Algebraic Algorithms for Sampling from Conditional Distributions

Persi Diaconis; Bernd Sturmfels

The Annals of Statistics, Vol. 26, No. 1 (Feb., 1998), 363-397.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199802%2926%3A1%3C363%3AAAFSFC%3E2.0.CO%3B2-J

The Annals of Statistics is currently published by Institute of Mathematical Statistics.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ims.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Mon Aug 23 10:33:43 2004

The Annals of Statistics
1998, Vol. 26, No. 1, 363–397

ALGEBRAIC ALGORITHMS FOR SAMPLING FROM CONDITIONAL DISTRIBUTIONS

BY PERSI DIACONIS[1] AND BERND STURMFELS[2]

Cornell University and University of California, Berkeley

We construct Markov chain algorithms for sampling from discrete exponential families conditional on a sufficient statistic. Examples include contingency tables, logistic regression, and spectral analysis of permutation data. The algorithms involve computations in polynomial rings using Gröbner bases.

1. Introduction. This paper describes new algorithms for sampling from the conditional distribution, given a sufficient statistic, for discrete exponential families. Such distributions arise in carrying out versions of Fisher's exact test for independence and goodness of fit. They also arise in constructing uniformly most powerful tests and accurate confidence intervals via Rao–Blackwellization. These and other applications are described in Section 2. As shown below, the new algorithms are a useful supplement to traditional asymptotic theory, which is useful for large data sets, and exact enumeration, which is useful for very small data sets.

The following example should motivate the general construction. Table 1 shows data gathered to test the hypothesis of association between birthday and deathday [Andrews and Herzberg (1985), page 429]. The table records the month of birth and death for 82 descendants of Queen Victoria. A widely stated claim is that birthday–deathday pairs are associated. The usual χ^2 test for independence is 115.6 on 121 degrees of freedom, suggesting no association. The classical rules of thumb for validity of the chi-square approximation (minimum 5 per cell) are badly violated here, and there are too many tables with these margins to permit exact enumeration. Figure 1 shows a probability–probability plot of the permutation distribution of the chi-square statistic versus the chi-square approximation (χ^2_{121}). The approximation is not particularly accurate. Indeed, the permutation probability of $\chi^2 \leq 115.6$ is 0.3208 versus 0.3775 for the approximation.

To illustrate the present approach, 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 $r_1, r_2, \ldots, r_I$ and column sums $c_1, c_2, \ldots, c_J$. Let

Received June 1993; revised April 1997.

[1] Also at Harvard University. Research supported in part by an NSF Grant.

[2] Resarch supported in part by NSF Grant and David and Lucile Packard Fellowship.

AMS 1991 *subject classifications.* 6E17, 13P10.

Key words and phrases. Conditional inference, Monte Carlo Markov chain, exponential families, Gröbner bases.

TABLE 1
Relationships between birthday and deathday

Month of birth	Month of death												
	Jan	**Feb**	**March**	**April**	**May**	**June**	**July**	**Aug**	**Sept**	**Oct**	**Nov**	**Dec**	**Total**
Jan	1	0	0	0	1	2	0	0	1	0	1	0	6
Feb	1	0	0	1	0	0	0	0	0	1	0	2	5
March	1	0	0	0	2	1	0	0	0	0	0	1	5
April	3	0	2	0	0	0	1	0	1	3	1	1	12
May	2	1	1	1	1	1	1	1	1	1	1	0	12
June	2	0	0	0	1	0	0	0	0	0	0	0	3
July	2	0	2	1	0	0	0	0	1	1	1	2	10
Aug	0	0	0	3	0	0	1	0	0	1	0	2	7
Sept	0	0	0	1	1	0	0	0	0	0	1	0	3
Oct	1	1	0	2	0	0	1	0	0	1	1	0	7
Nov	0	1	1	1	2	0	0	2	0	1	1	0	9
Dec	0	1	1	0	0	0	1	0	0	0	0	0	3
Total	13	4	7	10	8	4	5	3	4	9	7	8	82

$\mathscr{F}(\mathbf{r}, \mathbf{c})$ be the set of $I \times J$ arrays (x_{ij}) of nonnegative integers with the given row sums and column sums. Let

$$H = \prod_j \binom{c_j}{x_{1j} \cdots x_{Ij}} \Big/ \binom{N}{r_1 r_2 \cdots r_I}, \qquad N = \sum_{i=1}^{I} c_i = \sum_{j=1}^{J} r_j,$$

be the hypergeometric distribution on $\mathscr{F}(\mathbf{r}, \mathbf{c})$. This is the conditional distribution of the data, given the sufficient statistics (row/column sums) for the classical model of independence.

A Monte Carlo method for generating from H proceeds as follows. Let x be a table which satisfies the constraints. Modify x by choosing a pair of rows and a pair of columns at random. These intersect in four entries and x is modified as

$$\begin{matrix} + & - \\ - & + \end{matrix} \quad \text{or} \quad \begin{matrix} - & + \\ + & - \end{matrix} \quad \text{with probability } \tfrac{1}{2} \text{ each.}$$

The modification adds or subtracts 1 from each of the four entries as indicated. This does not 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 $\mathscr{F}(\mathbf{r}, \mathbf{c})$. By the usual Metropolis procedure (see Lemma 2.1) the chain is modified to give a connected, aperiodic, reversible Markov chain with stationary distribution H.

Figure 2 shows a histogram of the chi-square statistic for Table 1. Figure 1 and the counts reported above were derived from this chain. The 10^6 steps of the Markov chain took about three minutes to run on a p.c. As explained in Section 2, there are more direct methods for sampling from H for two-way tables but for three- and higher way tables the present approach seems to be the only one.

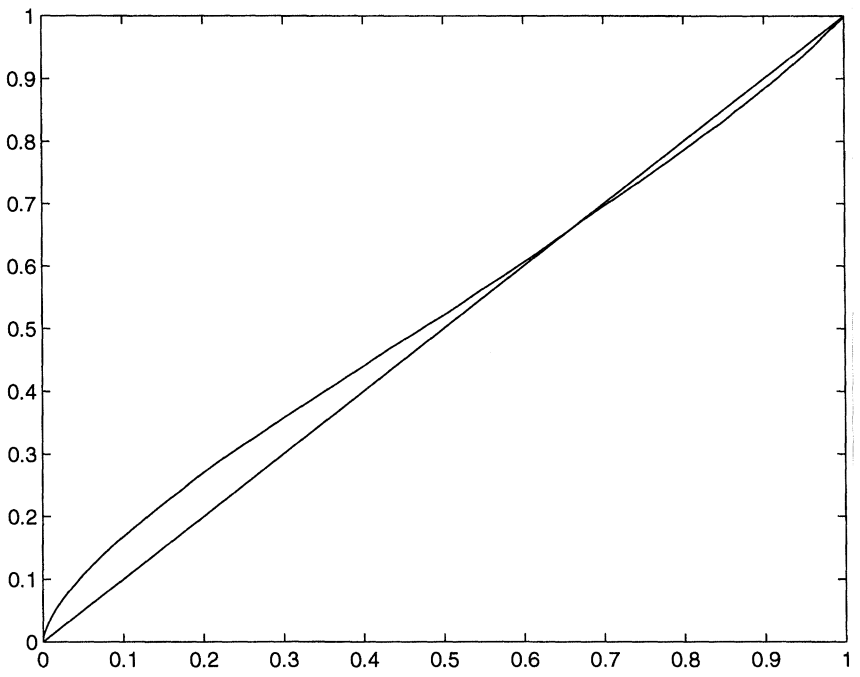


FIG. 1. *Probability–probability plot of the permutation distribution of the chi-square statistic versus chi-square* (121).

More generally, let $\mathscr{X}$ be a finite set. Consider the exponential family

$$P_\theta(x) = Z(\theta)e^{\theta \cdot T(x)}, \qquad \theta \in \mathbb{R}^d, \tag{1.1}$$

$Z(\theta)$ a normalizing constant with $T: \mathscr{X} \to \mathbb{N}^d - \{0\}$ (here $\mathbb{N} = \{0, 1, 2, \ldots\}$). If $X_1, X_2, \ldots, X_N$ are independent and identically distributed from (1.1), the statistic $t = T(X_1) + \cdots + T(X_N)$ is sufficient for θ. Let

$$\mathscr{Y}_t = \{(x_1, \ldots, x_N) \in \mathscr{X}^N : T(x_1) + \cdots + T(x_N) = t\}. \tag{1.2}$$

Under (1.1) the law of $X_1, \ldots, X_N$ given t is uniformly distributed over $\mathscr{Y}_t$. In natural problems it is difficult to enumerate $\mathscr{Y}_t$ effectively or sample from the uniform distribution on $\mathscr{Y}_t$.

