic random walk to check this approximation. A minimal ideal basis for this problem involves 8,569 basis elements. The walk was run, tilted to the hypergeometric distribution as in Lemma 2.2. A chi-square value was computed why 50 steps with 1000 values recorded in total. The observed value falls essentially at the median of the recorded values (their mean is 10.3). The values show good agreement with a chi-square (10) distribution as shown in Figure 4 below.

Figure 4

$\underline{P} - \underline{P}$ plot of 1000 random walk values of $\chi^{\mathbf{2}}$ versus a chi-square (10)

We conclude that the chi-square approximation is in good agreement with the conditional distribution and that the model (5.2) fits the data in Table 2.

In the remainder of Section 5B we give a combinatorial description of the basic moves to generate random data with fixed values of the statistics (5.3). We replace "12" by a parameter n, and assume without loss of generality that i ranges over $1 \le i \le n$.

DEFINITION. A *graded partition identity* with parameter n consists of $a_{\mathbf{1}}, \cdots, a_r, b_{\mathbf{1}}, \cdots, b_r$ with $a_i, b_j \in \mathbb{N},\ 1 \le a_i, b_j \le n$, and $a_{\mathbf{1}} + \cdots + a_r = b_{\mathbf{1}} + \cdots + b_r$. The identity is called *primitive* if no proper subset sum of the a_i equals a subset sum of the b_j. The *degree* of such an identity is $2r$.

For instance, for $n = 3$ it is easy to see that $1+3 = 2+2$ is the only primitive identity. When $n = 4$, the primitive identities are

$$3+3 = 2+4,\ 2+3 = 1+4,\ 1+3 = 2+2,\ 1+1+4 = 2+2+2,\ 3+3+3 = 1+4+4.$$

When $n = 5$ there are 16 primitive identities with two having maximal degree 8. When $n = 6$ there are 51 primitive identities, four having maximal degree 10. When $n = 7$, there are 127 primitive identities, two having maximal degree 12. Diaconis, Graham and Sturmfels (1993) prove that the maximal degree is $2(n-1)$, and that there are $\varphi(n-1)$ (Euler's function) primitive identities of maximal degree.

A graded partition identity corresponds to a "move" by removing an item from the $Y = 1$ row of columns $a_1, a_2, \cdots, a_r$ and adding an item to the $Y = 1$ row of columns $b_1, b_2, \cdots, b_r$ (or doing the opposite).

EXAMPLE: For $n = 4$, the move $3 + 3 + 3 = 1 + 4 + 4$ changes

	1	2	3	4			1	2	3	4
$n_1(i)$	10	20	30	40	to	$n_1(i)$	11	20	27	42
$n(i)$	100	100	100	100		$n(i)$	100	100	100	100

Diaconis and Sturmfels (1993) show that the moves corresponding to the primitive graded partition form a minimal ideal basis which is also a universal Gröbner basis. The bounds above allow the fiber walks of Section 2C to be carried out for relatively large n.

C. Multinomial response models. The following class of models does not fit under the umbrella of Section 5A. Let Y take values in $\{1, 2, \cdots, J\}$ with

$$P(Y = j \,|\, z) \;=\; C(z)e^{\alpha_j \cdot z}, \qquad C^{-1}(z) = \sum_{j=1}^{J} e^{\alpha_j z}.$$

Here $z = (z_1, \ldots, z_d)$ is a covariate vector from a fixed finite subset $\mathcal{A} \subset \mathbb{Z}^d$, and $\alpha_j \in \mathbb{R}^d$, $1 \le j \le J$. Data from a sample of size N can be presented as a $J \times n$ matrix u, whose entry $u(j, z)$ is the number of observed values j with covariate vector z. Let $U(z) = \sum_j u(j, z)$. The sufficient statistics are the column sums $\{U(z)\}_{z \in \mathcal{A}}$ and the $J \times d$ entries $\sum_{z \in \mathcal{A}} u(j, z_k) z_k$.