It is usual to recast this problem in terms of the hypergeometric distribution as follows. Write

$$t = \sum_{i=1}^{N} T(X_i) = \sum_x G(x)T(x) \quad \text{with } G(x) = \#\{i : X_i = x\}.$$

The counts $G(x)$ form a sufficient statistic for any independent identically distributed data. Define the set of all data sets with the given sufficient

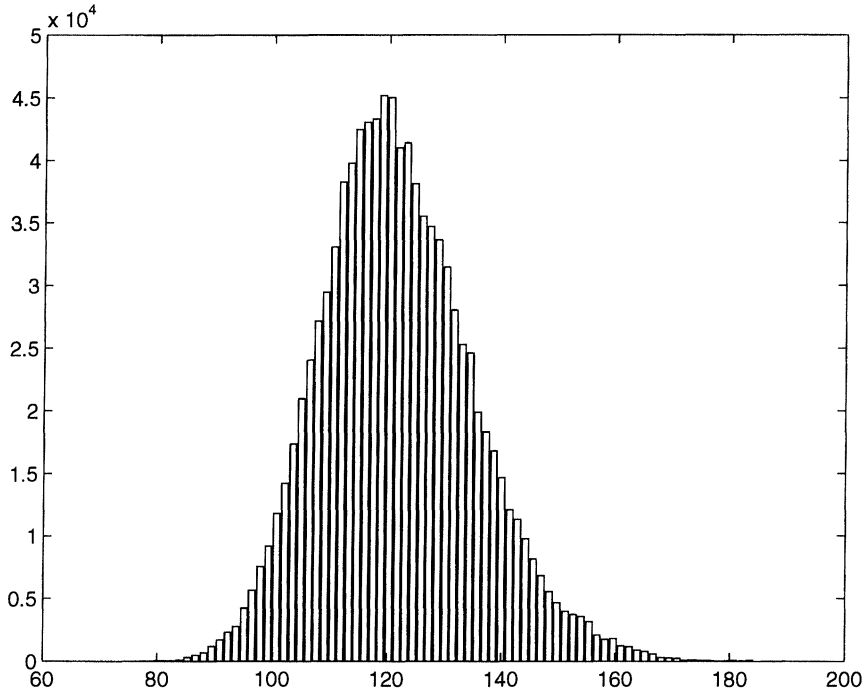


FIG. 2. *Histogram of the chi-square statistic for Table* 1.

statistic as

$$\mathscr{F}_t = \left\{ f: \mathscr{X} \to \mathbb{N}: \sum_x f(x)T(x) = t \right\}. \tag{1.3}$$

Since $T(x)$ is nonzero for all x and one always begins with one data set with sufficient t, $\mathscr{F}_t$ is finite and nonempty. This is assumed throughout. The image of the uniform distribution on $\mathscr{Y}_t$ under the map from $\mathscr{Y}_t$ to $\mathscr{F}_t$ is called the hypergeometric distribution

$$H_t(f) = \frac{N!}{|\mathscr{Y}_t|} \prod_x (f(x)!)^{-1}. \tag{1.4}$$

The problem is thus reduced to sampling from H_t on $\mathscr{F}_t$, given t.

For contingency tables, we have $\mathscr{X} = \{(i,j),\ 1 \le i \le I,\ 1 \le j \le J\}$. The usual model for independence has $T(i,j) \in \mathbb{N}^{I+J}$ a vector of length $I+J$ with two entries equal to one and the rest equal to zero. The ones in $T(i,j)$ are in position i in the first I coordinates and position j in the last J coordinates. The sufficient statistic t contains the row and column sums of the contingency table associated to N observations. The set $\mathscr{F}_t$ is all $I \times J$ tables with these row and column sums. The hypergeometric distribution (1.4) becomes the classical distribution at (1.1).

This paper gives methods for finding the analog of $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ moves for general exponential families.

DEFINITION. *A Markov basis* is a set of functions $f_1, f_2, \ldots, f_L$: $\mathscr{X} \to \mathbb{Z}$ (here $\mathbb{Z} = 0, \pm 1, \pm 2, \ldots$) such that

$$\text{(1.5)(a)} \qquad \sum_x f_i(x)T(x) = 0, \qquad 1 \le i \le L,$$

(1.5)(b) For any t and $f, f' \in \mathscr{T}_t$ there are $(\varepsilon_1, f_{i_1}), \ldots, (\varepsilon_A, f_{i_A})$ with $\varepsilon_i = \pm 1$,

$$f' = f + \sum_{j=1}^{A} \varepsilon_i f_{i_j} \quad \text{and} \quad f + \sum_{j=1}^{a} \varepsilon_j f_{i_j} \ge 0 \quad \text{for } 1 \le a \le A.$$

A Markov basis allows construction of a Markov chain on $\mathscr{T}_t$. From $f \in \mathscr{T}_t$, choose I uniformly in $\{1, 2, \ldots, L\}$ and $\varepsilon = \pm 1$. Form $f + \varepsilon f_I$. If $f + \varepsilon f_I$ is nonnegative, the chain moves there. In other cases the chain stays at f. Condition (1.5)(a) says $f + \varepsilon f_I$ is in $\mathscr{T}_t$. Condition (1.5)(b) says the chain is connected. The chain is modified to have stationary distribution H_t via an extra Metropolis coin-flip. Lemma 2.1 shows this gives an irreducible, aperiodic Markov chain with H_t as stationary distribution.

Section 2.1 lays out the stochastic underpinnings showing a variety of ways that the moves (1.5) can be used. In this paper we have used chi-square tests for goodness-of-fit but the conditional algorithms can be used to calibrate any test statistic. Section 2.2 contains a literature review along with a description of natural statistical problems where conditional calculations are useful. Section 2.3 gives pointers to rates of convergence literature. For example, the chain for tables described above requires (N^2) steps to reach stationarity while some of the speedups in Section 2.1 converge much more rapidly.

Our main contribution is a method for finding and understanding basic moves using tools from computational algebra. Section 3 shows how finding $\{f_1, \ldots, f_L\}$ in (1.5) is equivalent to finding generators for an ideal in a ring of polynomials. This allows us to use the rapidly expanding Gröbner basis technology.

Sections 4, 5 and 6 contain detailed treatments of special cases: contingency tables, logistic regression, and ranked data are treated. These illustrate the application to problems of testing, estimation and confidence intervals. They are more or less self-contained and may be read now for further motivation.

2. Basic stochastics. This section describes the stochastic and statistical background. In Section 2.1 we show how a variety of Markov chains can be constructed using the basic moves (1.5). In Section 2.2 we review the statistical literature on conditional inference and the various approaches that have been used to approximate the conditional distribution. In Section 2.3 we

give an overview of available results on the rate of convergence of the chains to their stationary distribution.

2.1. *Markov chains.* We first show how to set up a Markov chain for a general distribution on $\mathscr{T}_t$.

LEMMA 2.1. *Let $\sigma(g)$ be a positive function on $\mathscr{T}_t$ of* (1.3). *Given functions $f_1, \ldots, f_L$ satisfying* (1.5), *generate a Markov chain on $\mathscr{T}_t$ by choosing I uniformly in $\{1, 2, \ldots, L\}$ and $\varepsilon = \pm 1$ with probability $\frac{1}{2}$ independent of I. If the chain is currently at $g \in \mathscr{T}_t$, it moves to $g + \varepsilon f_I$ (provided this is nonnegative) with probability*

$$\min\left\{\frac{\sigma(g + \varepsilon f_I)}{\sigma(g)}, 1\right\}.$$

In all other cases the chain stays at g. This is a connected, reversible, aperiodic Markov chain on $\mathscr{T}_t$ with stationary distribution proportional to $\sigma(g)$.

PROOF. Call the chain described $K(g, \tilde{g})$. It is easy to check that $\sigma(g)K(g, \tilde{g}) = \sigma(\tilde{g})K(\tilde{g}, g)$. Condition (1.5)(b) shows that the chain is connected. Since there is some holding probability (iterate $g \mapsto g + f_1$ sufficiently often to get a negative coordinate), we are done. □

REMARKS. A useful class of measures on $\mathscr{T}_t$ is specified by choosing a function $\omega_x\colon \mathbb{N} \to \mathbb{R}^+$ for each $x \in \mathscr{X}$. For $g \in \mathscr{T}_t$, define $\sigma(g) = \prod_x \omega_x(g(x))$. For example, 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 generating confidence regions. Taking $\theta_x \equiv 1$ gives they hypergeometric distribution of (1.4). For this class of measures, the ratio $\sigma(\tilde{g})/\sigma(g)$ involves only a few terms in the product if $\tilde{g}$ and g differ in only a few terms. This always seems to happen, and we have found this method effective in the examples of Sections 4–6.