EXAMPLE: For the model $P(Y = j \,|\, i) = C(i)e^{\alpha_j + \beta_j i}$, $j = 1, 2, 3$, $i = 1, 2, \cdots, n$, the data would be a $3 \times n$ array with entries $u(j, (1, i))$. The sufficient statistics are $\{U(1,1), U(1,2), \cdots, U(1,n)\}$ and

$$\begin{array}{ll} \sum_i u(1,(1,i)), & \sum_i i \cdot u(1,(1,i)), \\ \sum_i u(2,(1,i)), & \sum_i i \cdot u(2,(1,i)), \\ \sum_i u(3,(1,i)), & \sum_i i \cdot u(3,(1,i)). \end{array}$$

We found that the full algebraic machinery of Section 3 is needed to deal with multiple logistic regression. There seems to be no set of "obvious" generators which connect the little fibers. As an example, for trivariate logistic regression with $n = 5$ a minimal set of generators has 96 elements:

$$\text{21 of degree 4 such as} \quad \begin{pmatrix} 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & -1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\text{51 of degree 6 such as} \quad \begin{pmatrix} 0 & 1 & 0 & -3 & 2 \\ 0 & -1 & 0 & 3 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\text{18 of degree 8 such as} \quad \begin{pmatrix} -1 & -2 & 0 & -2 & 1 \\ 2 & -3 & 0 & 1 & 0 \\ -1 & 1 & 0 & 1 & -1 \end{pmatrix},$$

$$\text{6 of degree 10 such as} \quad \begin{pmatrix} 0 & -1 & 0 & 3 & -2 \\ 2 & -2 & 0 & -2 & 2 \\ -2 & 3 & 0 & -1 & 0 \end{pmatrix}.$$

6. Spectral Analysis.

A version of spectral analysis suitable for permutation data was introduced in (Diaconis 1989). This generalizes the usual discrete Fourier transform analysis of time series. An introduction by example is given in Section 6A. In Section 6B we prove that appropriate Markov chains can be found with Gröbner bases having small degree. This uses a result of Stanley (1980), and the connection between Gröbner bases and triangulations of the convex polytope $conv\{T(x) : x \in \mathcal{X}\}$ developed in (Sturmfels 1991).

A. Spectral analysis of permutation data. Let S_n denote the group of permutations of n items. A data set consists of a function $f : S_n \to \mathbb{N}$, where $f(\pi)$ is the number of people choosing the permutation π. One natural summary of f is the $n \times n$-matrix $t = (t_{ij})$, where t_{ij} is the number of people ranking item i in position j. This is only a partial summary, since $n!$ numbers are compressed into $n^{\mathbf{2}}$ numbers. A sequence of further summaries was described in Diaconis (1989). These arise from a decomposition

$$L(S_n) = V_{\mathbf{0}} \oplus V_{\mathbf{1}} \oplus V_{\mathbf{2}} \oplus \cdots \oplus V_k.$$

On the left is $L(S_n)$, the vector space of all real-valued functions on S_n. On the right is an orthogonal direct sum of subspaces of functions. The summary t amounts to the projection onto $V_{\mathbf{0}} \oplus V_{\mathbf{1}}$. It is natural to look at the squared length of the projection of the original data set f into the other pieces to help decide if further projections need be considered.

As an example, Croon (1989) reports responses of 2,262 German citizens who were asked to rank order the desirability of four political goals:

1. maintain order
2. give people more say in government
3. fight rising prices
4. protect freedom of speech.

The data appear as

1234	137	2134	48	3124	330	4123	21	
1243	29	2143	23	3142	294	4132	30	
1324	309	2314	61	3214	117	4213	29	
1342	255	2341	55	3241	69	4231	52	
1423	52	2413	33	3412	70	4312	35	
1432	93	2431	59	3421	34	4321	27	
	875		279		914		194	2262

Thus 137 people ranked item 1 first, 2 second, 3 third and 4 fourth. The marginal totals show people thought item 3 most important (914 ranked it first). The first order summary $t = (t_{ij})$ is the 4×4-matrix

		item		
	875	279	914	194
	746	433	742	341
Position	345	773	419	725
	296	777	187	1002

The first row shows the number of people ranking a given item first. The last row shows the number of people ranking a given item last. Here we see what appears to be some "hate vote" for items 2 and 4, an indication that people vote against these items.

The data was collected in part to study if the population could be usefully broken into "liberals" who might favor items 2 and 4, and "conservatives" who might favor items 1 and 3. To investigate further we give the decomposition of the space of all functions $L(S_4)$ into an orthogonal direct sum

	$L(S_4)$	$=$	V_0	$\oplus$	V_1	$\oplus$	V_2	$\oplus$	V_3	$\oplus$	V_4
dim	24		1		9		4		9		1
length^2			462		381		268		48		4

Here, V_0 is the 1-dimensional space of constant functions, V_1 is a 9-dimensional space of "first order functions" spanned by $\pi \longmapsto \delta_{i\pi(j)}$ and orthogonal to V_0. The projection of f onto $V_0 \oplus V_1$ is equivalent to the first order summary given above. The space V_2 is a space of "unordered second order functions" spanned by $\pi \longmapsto \delta_{\{i,i'\},\{\pi(j),\pi(j')\}}$ and orthogonal to $V_0 \oplus V_1$. The space V_3 contains "ordered second order functions" and V_4 is a 1-dimensional space recording the differences between even and odd permutations. Further details are in Diaconis (1988, 1989).

Below each subspace is shown the squared length of the projection of the original data f. The first two subspaces V_0 and V_1 pick up much of the total sum of squares. The projection onto V_2 has norm 268 which seems moderately large. To investigate if this 268

is forced by the first order statistics or an indication of interesting structure, we performed the following experiment: Using a random walk detailed below, 100 independent data sets $f : S_4 \to \mathbb{N}^4$ with the same first order summary $t = (t_{ij})$ were chosen from the uniform distribution. For each data set, the squared length of its projection onto V_2 was calculated. The median squared length was 244 with upper and lower quantiles 268 and 214. We see that the moderately large value 268 is typical of data sets with first order statistics t and nothing to get excited about. For further analysis of this data see Bökenholt (1993).

The random walk was based on a Gröbner basis formed in the following way. Let $\mathcal{X} = S_4$, and let $T(\pi)$ be the 4×4 permutation matrix with (i, j)-entry $\delta_{i\pi(j)}$; this is one if item j is ranked in position i and zero otherwise. Given a function $f : \mathcal{X} \to \mathbb{N}$, then the 4×4-matrix

$$t \quad = \quad \sum_{\pi \in S_n} f(\pi) T(\pi)$$

is the first order summary reported above. We identify f with the monomial $\prod_\pi x_\pi^{f(\pi)}$ in the variables $x_\pi = [\pi_1 \pi_2 \pi_3 \pi_4]$, $\pi \in \mathcal{X}$. The permutation group was ordered using lex order ($1234 > 1243 > \cdots > 4321$). Then grevlex order was used on monomials in $K[\mathcal{X}]$. The computer program MACAULAY found a Gröbner basis containing 199 binomials. There were 18 quadratic relations (example [3421][4312] - [3412][4321]); 176 cubic relations (example, [4123][4231][4312] - [4132][4213][4321]) and five quartic relations (example, [1342][2314][2431][3241] - [1234][2341]^2[3412]). The walk was performed by repeatedly choosing a relation at random and adding and subtracting from the current function according to the relation or its negative. The walk was sampled every thousand steps until 100 functions had accumulated.

It is worth recording that a similar undertaking for S_5 led to a huge number of Gröbner basis elements (1,050 relations of degree 2 and 56,860 of degree 3). This is why the fiber walks of Section 2C were developed. These walks work for permutation data sets up to S_{10} or beyond. Their feasibility is guaranteed by the bound to be derived in Threorem 6.1.

B. Toric ideals for permutation data. We write x_π for the indeterminate associated with $\pi \in \mathcal{X} = S_n$ and $t_{i,j}$ for the indeterminate associated with the entries in the permu-

tation matrix. The ring homomorphism φ_T of Section 3 here becomes

$$\begin{array}{rcl} \varphi : K[\mathcal{X}] & \longrightarrow & K[\,t_{ij},\ 1 \le i,j \le n] \\ x_\pi & \longmapsto & \prod_{i=1}^{n} t_{i,\pi(i)} \end{array} \tag{6.1}$$

We are interested in (Gröbner) bases for the ideal $\mathcal{I} = \ker(\varphi)$. The main result is

THEOREM 6.1. Let $\succ$ be any of the $(n!)!$ graded reverse lexicographic term orders on $K[\mathcal{X}]$. The reduced Gröbner bases consists of homogeneous monomial differences of degree $\le n$.