As a nonstandard example, Table 2 gives a 4×4 contingency table [data of Snee (1974)]. The chi-square statistic for this table is $\chi^2 = 138.29$ on

TABLE 2
A 4×4 contingency table

	Hair color				
Eye 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

9 degrees of freedom. Diaconis and Efron (1985) were interested in the distribution of χ^2 under the *uniform* distribution on $\mathscr{T}_t$ [thus $\sigma(g) = 1$]. They labored long and hard to determine the proportion of tables with the same row and column sums as Table 2 having $\chi^2 \leq 138.29$. Their best estimate using a combination of asymptotics and Monte Carlo was "about 10%." Figure 3 shows a histogram from a Monte Carlo run using Lemma 2.1 with $\sigma \equiv 1$. In the run, 18.31% of all tables had $\chi^2 \leq 138.29$.

The algorithm needs no Metropolis step and simply involves the $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ moves described in the Introduction. As an indication of the sizes of the state spaces involved, we note that Des Jardins has shown there are exactly 1,225,914,276,276,768,514 tables with the same row and column sums as Table 2. See Diaconis and Gangolli (1995) for more on this. Holmes and Jones (1995) have introduced a quite different method for uniform generation which gives similar results for this example.

Lemma 2.2 shows how to use the moves (1.5) as directions in $\mathscr{T}_t$ to make longer steps.

LEMMA 2.2. *Give $f_1, \ldots, f_L$ satisfying* (1.5) *on $\mathscr{T}_t$, generate a Markov chain on $\mathscr{T}_t$, by choosing I uniformly in $\{1, 2, \ldots, L\}$. If the chain is currently at $g \in \mathscr{T}_t$, determine the set of $j \in \mathbb{Z}$ such that $g + jf_I \in \mathscr{T}_t$. Choose j in this set*

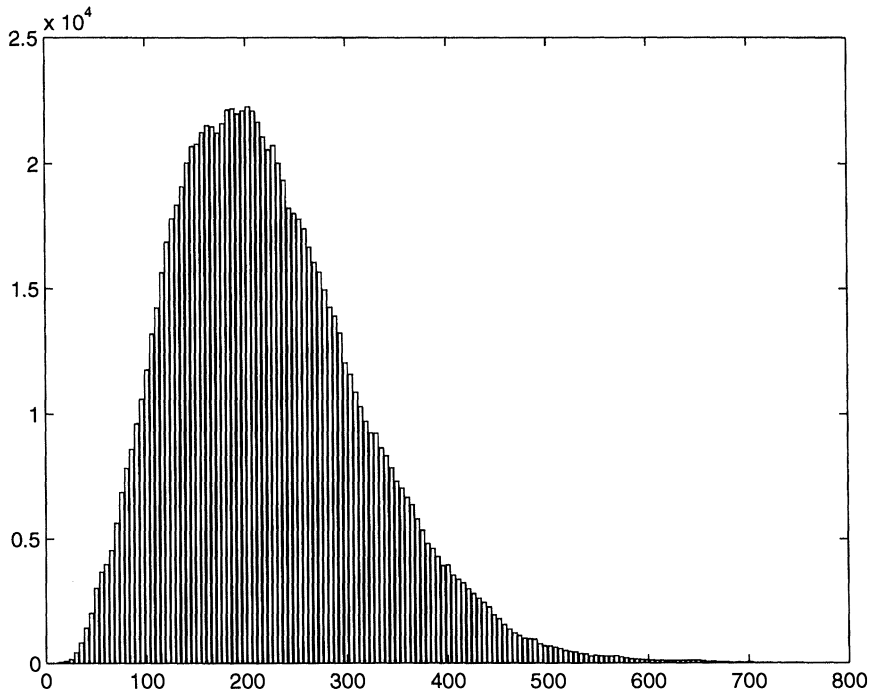


FIG. 3. *Histogram from a Monte Carlo run using Lemma* 2.1 *with* $\sigma \equiv 1$.

with probability proportional to

$$\prod_{x \in C_I} \frac{1}{[g(x) + jf_I(x)]!} \tag{2.1}$$

with $C_I = \{x\colon f_I(x) \neq 0\}$. *This is a connected, reversible, aperiodic Markov chain on* $\mathscr{T}_t$ *with stationary distribution* H_t *of* (1.4).

PROOF. The chain described is a modification of the popular Gibbs sampler or hit and run algorithm. It is easy to see that the product (2.1) is proportional to the stationary distribution constrained to the line $\{g + jf_I\}_{j \in \mathbb{Z}} \cap \mathscr{T}_t$. Hence the chain is reversible with respect to H_t. From (1.5)(b), the chain is connected and again has some holding states and so is aperiodic. This completes the proof. □

REMARKS. For contingency tables, the algorithm of Lemma 2.2 becomes: pick a pair of rows and a pair of columns at random. This delineates a 2×2 subtable. Replace it by a 2×2 table with the same margins, chosen from the hypergeometric distribution. This is easy to do; such a 2×2 table being determined by its (1, 1) entry.

Lemma 2.2 works as well for a general measure σ; just replace (2.1) by $\prod_{x \in C_I} \sigma(g(x) + jf_I(x))$. For example, in a contingency table, if $\sigma \equiv 1$, the 2×2 table is replaced by a uniformly chosen 2×2 table with the same margins.

Sampling from (2.1) can itself be done by running a Markov chain in j. We recommend the directed Metropolis chains of Diaconis, Holmes and Neale (1997).

FINAL REMARKS. (i) Diaconis, Eisenbud and Holmes (1997) have used $\{f_i\}_{i=1}^L$ to run a walk directly on the data space $\mathscr{Y}_t$. This seems useful for sparse problems.

(ii) There is a completely different use of the moves $\{f_i\}_{i=1}^L$ to applied probability problems. For example, consider $\mathscr{T}_t$ as the set of all $I \times I$ tables with all row and column sums equal to m (magic squares). A variety of applied probability questions can be asked. Pick a table in $\mathscr{T}_t$ at random; what is the distribution of the number of 2's (or m's or ...)? Stein's method [see Stein (1986)], is an effective tool to deal with such nonstandard problems. A basic ingredient of Stein's method is an exchangeable pair (X, X') which is marginally uniform on $\mathscr{T}_t$. This is exactly what the Markov chain of Lemma 2.1 provides: choose X uniformly in $\mathscr{T}_t$ and let X' be one step in the chain (with $\sigma \equiv 1$). Holmes (1995) has used this approach to prove that for I large and $m = 2$, the number of twos in a random table is approximately Poisson(1).

(iii) All the algorithms in this paper are for discrete exponential families. The basic ideas can be adapted to more general spaces. For example, in testing goodness-of-fit to a gamma family with unknown location and scale,

one needs to generate from the uniform distribution on $\{(x_1, \ldots, x_N);\ x_i \in \mathbb{R}_+,\ \Sigma x_i = m,\ \Pi x_i = p\}$ for fixed m and p. Given x satisfying the constraints, choose three coordinates at random, change one of them by a small amount [say, uniformly chosen in $(-a, a)$] and then solve for the unique value of the other two coordinates to satisfy the constraints.

(iv) It is worth recording why one "obvious" approach, using a lattice basis to move around on $\mathscr{T}_t$, doesn't work. Given a statistic $T\colon \mathscr{X} \to \mathbb{N}^d$, define a $d \times |\mathscr{X}|$ matrix A with columns $T(x)$. Then $\mathscr{T}_t = \{g\colon \mathscr{X} \to N\colon Ag = t\}$. It is easy to find a basis of integer vectors $V_1, \ldots, V_b$ for $\ker A$ (using, e.g., the Hermite normal form of A [Schrijver (1986), page 45]. Then, for every $g, g' \in \mathscr{T}_t$, $g' = g + \sum_{j=1}^{b} a_j V_j$, for some integers a_j. This suggests a simple Markov chain: from g, choose $A_1, A_2, \ldots, A_b$ independently from some fixed measure on $\mathbb{N}$ [e.g., Poisson(θ)]. Try to move to $g' = g + \sum_{j=1}^{b} \varepsilon_j A_j V_j$ with $\varepsilon_j = \pm 1$ symmetric and independent of A_j. If $g' \in \mathscr{T}_t$ the walk moves there. If not, the walk stays at g.