PROOF: We fix one of the $(n!)!$ linear orders on S_n and let $\succ$ denote the resulting graded reverse lexicographic term order. Let Ω be the convex polytope of $n \times n$ doubly stochastic matrices (the Birkhoff polytope). This is the convex hull of the vectors $T(\pi)$ in $\mathbb{R}^{n^2}$. There is a close relation between triangulations of the convex hull of $T(x)$ and Gröbner bases. This is developed by Sturmfels (1991). It allows us to use results of Stanley (1980) on triangulations of Ω. The first step is to show that

(6.2) the initial ideal init $(\mathcal{I})$ is generated by square-free monomials.

Stanley (1980, Example 2.11(b)) has shown that the Birkhoff polytope Ω is *compressed*. This means that the *pulling triangulation* of Ω, which is determined by sequentially coning over vertices of Ω in the specified linear order, results in a decomposition into simplices of unit volume. Sturmfels (1991, Corollary 5.2) has shown that pulling triangulations correspond to grevlex initial ideals. Under this correspondence, triangulations into unit simplicies are identified with square-free initial ideals. This completes the proof of (6.2).

To prove the theorem, let $\mathcal{X}^f = \prod_\pi x_\pi^{f(\pi)}$ be one of the minimal square-free generators of the initial monomial ideal init$(\mathcal{I})$. Such a monomial is called *minimally non-standard*. (A monomial is *standard* if it does not lie in init$(\mathcal{I})$, it is non-standard otherwise and minimally non-standard if no proper divisor lies in init$(\mathcal{I})$). Let $\mathcal{X}^f - \mathcal{X}^g \in \mathcal{I}$ be a relation having leading monomial $\mathcal{X}^f$. The monomials $\mathcal{X}^f$ and $\mathcal{X}^g$ must be relatively prime. For, if x_π were a common factor then n$(\mathcal{X}^f - \mathcal{X}^g)/x_\pi \in \mathcal{I}$, because $\mathcal{I} = ker(\phi)$ is a prime ideal, and then $\mathcal{X}^f/x_\pi \in$ init$(\mathcal{I})$ which contradicts our choice.

Let x_π be the smallest variable which divides the trailing term $\mathcal{X}^g$. Then x_π does not divide the leading term $\mathcal{X}^f$. On the other hand,

$$\varphi(x_\pi) = \prod_{i=1}^{n} t_{i,\pi(i)} \qquad \text{divides} \qquad \varphi(\mathcal{X}^g) = \varphi(\mathcal{X}^f) = \prod_{\sigma \in S_n} \prod_{i=1}^{n} t_{i,\sigma(i)}^{f(\sigma)}.$$

Hence, for each $i \in \{1, \ldots, n\}$ there exists a permutation σ with $\sigma(i) = \pi(i)$ and $f(\sigma) \geq 1$. Let $\mathcal{X}^{f'}$ denote the product (without repetitions) of the corresponding n variables x_σ. By construction, $\mathcal{X}^{f'}$ is a monomial of degree $\leq n$ which divides $\mathcal{X}^f$. Moreover, in the chosen ordering, the variable x_π is smaller than any of the variables appearing in $\mathcal{X}^{f'}$.

We claim that $\mathcal{X}^{f'}$ is not standard. Consider the monomial $\varphi(\mathcal{X}^{f'})/\varphi(x_\pi)$ in the variables t_{ij}. Its exponent matrix is non-negative with all row and column sums equal. Birkhoff's theorem implies it is a non-negative integer linear combination of permutation matrices. Hence, $\varphi(\mathcal{X}^{f'})/\varphi(x_\pi)$ is a monomial which lies in the image of the ring map φ. Let $\mathcal{X}^h$ be any preimage. Then $\mathcal{X}^{f'} - x_\pi \cdot \mathcal{X}^h$ lies in $\mathcal{I}_n$. Here $\mathcal{X}^{f'}$ is the grevlex leading term since all of its variables are higher than x_π.

We conclude that $\mathcal{X}^{f'}$ is standard and is a factor of the minimally non-standard monomial $\mathcal{X}^f$. Therefore $\mathcal{X}^f = \mathcal{X}^{f'}$ is a monomial of degree $\leq n$. This shows that $init(\mathcal{I})$ is generated by square-free monomials of degree $\leq n$. The reduced Gröbner basis for $\mathcal{I}$ is given by $\mathcal{X}^{f_i} - \mathcal{X}^{g_i}$, where the $\mathcal{X}^{f_i}$ are the minimal generators of $init(\mathcal{I})$ and the $\mathcal{X}^{g_i}$ are standard (cf. Cox, Little, O'Shea (1992, Section 2.5)). □

REMARKS: 1. The conclusion of Theorem (6.1) and fact (6.2) only hold for graded reverse lexicographic order. Other term orders can require much larger Gröbner bases.

2. Stanley's result, used to prove (6.2), has the following direct combinatorial interpretation: let t be any $n \times n$ matrix with non-negative integer entries and constant row and column sums. Order the permutation group S_n and repeatedly subtract the associated permutation matrices until this leads to negative entries. Any order will end in the zero matrix without getting stuck. In fact, this combinatorial process is equivalent to the normal form reduction with respect to the above reduced Gröbner basis $\{\mathcal{X}^{f_i} - \mathcal{X}^{g_i}\}$.

FINAL REMARK: The random walk was used above to quantify a small part of the data analysis. A similar walk would be used to give an indication of the variability of the 2nd

order effects determined by the projection onto $V_{\mathbf{2}}$ (see the example in Diaconis (1989, Section 2). Similar analysis could be carried out for analyses conditional on the projection on other pieces. Finally, there are other settings where these ideas can be used: homogeneous spaces (such as partially ranked data) and other groups (such as $\mathbb{Z}_{\mathbf{2}}^d$ used for panel studies or item analysis), see Diaconis (1988, Chapter 7).

Acknowledgements. We thank Anders Björner for arranging the Combinatorics Year 1991/92 at the Mittag-Leffler Institute, Stockholm, which allowed this work to begin. Thanks to David des Jardins, Anil Gangolli, Ron Graham, David Johnson, Steffan Lauritzen, Bruce Levin, Juin Liu, Brad Mann, Mike Stillman, Rekha Thomas, Alan Zaslowzki and Günter Ziegler for their help. Both authors acknowledge partial support by the National Science Foundation. The second author is also supported by a David and Lucile Packard Fellowship.

References

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

2. Agresti, A. (1992). A Survey of Exact Inference for Contingency Tables. *Statist. Sci.* **7**, 131-177.

3. Aldous, D. (1987). On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing. *Prob. in Eng. and Info. Sci.* **1**, 33-46.

4. Baglivio, J., Olivier, D. and Pagano, M. (1988). Methods for the analysis of contingency tables with large and small cell counts. *Jour. Amer. Statist. Assoc.* **83**, 1006-1013.

5. Baglivio, J., Olivier, D. and Pagano, M. (1992). Methods for Exact Goodness-of-fit Tests. *Jour. Amer. Statist. Assoc.* **87**, 464-469.

6. Baglivio, J., Olivier, D. and Pagano, M. (1993). Analysis of Discrete Data: Rerandomization Methods and Complexity. Technical report, Dept. of Math., Boston College.

7. Barndorff-Nielsen, O. (1978), *Information and Exponential Families in Statistical Theory.* Wiley, New York.

8. Bayer, D. and Stillman, M. (1987). A theorem on refining division orders by the reverse lexicographic order. *Duke J. Math.* **55**, 321-328.

9. Bayer, D., and Stillman, M. (1989). MACAULAY: A computer algebra system for algebraic geometry, available via *anonymous ftp* from `zariski.harvard.edu`.

10. Birch, B.W. (1963). Maximum likelihood in three-way contingency tables. *Jour. Roy. Statist. Soc.*, **B 25**, 220-233.

11. Bishop, Y., Fienberg, S., Holland, P. (1975). *Discrete Multivariate Analysis*, MIT Press, Cambridge.

12. Björner, A.. Las Vergnas, M., Sturmfels, B., White, W., and Ziegler, G. (1993). *Oriented Matroids*, Cambridge University Press, Cambridge.

13. Bokowski, J. and Richter, J. (1990). On the finding of final polynomials. *European Jour. Combinatorics* **11**, 21-34.

14. Bokowski, J. and Richter-Gebert, J. (1991). On the classification of non-realizable oriented matroids, part II: preprint, T.H. Darmstadt.