We have tried this idea in half a dozen problems and found it does not work well. For example, take 10×10 tables with all row and column sums equal to 2. A lattice basis as above can be taken as $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ moves for all sets of four adjacent squares. In repeated runs, the lattice basis walk required millions of steps to converge while the walk described in the Introduction converged after a few hundred steps. Further, finding a choice of the Poisson parameter θ so that the chain moved at all was a remarkably delicate operation.

(v) The ideas above can be used to solve large problems by working on smaller pieces through a procedure we call a fiber walk. Let $\{f_i\}_{i=1}^{L}$ be a Markov basis. Write $f_i^+ = \max\{f_i, 0\}$, $f_i^- = \max\{-f_i, 0\}$ so $f_i = f_i^+ - f_i^-$. Let $\deg f_i = \max\{\Sigma_x f_i^+(x), \Sigma_x f_i^-(x)\}$. Let $D = \max_i\{\deg f_i\}$. For contingency tables, the basic $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ moves have degree 2 so $D = 2$. As will emerge, we can get bounds on the degree without knowing $\{f_i\}$. Let D_* be the minimum degree over all generating sets. Just known an upper bound $D_* \le d_*$ allows a walk to be constructed on $\mathscr{Y}_t$, the big fiber of (1.2). The walk is simple: from $y \in \mathscr{T}_t$, choose d_* coordinates at random. Calculate t_*, the sum of $T(x)$ over the chosen coordinates. Now, choose uniformly at random from the set of d_* tuples with the given value of t_* and replace the d_* tuples chosen with the freshly chosen set. It is easy to see that this walk gives a symmetric, connected, aperiodic Markov chain on $\mathscr{Y}_t$. It follows that the image walk on $\mathscr{T}_t$ has the hypergeometric distribution.

As an example, Section 6 describes some statistical problems involving ranked data. For five ranked items, there are too many variables to easily find a Markov basis. On the other hand, in Section 6.2 we are able to show that $d_* = 5$. We may run a walk by choosing 5-tuples of permutations, computing their 5×5 permutation matrices and choosing a fresh 5-table by direct enumeration. The required calculations are quite feasible.

A crucial ingredient for this version of the algorithm is a bound on D_*. We describe general bounds in Section 3.3 and specific bounds in Sections 4–6.

2.2. *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 closely related literature.

As with so many topics of inferential interest, conditioning was first studied by R. A. Fisher. He systematically used the conditional distribution of the data given a sufficient statistic as a basis for tests of a model in *Statistical Methods for Research Workers* (1925). Even earlier, Fisher, Thornton and Mackenzie (1922) based a test on the fact that if $X_1, \ldots, X_n$ are independent Poisson variates then the distribution of $X_1, \ldots, X_n$ given $X_1 + \cdots + X_n = h$ is like the box counts of h balls dropped at random into n boxes. He returned to this in Fisher (1950), showing how the exact count of partitions gives a useful supplement to the asymptotic chi-square approximation. Fisher 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.

In independent work, Neyman (1937) introduced conditioning as a way of deriving optimal tests and confidence intervals for exponential families. Roughly, to test if one component of a vector of parameters is zero, one uses the conditional distribution of the corresponding component of the sufficient statistic given the rest of the statistic. This has evolved into a unified theory described in Chapters 3 and 4 of Lehmann (1986). Here conditioning is used as a device for getting rid of nuisance parameters. The overall tests are unconditional. An example where the present techniques are used in this way is in Section 5. This decision theoretic use of conditioning has a healthy development [see, e.g., Farrell (1971) or Cohen, Kemperman and Sakrowitz (1994).]

There is far more to the conditional controversy than the above applications. Fortunately, there are good surveys available. Cox (1958, 1988), Kiefer (1977), Efron and Hinkley (1978) and Brown (1990) have been influential papers which have extensive literature reviews. Lehmann (1986), Chapter 10 gives a splendid overview of the inferential issues. Agresti (1992) surveys contingency tables and Reid (1995) surveys the approximation problem.

On the computational side, there has become a growing awareness that the usual asymptotic approximation of mathematical statistics can be poor for moderate sample sizes. Clear examples in a contingency table setting are 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 two-way tables are explicitly described in Aldous (1987), Gangolli (1991) and Glonek (1987). Apparently Darroch suggested the idea in the late 1970's. 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 at the first r_1 places; the number of entries between $c_1 + \cdots + c_{j-1} + 1$ and $c_1 + \cdots + c_j$ is n_{1j}, $1 \le j \le J$. The number of such entries in the next r_2 places is r_{2j}, and so

on. We have compared the output of this exact Monte Carlo procedure with the random walk procedure for a variety of tables and found they produce virtually identical results for two-way tables.

Closely related is a combinatorial method for carrying out an exact test for Hardy–Weinberg equilibrium. Guo and Thompson (1992) give a random walk approach which can be seen as a special case of the general algorithm. See Section 4.3. Lange and Lazzeroni (1997) give a different random walk, which comes with a guarantee. Besag and Clifford (1989) discuss a similar method for testing the Rasch model with binary matrices.

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 STATXACT, 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., $\tilde{\theta}$ the maximum likelihood estimator), Edgeworth or saddle point approximations to the probability $P_{\tilde{\theta}}(t)$ and $P_{\tilde{\theta}}(x, t)$ 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 Kong and Levin (1993) and 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 (1994, 1996) are a recent contribution in this direction.

2.3. *Rates of convergence.* The Markov chains described in Section 2.1 require some running time to reach their stationary distribution. There has been active work in computing sharp rates of convergence for such discrete chains. Roughly, for a variety of chains, theory shows that order γ^2 steps are necessary and suffice for convergence in total variation. Here γ is the diameter of the underlying graph, which has as vertices the points of the state space ($\mathscr{T}_t$ in our examples) and an edge from f to $\hat{f}$ if $\hat{f}$ can be reached in one step from f.

The theory has been most carefully worked out for contingency tables with uniform stationary distribution and steps $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ or $\begin{smallmatrix} - & + \\ + & - \end{smallmatrix}$ as described in the Introduction. Here is a typical result.

THEOREM [Diaconis and Saloff-Coste (1995a)]. *Fix I, J and positive integer $\mathbf{r} = (r_1, \ldots, r_I)$, $\mathbf{c} = (c_1, \ldots, c_J)$ with $\Sigma c_j = \Sigma r_i = N$. Let $\mathscr{T}(\mathbf{r}, \mathbf{c})$ be the set*

of all $I \times J$ tables with row/column sums $\mathbf{r}, \mathbf{c}$. Let U be the uniform distribution on $\mathscr{F}(\mathbf{r}, \mathbf{c})$. Let $K(x, y)$ be the $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ walk described in the Introduction. Then,

$$\|K_x^k - U\|_{TV} \leq A_1 e^{-A_2 c} \quad \text{for } k = c\gamma^2,\ c > 0.$$

Here $x \in \mathscr{F}(\mathbf{r}, \mathbf{c})$ is any starting state and A_1, A_2 are explicit constants which depend on I, J but not otherwise on $\mathbf{r}, \mathbf{c}$. Further γ, the diameter of the graph, satisfies $\gamma \leq N/2$.

Conversely, there is $x \in \mathscr{F}(r, c)$ and constants A_3, A_4 as above such that

$$\|K_x^k - U\|_{TV} \geq A_3 e^{-A_4 c} \quad \text{for } k = c\gamma^2.$$

The theorem shows that order γ^2 steps are necessary and sufficient to achieve stationarity. The constants A_i grow exponentially in $D = I \cdot J$ being roughly $(D/4)^{D/4}$. For small tables (e.g., 4×4) this gives reasonable rates. For example, for Table 2, it suggests 100,000 $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ steps are necessary and suffice for convergence.

For large size tables, there is an alternative result due to Chung, Graham and Yau (1996) which has similar conclusions—order (diam)^2 steps necessary and sufficient—but constants which do not depend badly on dimension. Their result does require restrictions on the row/column sums being sufficiently large. A similar result, proved by different methods, is due to Dyer, Kannan and Mount (1995). Presumably the technical difficulties blocking a unified result will soon be overcome. For the present, order (diameter)^2 with diameter $\leq N/2$ is a useful heuristic with quite a bit of technical back-up.

The analysis above is all for the *local* algorithm based on $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ moves. The algorithm described in Lemma 2.2 obviously gets random much more rapidly, at least for non sparse tables, such as Table 2. (For sparse tables it is not possible to move very far.) It is one of the challenging open problems of the theory to prove this. The algorithms of Lemma 2.2 are very similar to the continuous hit and run algorithms of Belisle, Romeijn and Smith (1993). There, Doeblin's condition gives reasonable results. It should be possible to modify the proofs there for the discrete case.