15. Bökenholt, U. (1993). Applications of Thurstonian models to ranking data. In M.

Fligner, J. Verducci (ed.) *Probability Models and Statistical Analysis for Ranking Data*, pp. 157-172. Lecture notes in statistics 80, Springer Verlag, New York.

16. Brown, L.D. (1990). An ancillarity paradox which appears in multiple linear regression. *Ann. Statist.* **18**, 471-538.

17. Christensen, R. (1990). *Log-Linear Models*, Springer-Verlag, New York.

18. Conti, P. and Traverso, C. (1991), Buchberger algorithm and integer programming, Proceedings AAECC-9 (New Orleans), *Springer Lecture Notes in Computer Science*, **539**, pp. 130-139.

19. Cox, D. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357-372.

20. Cox, D., Little, J., and O'Shea, D. (1992). *Ideals, Varieties, and Algorithms*, Springer-Verlag, New York.

21. Croon, M. (1989). Latent class models for the analysis of rankings. In G. De Solte, H. Feger, K.C. Klauer (eds.) *New developments in psychological choice modeling*, pp. 99-121. Elsevier: Holland.

22. Darroch, J., Lauritzen, S. and Speed, T. (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics* **8**, 522-539.

23. Darroch, J. and Speed, T. (1983). Additive and multiplicative models and interactions. *Annals of Statistics* **11**, 724-738.

24. Diaconis, P. (1989). A generalization of spectral analysis with application to ranked data, *Annals of Statistics* **17**, 949-979.

25. Diaconis, P. (1988). *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics. Hayward, CA.

26. Diaconis, P. and Efron, B. (1980). Testing for independence in a two-way table: New interpretations for the chi-square statistic. *Ann. of Statistics* **13**, 845-905.

27. Diaconis, P. and Efron, B. (1987) Probabilistic-geometric theorems arising from the analysis of contingency tables. In A. Gelfand (ed.), *Contributions to the Theory and Application of Statistics. A Volume in Honor of Herbert Solomon.* Academic Press, Inc., New York.

28. Diaconis, P. and Freedman, D. (1987). A dozen deFinetti-style results in search of a theory. *Ann. Inst. Henri Poincaré* **23**, 397-423.

29. Diaconis, P., Graham, R.L., and Sturmfels, B. (1993). Primitive partition identities. Technical Report, Dept. of Mathematics, Harvard University.

30. Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains, *Ann. Appl. Prob.* **1**, 36-61.

31. Diaconis, P. and Sturmfels, B. (1993). Universal Gröbner bases of toric ideals. Technical Report, Dept. of Mathematics, Harvard University.

32. Efron, B. and Hinkley, D. (1978). Assessing the accuracy of the MLE: Observed versus expected Fisher information (with discussion). *Biometrika* **65**, 457-487.

33. Fisher, R. (1925). *Statistical Methods for Research Workers*. 1st ed. (14th ed. 1970). Oliver and Boyd, Edinburgh.

34. Fulton, W. (1993). *Introduction to Toric Varieties*, Princeton University Press.

35. Gangolli, A. (1991). Convergence bounds for Markov chains and applications to sampling. Ph.D. Thesis, Dept. of Computer Science, Stanford University.

36. Goodman, L. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *Jour. Amer. Statis. Assoc.* **65**, 226-256.

37. Glonek, G. (1987). Some Aspects of Log Linear Models. Ph.D. Thesis, School of Mathematical Sciences, Flinders University of South Australia.

38. Good, I.J. (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Annals of Statistics* **4**, 1159-1189.

39. Good, I.J. (1979). Bayes-Billard-Ball argument extended to multinomials. *Jour. Statist. Comput. and Simulation* **9**, 161-163.

40. Good, I.J. (1977). The enumeration of arrays and a generalization related to contingency tables. *Discrete Math.* **19**, 23-45.

41. Haberman, S. (1978). *Analysis of Qualitative Data*. Vols. 1,2, Academic Press, Orlando.

42. Hammersly, I. and Handscomb, D. (1964). Monte Carlo Methods. Wiley, New York.

43. Harris, J. (1992). *Algebraic Geometry*: A first course. Springer, New York.

44. Jerrum, M., Valient, L. and Vazirani, V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Computer Sci.* **43**, 169-188.

45. Jensen, J. (1991). Uniform saddlepoint approximations and log-convex densities.

Jour. Roy. Statist. Soc. B, 157-172.

46. Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *Jour. Amer. Statist. Assoc.* **72**, 789-827.

47. Kolassa, J. and Tanner, M. (1993). Approximate conditional inference in exponential families via the Gibbs sampler. Technical Report, Dept. of Statistics, University of Rochester.

48. Kong, F. (1993). Edgeworth expansions for conditional distributions in logistic regression models. Technical report, Dept. of Statistics, Columbia University.

49. Kong, F. and Levin, B. (1993). Edgeworth expansions for the sum of discrete random vectors and their applications in generalized linear models. Technical report, Dept. of Statististics, Columbia University.

50. Larntz, K. (1978). Small-sample comparison of exact levels for chi-squared goodness-of-fit statistics. *Jour. Amer. Statist. Assoc.* **73**, 253-263.

51. Lauritzen, S. (1993). *Graphical Association Models.* Unpublished book manuscript.

52. Lehmannn, E. (1986). *Testing Statistical Hypotheses*, 2nd ed., Wiley, New York.

53. Levin, B. (1983). On calculations involving the maximum cell frequency. *Comm. Statist.*

54. Levin, B. (1992). Tests of odds ratio homogeneity with improved power in sparse fourfold tables. *Commun. Statist. Th.-method* **21**, 1469-1500.

55. Martin, Löf, P. (1974). Exact tests, confidence regions and estimates. Pp. 121-138 in *Proceedings of the Conference on Foundational Questions in Statistical Inference*, O. Barndorff-Nielsen, P. Blaesild, G. Scholl (eds.), Memoirs No. 1, Dept. of Theoretical Statistics, Institute of Mathematics, Aarhus University.

56. Mayr, E. and Meyer, A. (1982). The complexity of the word problem for commutative semigroups and polynomials ideals. *Adv. Math.* **46**, 305-329.

57. McCullogh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential family models. *International Statistical Review* **53**, 61-67.

58. McCullogh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Jour. Amer. Statist. Assoc.* **81**, 104-107.

59. Mehta, C. and Patel, N. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Jour. Amer. Statist. Assoc.* **78**, 427-434.

60. Mehta, C. and Patel, N. (1992). STATEXACT

61. Odoroff, C. (1970). A comparison of minimum logit chi-square estimation and maximum likelihood estimation in $2 \times 2 \times 2$ and $3 \times 2 \times 2$ contingency tables: tests for interaction. *Jour. Amer. Statist. Assoc.* **65**, 1617-1631.

62. Plackett, R. (1977). The marginal totals of a 2×2 table. *Biometrika* **64**, 37-42.

63. Savage, L. (1976). On Rereading R.A. Fisher. (with discussion) *Annals of Statistics* **4**, 441-500.

64. Sinclair, A. (1993). Algorithms for Random Generation and Counting: a Markov Chain Approach. Birkhäuser, Boston.

65. Skovgaard, I. (1987). Saddlepoint expansions for conditional distributions. *Jour. Appl. Prob.* **24**, 875-887.

66. Snee (1974). Graphical display of two-way contingency tables. *Amer. Statist.* **38**, 9-12.

67. Stanley, R. (1980). Decomposition of rational convex polytopes, *Annals of Discrete Math.* **6**, 333-342.

68. Sturmfels, B. (1991). Gröbner Bases of Toric Varieties. *Tôhoku Math. Journ.* **43**, 249-261.

69. Sturmfels, B. (1992). Asymptotic Analysis of Toric Ideals. *Memoirs Fac. Sci. Kyushu Univ., Ser. A* **46**, 217-228.

70. Thomas, R. (1993). A geometric Buchberger algorithm for integer programming. Technical report, Dept. of Operations Reseach, Cornell University.

71, Weispfenning, V. (1987). Admissible orders and linear forms, *ACM SIGSAM Bulletin* **21**, 16–18.

72. Yarnold, J. (1970). The minimum expectation in X^2 goodness of fit tests and the accuracy of approximations for the null distribution. *Jour. Amer. Statist. Assoc.* **65**, 864-886.

73. Yates, F. (1984). Tests of significance for 2×2 contingency tables. *Jour. Royal Statistic. Soc. A*, **147**, 426-463.