The discussion above has all been for tables. There is much to be done in adapting the available machinery, such as the Poincaré, Nash and log Sobolev inequalities used by Diaconis and Saloff-Coste (1995b, 1996a, b), to handle more general problems. Virag (1997) is a useful contribution to this program.

The approaches above use eigenvalues to bound the rate of convergence. There is every hope of using coupling as in Hernek (1997) or stopping times as in Propp and Wilson (1996) to get useful bounds.

3. Some algebra. In this section we show how to compute a Markov basis using tools from computational algebra. This is not familiar in statistical work but we can assure the reader that all we need is long division of polynomials. The first two chapters of the marvelous undergraduate book by Cox, Little and O'Shea (1992) is more than enough background. In Section 3.1

we show how finding a Markov basis is equivalent to finding a set of generators of an ideal in a polynomial ring. In Section 3.2 we show how to represent this ideal in a way suitable for computation in *MATHEMATICA* or *MAPLE*.

3.1. *Markov bases and ideals.* Throughout, $\mathscr{X}$ is a finite set and T: $\mathscr{X} \to \mathbb{N}^d - \{0\}$ is given. For each $x \in \mathscr{X}$ introduce an indeterminate also denoted x. Consider the ring of polynomials $k[\mathscr{X}]$ in these indeterminates where k is any field (e.g., the real field $\mathbb{R}$ on $\mathbb{F}_2$, the field of two elements). A function g: $\mathscr{X} \to \mathbb{N}$ will be represented as a monomial $\prod_{x \in \mathscr{X}} x^{g(x)}$. This monomial is denoted $\mathscr{X}^g$. The function T: $\mathscr{X} \to \mathbb{N}^d$ is represented by the homomorphism

$$\varphi_T\colon k[\mathscr{X}] \to k[t_1, \ldots, t_d]$$
$$x \mapsto t_1^{T(x)_1} t_2^{T(x)_2} \cdots t_d^{T(x)_d}.$$

Here $T(x)_i$ denotes the ith coordinate of $T(x) \in \mathbb{N}^d$ and the map φ_T is defined on products and sums by multiplicativity and linearity [so $\varphi_T(x^2) = (\varphi_T(x))^2$, $\varphi_T(x + y) = \varphi_T(x) + \varphi_T(y)$, etc.]. Our basic object of study is $\mathscr{I}_T = \{p \in k[\mathscr{X}]\colon \varphi_T(p) = 0\}$, the kernel of φ_T.

In Theorem 3.1 we will show that a set of generators for $\mathscr{I}_T$ (that is, a set of polynomials in $\mathscr{I}_T$ that generate $\mathscr{I}_T$ as an ideal in $k[\mathscr{X}]$) corresponds to a Markov basis. To state this correspondence we need the following notation. Any function f: $\mathscr{X} \to \mathbb{Z}$ can be written as the difference between two functions f^+ and f^-, $\mathscr{X} \to \mathbb{N}$ having disjoint support: $f^+(x) = \max(f(x), 0)$, $f^-(x) = \max(-f(x), 0)$. Observe that $\sum f(x)T(x) = 0$ if and only if the monomial difference $\mathscr{X}^{f^+} - \mathscr{X}^{f^-}$ is in $\mathscr{I}_T$. A basic result is the following theorem.

THEOREM 3.1. *A collection of functions $f_1, f_2, \ldots, f_L$ is a Markov basis* (1.5) *if and only if the set*

$$\mathscr{X}^{f_i^+} - \mathscr{X}^{f_i^-}, \qquad 1 \le i \le L$$

generates the ideal $\mathscr{I}_T$.

PROOF. The proof proceeds in two stages. Let $\mathscr{I}'$ be the ideal generated by the monomial differences

$$\mathscr{X}^{f^+} - \mathscr{X}^{f^-}; \qquad \sum f(x)T(x) = 0. \tag{3.1}$$

We first show that $\mathscr{I}' = \mathscr{I}_T$. It is clear that $\mathscr{I}' \subseteq \mathscr{I}_T$ (since each generator is in $\mathscr{I}_T$). To prove the converse, fix a total order of the set of all monomials by linearly ordering the variables and declaring one monomial 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 $\mathscr{I}' \subsetneq \mathscr{I}_T$. Let $p \in \mathscr{I}_T - \mathscr{I}'$ have its largest monomial $\mathscr{X}^\alpha$ a minimum. Since $\varphi_T(p) = 0$, there must be a second monomial $\mathscr{X}^\beta$ in p such that $\varphi_T(\mathscr{X}^\beta) = \varphi_T(x^\alpha)$. Factor out common variables writing $\mathscr{X}^\alpha - \mathscr{X}^\beta = \mathscr{X}^\gamma(\mathscr{X}^{\alpha'} - \mathscr{X}^{\beta'})$ with α' and β' having disjoint support. Clearly $\varphi_T(\mathscr{X}^{\alpha'}) =$

$\varphi_T(\mathscr{X}^{\beta'})$. Setting $h(x) = \alpha'(x) - \beta'(x)$, we have $\Sigma h(x)T(x) = 0$ and so $\mathscr{X}^{\alpha} - \mathscr{X}^{\beta} = \mathscr{X}^{\gamma}(\mathscr{X}^{h^+} - \mathscr{X}^{h^-}) \in \mathscr{I}'$. Subtracting a multiple of $\mathscr{X}^{\alpha} - \mathscr{X}^{\beta}$ from p, we get a polynomial in $\mathscr{I}_T - \mathscr{I}'$ with a smaller leading monomial. This provides $\mathscr{I}_T = \mathscr{I}'$.

To prove the theorem, Let $\mathscr{B} = \{\mathscr{X}^{f_i^+} - \mathscr{X}^{f_i^-}; 1 \le i \le L\}$. Property (1.5)(a) is equivalent to $\mathscr{B} \subset \mathscr{I}_T$. Thus it must be shown that (1.5)(b) holds if and only if $\mathscr{B}$ generates $\mathscr{I}_T$. Assume (1.5)(b) holds. By what was proved above it is enough to show that for any $f: \mathscr{X} \to \mathbb{Z}$ with $\Sigma_x f(x)T(x) = 0$, the monomial difference $\mathscr{X}^{f^+} - \mathscr{X}^{f^-}$ is in the ideal generated by $\mathscr{B}$. Use (1.5)(b) with $g = f^+$ and $g' = f^-$. We have

$$g + \sum_{j=1}^{A} \varepsilon_j f_{i_j} = g' \quad \text{with } g + \sum_{j=1}^{a} \varepsilon_j f_{i_j} \ge 0,\ 1 \le a \le A.$$

If $A = 1$ and, say $\varepsilon_1 = 1$, 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 $\mathscr{X}^{f^+} - \mathscr{X}^{f^-} = -(\mathscr{X}^{f_{i_1}^+} - \mathscr{X}^{f_{i_1}^-}) \in \mathscr{I}_T$. A similar argument works if $A = 1$ and $\varepsilon_1 = -1$. In the general case $A > 1$. By induction on A, the monomial differences $\mathscr{X}^{f^+} - \mathscr{X}^{f^+ + \varepsilon_1 f_{i_1}}$ and $\mathscr{X}^{f^- + \varepsilon_1 f_{i_1}^+} - \mathscr{X}^{f^-}$ lie in the ideal generated by $\mathscr{B}$. So does their sum.

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

$$\mathscr{X}^{g} - \mathscr{X}^{g'} = \sum_{j=1}^{L} \varepsilon_j \mathscr{X}^{h_j}(\mathscr{X}^{f_{i_j}^+} - \mathscr{X}^{f_{i_j}^-}), \qquad \varepsilon_j \in \{\pm 1\}.$$

Here h_j: $\mathscr{X} \to \mathbb{N}$ and the polynomial on the right has coefficients plus or minus 1 since the proof that $\mathscr{I}' = \mathscr{I}_T$ above works over $\mathbb{Z}$ so any integer polynomial with $\varphi_T(p) = 0$ can be written as an integer polynomial combination of $\mathscr{X}^{f^+} - \mathscr{X}^{f^-}$ with $\Sigma f(x)T(x) = 0$. If $A = 1$, the identity above translates directly into (1.5)(b). For $A > 1$, proceed by induction. From the identity, $\mathscr{X}^{g} = \mathscr{X}^{h_r}\mathscr{X}^{f_{i_r}^{\pm}}$ for some r, say, $\mathscr{X}^{g} = \mathscr{X}^{h_r}\mathscr{X}^{f_{i_r}^-}$. Then $g - f_{i_r}^-$ is nonnegative and so $g + f_{i_r}$ is nonnegative. Subtracting $\mathscr{X}^{h_r}(\mathscr{X}^{f_{i_r}^+} - \mathscr{X}^{f_{i_r}^-})$ from both sides and using $h_r + f_{i_r}^+ = g + f_{i_r}$, we get an expression for $\mathscr{X}^{g+f_{i_r}} - \mathscr{X}^{g'}$ having length $A - 1$. By induction, $g + f_{i_r}$ can be connected to g' by allowable steps so (1.5)(b) holds for all g, g'. □

REMARKS. (i) The Hilbert basis theorem says that any ideal in a polynomial ring has a finite generating set. Applying this to the ideal $\mathscr{I}_T$ we see that Markov bases exist for any statistic T. We show how to compute such a basis explicitly in Section 3.2.

(ii) We have chosen to work with T taking values in $\mathbb{N}^d$. Essentially the same arguments work if $T(x) \in \mathbb{Z}^d$; just map from $k[\mathscr{X}]$ into $k[t_1, \ldots, t_d, t_1^{-1}, \ldots, t_d^{-1}]$.

3.2. *Algorithms for computing a Markov basis with examples.* Theorem 3.1 reduces the problem to computing a generating set for the ideal $\mathscr{I}_T \subseteq k[\mathscr{X}]$.

We show how to give a finite description of $\mathcal{I}_T$ which can then be read into computer algebra systems such as *AXIOM*, *MAPLE*, *MACSYMA*, *MATHEMATICA*. All the examples in this paper were computed using the program *MACAULAY* of Bayer and Stillman (1989). An updated version *MACAULAYII* due to Grayson and Stillman is fast and avaialble at no cost (http://math.uiuc.edu/Macaulay2).

A crucial ingredient here is an ordering on monomials. We use grevlex order; consider $\alpha, \beta \in \mathbb{N}^n$. Declare $\mathcal{X}^\alpha \prec \mathcal{X}^\beta$ if either $\sum \beta_i > \sum \alpha_i$ or $\sum \beta_i = \sum \alpha_i$ and the first nonvanshing difference, working from the right, has $\beta_i - \alpha_i < 0$. Thus $(1, 0, 1) \prec (0, 2, 0)$ in grevlex order. Of course, implicit in this order is an ordering on the basic variables $x \in \mathcal{X}$. This will be made explicit in examples below. See Cox, Little and O'Shea (1992), Chapter 2, for background. This ordering allows us to define the initial term init(p) of a polynomial.

Let $\mathcal{I}$ be an ideal in $k[\mathcal{X}]$ with an ordering on monomials as above. A *Gröbner basis for* $\mathcal{I}$ is a set of polynomials $\{p_1, p_2, \ldots, p_L\} \subset \mathcal{I}$ such that the ideal generated by $\{\text{init}(p_1), \ldots, \text{init}(p_L)\}$ equals the ideal generated by $\{\text{init}(p);\ p \in \mathcal{I}\}$. A Gröbner basis generates $\mathcal{I}$ and there is a computationally feasible algorithm for finding Gröbner bases in the computer systems above. A Gröbner basis is *minimal* if no polynomial can be deleted. It is *reduced* if for a each pair i, j, no term of p_i is divisible by init p_j. Fixing a term order, there is a unique reduced Gröbner basis. The following algorithm is an easy-to-implement way of finding this basis.

THEOREM 3.2. *Let $\mathcal{X}$ be a finite set. Let T: $\mathcal{X} \to \mathbb{N}^d$ be given. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_d\}$. Given an ordering for $\mathcal{X}$, extend it to an elimination ordering for $\mathcal{X} \cup \mathcal{T}$ with $t \succ x$ for all $x \in \mathcal{X}$, $t \in \mathcal{T}$ in $k[\mathcal{X}, \mathcal{T}]$. Define $\mathcal{I}_T = \{x - \mathcal{T}^{T(x)}$, $x \in \mathcal{X}\}$. Then $\mathcal{I}_T = \mathcal{T}_T \cap k[\mathcal{X}]$ and the reduced Gröbner basis for $\mathcal{I}_T$ can be found by computing a reduced Gröbner basis for $\mathcal{I}_T$ and taking those output polynomials which only involve $\mathcal{X}$.*

The proof is a straightforward application of the elimination theorem from Cox, Little, and O'Shea (1992), pages 114, 128. The method is a special case of the implicitization algorithm.

EXAMPLE 3.3. Consider finding a basis for the case of 3×3 contingency tables. Using the computer system Maple, the following 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*z3,  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 $y1, y2, y3, z1, z2, z3$, we are left with nine basic moves of the type $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$.

REMARK 3.4. For $I \times J$ contingency tables, $k[\mathscr{X}]$ is the ring of polynomial functions on a generic $I \times J$ matrix. The ideal $\mathscr{I}_T$ is the ideal generated by the 2×2 minors. Using row major order on the variables $x_{11} > x_{12} > \cdots > x_{1J} > x_{21} > \cdots > x_{IJ}$, the algorithm of Theorem 3.2 produces the $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ moves of the Introduction. See Sturmfels (1991), page 260. These determinantal ideals have been the object of intense study by algebraists and geometers. Sturmfels (1996) gives further discussion and references.

Two other sets of moves are worth mentioning for this example. Let K_{IJ} be the complete bipartite graph on I and J nodes. So K_{23} appears as shown in Figure 4. Any cycle in K_{IJ} 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. These moves, algebraically interpreted, are a Gröbner basis for *any* ordering of the variables. These universal Gröbner bases are discussed in Sturmfels (1996), Chapter 7.

Here is an interesting statistical application. Contingency tables sometimes have forced zero entries: 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, Fienberg and Holland (1975) or Haberman (1978) Chapter 7 for discussion and example. To do a random walk on tables with restricted positions, just delete the edges of K_{IJ} corresponding to the restrictions and use the cycles in the remaining graph. An amusing consequence of the connectedness of this algorithm is that if there are no circuits, the remaining table is uniquely determined by its margins. The use of universal Gröbner bases to handle forced zeros extends to the general set-up.

A second set of moves consists of using only the $(1, 1)$ entry coupled with the $(i, j), (1, i), (j, 1)$ entries, $2 \le i \le I$, $2 \le j \le J$. These moves fail to connect for all tables but Gloneck (1987) shows they connect if all the row and column sums are at least 2. Extensions and variants of Gloneck's result using the primary decomposition of one ideal in a second are in Diaconis, Eisenbud and Sturmfels (1996). Curiously, these same nonconnecting moves are used by Kolassa and Tanner (1994), who failed to worry about connectedness. It is not

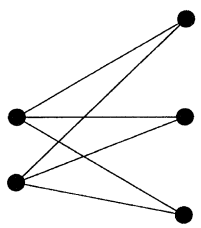


FIG. 4.

serious for $I \times J$ tables, but they make the same error for three- and higher tables where things are much more serious. Their error consists in assuming that a lattice basis connects staying positive. This is simply false.

REMARK. Theorem 3.2 is a useful way of finding reduced Gröbner bases. It may need many new variables. There are several alternatives. See Sturmfels (1996), Section 12.A for details.

4. Contingency tables. Two-way tables have been used as a running example in previous sections. In this section we treat three- and higher way tables. Apparently the random walks presented here are the only way of generating from the hypergeometric distribution for most cases. Section 4.1 treats an example of "no three-way interaction." Section 4.2 briefly discusses hierarchial, graphical, and decomposable models. Section 4.3 treats Hardy–Weinberg equilibrium. There is a vast modern literature on contingency tables. Argesti (1990), Bishop, Fienberg and Holland (1975), Christensen (1990) and Haberman (1978) give surveys of the literature.

4.1. *No three-factor interactions.* 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}. The "no three-factor interaction" model specifies constant log odds:

$$\frac{p_{111}p_{ij1}}{p_{i11}p_{1j1}} = \frac{p_{11k}p_{ijk}}{p_{i1k}p_{1jk}} \qquad 2 \le i \le I, 2 \le j \le J, 2 \le k \le K. \tag{4.1}$$

Sufficient statistics for this model are all "line sums." If the table entries are N_{ijk}, the line sums are $N_{.jk}, N_{i\cdot k}, N_{ij\cdot}$, where, for example, $N_{.jk} = \Sigma_i N_{ijk}$. Tests for this model are described by Birch (1963) or Bishop, Fienberg and Holland (1975). We first treat an example and then return to the general case.

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 $3 \times 3 \times 3$ array shown as Table 3 below.

The first variable is type of Christian (Northern Protestant, Southern Protestant, Catholic). The second variable is education: low (less than 9

TABLE 3
Attitudes toward abortions among white Christian subjects

	Northern Protestant			**Southern Protestant**			**Catholic**		
	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

years), medium (9 through 12 years), high (more than 12 years). The third variable is attitude to nontherapeutic abortion (positive, mixed, negative). The data are treated as a simple random sample of size 1,055 from the U.S. population in 1972.

The maximum likelihood estimates of the cell entries under the model (4.1) are found by iterative proportional fitting to be

12.01	14.43	39.58	9.44	12.25	40.27	6.55	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 13.37. 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 in Lemma 2.1 to get a hypergeometric sample with the same line sums. The walk was based on 110 moves described below. After 50,000 burn-in steps, the walk was run for 100,000 steps sampling every 50 steps for a total of 2,000 values.

We conclude that the algorithm works easily and well, that the chi-square approximation seems good (there is a small systematic bias upward in Figure 2), and that the no three-way interaction model fits these data. Haberman

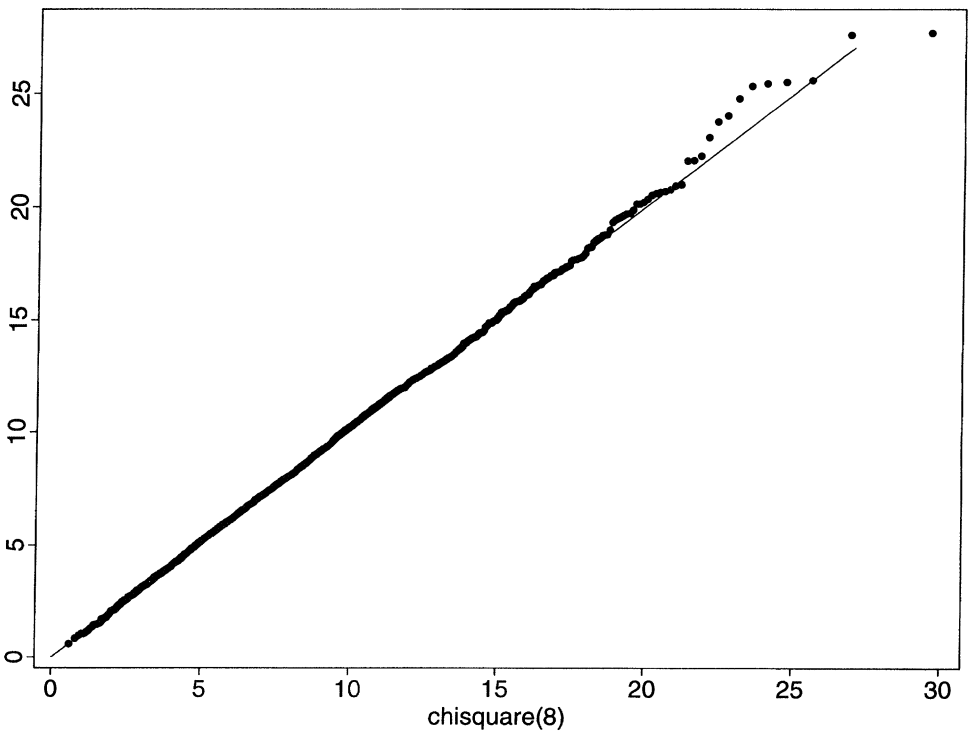


FIG. 5. *A p − p plot of random walk values of chi-square versus chi-square* (8).

(1978), Section 4.2 presents further analysis with data from subsequent years.

We turn next to the moves needed to perform a random walk on a $3 \times 3 \times 3$ table with fixed line sums. It is natural to consider basic $2 \times 2 \times 2$ moves such as

$$\begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & + & - & 0 & - & + \\ 0 & 0 & 0 & 0 & - & + & 0 & + & -. \end{array} \tag{4.2}$$

There are 27 such moves; alas, the chain they generate is not connected. Using the program *MACAULAY*, we ran the basic algorithm of Theorem 3.2. This involved computations in a polynomial ring with 54 variables (27 variables x_{ijk} for the table entries and 27 variables $y^1_{ij}, y^2_{ik}, y^3_{jk}$ for the line sums). We found that a minimal set of generators consists of the 27 moves as in (4.2) and 54 moves of degree 6 like

$$\begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 0 & - & + & 0 & + & - \\ 0 & 0 & 0 & + & 0 & - & - & 0 & + \\ 0 & 0 & 0 & - & + & 0 & + & - & 0. \end{array} \tag{4.3}$$

The pattern in the last two layers can be permuted in six ways and the two layers placed in nine ways. This gives 54 moves.

In carrying out the computation, the cells (i, j, k) were ordered lexicographically and grevlex was used for a term order on the monomnials in the x_{ijk}. The reduced Gröbner basis for this order contains 110 basic moves: the 27 + 54 minimal generators plus

$$\text{(4.4)}\quad \text{28 relations of degree 7 like} \qquad \begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & + & 0 & - & - & 0 & + \\ 0 & - & + & - & + & 0 & + & 0 & - \\ 0 & + & - & 0 & - & + & 0 & 0 & 0 \end{array}$$

$$\text{(4.5)}\quad \text{1 relation of degree 9} \qquad \begin{array}{ccc|ccc|ccc} -2 & + & + & + & 0 & - & + & - & 0 \\ + & 0 & - & 0 & 0 & 0 & - & 0 & + \\ + & - & 0 & - & 0 & + & 0 & + & - \end{array}$$

We conclude by reporting what we know for larger tables with fixed line sums $N_{\cdot jk}, N_{i\cdot k}, N_{ij\cdot}$. There is a neat description of the moves for $2 \times J \times K$ tables. For a $2 \times n \times n$ table, consider the move

$$\begin{array}{cccccc|cccccc} + & - & 0 & 0 & \cdots & 0 & - & + & 0 & 0 & \cdots & 0 \\ 0 & + & - & 0 & \cdots & 0 & 0 & - & + & 0 & \cdots & 0 \\ 0 & 0 & + & - & \cdots & 0 & 0 & 0 & - & + & \cdots & 0 \\ \vdots & & & & & & \vdots & & & & & \\ 0 & 0 & \cdots & 0 & + & - & 0 & 0 & \cdots & 0 & - & + \\ - & & \cdots & 0 & 0 & + & + & 0 & \cdots & & 0 & - \end{array} \tag{4.6}$$

The product of the symmetric groups $S_n \times S_n$ acts on the rows and columns. This gives $(n-1)!n!/2$ distinct permutations of (4.6). Call these *basic moves of degree* $2n$. For $n \le J \le K$, any of these basic moves can be placed in a $2 \times J \times K$ array. There are $\binom{J}{n}\binom{K}{n}$ distinct ways to do this. Al-

together this gives $\sum_{n=2}^{J}((n-1)!n!/2)\binom{J}{n}\binom{K}{n}$ moves. We have shown that these moves form a minimal generating set which is at the same time a universal Gröbner basis.

Call the ideal associated to the fixed line sum problem $\mathcal{I}(I, J, K)$. A binomial in $\mathcal{I}(I, J, K)$ is *critical* if it cannot be written as a polynomial linear combination of binomials of lower degree. Thus the moves corresponding to critical binomials are needed to get a connected walk. The *type* of a critical binomial is the size of the smallest three-way table which supports it. We give two nontrivial examples of critical binomials. These show that basic moves for $2 \times J \times K$ tables do not generate $\mathcal{I}(I, J, K)$ for large I, J, K.

A critical relation of type $4 \times 4 \times 6$ is

$$\begin{aligned} &x_{131}x_{241}x_{142}x_{322}x_{123}x_{433}x_{214}x_{344}x_{235}x_{415}x_{316}x_{426} \\ &\quad - x_{141}x_{231}x_{122}x_{342}x_{133}x_{423}x_{244}x_{314}x_{215}x_{435}x_{416}x_{326}. \end{aligned} \tag{4.7}$$

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

$$\begin{aligned} &x_{111}x_{361}x_{132}x_{342}x_{153}x_{323}x_{124}x_{214}x_{225}x_{335} \\ &\quad \times x_{356}x_{266}x_{147}x_{257}x_{318}x_{248}x_{169}x_{239} \\ &\quad - x_{161}x_{311}x_{142}x_{332}x_{123}x_{353}x_{114}x_{224}x_{325}x_{235} \\ &\quad \times x_{256}x_{366}x_{157}x_{247}x_{218}x_{348}x_{139}x_{269}. \end{aligned} \tag{4.8}$$

We briefly explain the derivation of (4.7) and (4.8). First note that we get zero after deleting the third subscript. This amounts to a nontrivial 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, Las Vergnas, Sturmfels, White and Ziegler (1993), Section 8.5]. They encode projective incidence theorems or nonrealizability 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)].

The following result is shown in Sturmfels (1996), 14.14.

PROPOSITION 4.1. *Given any triple of integers $K \geq J \geq I \geq 2$ there exists a critical relation of type $K' \times J' \times I'$ for some integers $K' \geq K$, $J' \geq J$, $I' \geq I$.*

None of this says that it is impossible to find some "nice" set of generators for $\mathcal{I}(I, J, K)$; it only says that the simple moves we found so far do not suffice. Of course, in any specific case, one can always ask the computer to find moves.

As a final topic, we give the best bounds we have on the degree of binomials needed to generate $\mathcal{I}(I, J, K)$. Let T_{IJK} represent the linear map which takes a three-way table to the set of line sums. It is easy to see that the kernel of T_{IJK} has rank $(I-1)(J-1)(K-1)$. Therefore, the rank of T_{IJK} equals $r = IJ + IK + JK - I - J - K - 1$. Let $D(I, J, K)$ denote the largest absolute value of any $r \times r$ minor of the matrix T_{IJK}.

4.2. *For* $3 \le I \le J \le K$, *we have the following.*

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

(b) D *satifies* $\min(I, J, K) - 1 \le D(I, J, K) \le 3^{r/2}$.

PROOF. Part (a) is proved in Sturmfels (1991). To prove the upper bound in (b), note that $D(I, J, K)$ is the determinant of an $r \times r$ matrix which has at most three ones and otherwise zeros in each column. Hadamard's inequality now gives the result. For the lower bound in (b), use the fact that $D(I, J, K)$ is an upper bound for the degree of any variable in a circuit of $\mathcal{A}(I, J, K)$; a circuit of $\mathcal{A}(I, J, K)$ is a binomial with minimal support [Sturmfels (1996)]. The following binomial is a circuit for the $I \times I \times I$ table:

$$\begin{aligned} &\prod_{i=1}^{I} x_{iii} \prod_{j=2}^{I-1} (x_{j11} \cdot x_{jjj+1}) \prod_{k=2}^{I} (x_{I1k} x_{Ik1}) \\ &\quad = x_{I11}^{I-1} x_{1I1} \prod_{i=1}^{I-1} x_{1Ii+1} \prod_{j=2}^{I-1} (x_{j1j+1} x_{jj1}) \prod_{k=2}^{I} x_{Ikk}. \end{aligned} \tag{4.9}$$

The variable x_{I11} appears with degree $I - 1$ in the circuit (4.9). So we are done.

□

4.2. *Log-linear models.* These are models for multiway contingency tables. The index set is $\mathcal{X} = \prod_{\gamma \in \Gamma} I_\gamma$ with Γ indexing the various categories and I_γ the set of values in a category. Let $p(x)$ be the probability of falling into cell $x \in \mathcal{X}$. A log-linear model can be specified by writing

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

The sum ranges over subsets $a \subseteq T$ and the notation $\varphi_a(x)$ means the function φ_a only depends on x through coordinates in a. Thus $\varphi_\emptyset$ is a constant and φ_r is a completely general function. Specifying $\varphi_a \equiv 0$ for some class of sets determines a model.

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 $\Gamma = \{a, b, c\}$, the class $\mathcal{C} = \{(a, b), (a, c), (b, c)\}$ defines the no three-way interaction model of Section 4.1.

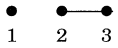
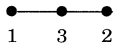
The sufficient statistics for a hierarchical model are $\{N(i_c)\}$ with c ranging over $\mathcal{C}$, $i_c \in \prod_{\gamma \in c} I_\gamma$ and $N(i_c)$ the sum over all x that agree with i_c in the coordinates determined by c. This falls into the class of problems covered by the basic set-up of this paper.

Hierarchical models have unique maximal likelihood estimates which can be effectively computed using Newton–Raphson or the iterated proportional

fitting method. This leads to estimates $\hat{p}_{\mathscr{E}}(x)$. If $\mathscr{E} \subset \mathscr{D}$ are two generating classes, an exact test for adequacy of model $\mathscr{E}$ within $\mathscr{D}$ may be based on the conditional distribution (under $\mathscr{E}$) of the chi-square statistic.

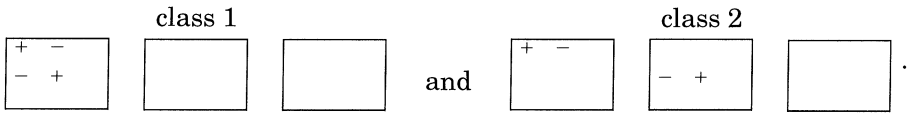
$$\sum_x \frac{\left(N\hat{p}_{\mathscr{E}}(x) - N\hat{p}_{\mathscr{D}}(x)\right)^2}{N\hat{p}_{\mathscr{E}}}.$$

Graphical models [see Lauritzen (1996)] are a subclass of hierarchical models obtained from a graph with vertex set Γ and edge set E. The generating class $\mathscr{E}$ is the cliques of the graph (maximal complete subgraphs). These models are 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. For example, on three points, the following models are graphical:

	complete independence	one-variable independent	conditional independence
			
model p_{ijk}	$p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k}$	$p_{i\cdot\cdot}p_{\cdot jk}$	$p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k}$
sufficient statistics	$N_{i\cdot\cdot}, N_{\cdot j\cdot}, N_{\cdot\cdot k}$	$N_{i\cdot\cdot}, N_{\cdot jk}$	$N_{i\cdot k}, N_{\cdot jk}$

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*. These arise from graphs for which any cycle of length 4 contains a chord. Decomposable models allow closed form maximum likelihood estimates and simple algorithms for generating from the hypergeometric distribution. The three models pictured above are decomposable. We briefly describe the moves for a random walk for these models.

Complete independence. There are two classes of moves which are depicted as



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

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

These generate an irreducible Markov chain. The ring map φ_T of Section 2 sends x_{ijk} to $u_i v_j w_k$. The associated ideal $\mathscr{I}_T$ 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).

One-variable independent. There are three choices possible. For definiteness, say that the variable i is independent of (j, k). The sufficient statistics are $N_{i..}$ and $N_{.jk}$, $1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$. An easy-to-implement Markov chain identifies the pairs (j, k) with a new variable l, $1 \le l \le L = JK$. Now consider the table as an I by L array and use the two-dimensional $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ moves of the Introduction.

Conditional independence. Again there are three choices. For definiteness, say variables i and j are conditionally independent given k. Then the sufficient statistics are $N_{i.k}$ and $N_{.jk}$. Here, for each fixed value of k, one has a two-dimensional face with k fixed. The walk proceeds as k independent walks in each of these k tables.

4.3. *Hardy–Weinberg equilibrium.* In common genetics problems N ordered pairs with values in $\{(i, j), 1 \le i \le j \le n\}$ are observed. These give rise to counts N_{ij}: the number of times (i, j) appears. The Hardy–Weinberg model assumes there are parameters p_i, $1 \le i \le n$, $p_1 + \cdots + p_n = 1$ such that the chance of the pair (i, j) is $2p_i p_j$ if $i \ne j$ and p_i^2 if $i = j$. This model can be derived as the equilibrium distribution for a large population of alleles with no mutation or migration. The chance of observing $\{N_{ij}\}$ is proportional to

$$\prod_{i=1}^{n} p^{N_{i\cdot}}, \qquad N_{i\cdot} = N_{ii} + \sum_{j=1}^{n} N_{in}.$$

A test of the model can be based on the conditional distribution of N_{ij} given $N_{1\cdot}, N_{2\cdot}, \ldots, N_{n\cdot}$.

Guo and Thompson (1992) describe the background and give examples to show that asymptotic approximations can perform poorly for sparse tables. They develop a Monte Carlo approach using moves akin to the $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ moves of the Introduction. We show below that their moves arise from a well-known Gröbner basis for the ideal generated by the 2×2 minors of a symmetric matrix. We further show how to generalize these to larger subsets [e.g., (i, j, k)] with restrictions on the number of types.

Let $\tilde{\mathscr{X}} = \{\mathbf{i} = (i_1, i_2, \ldots, i_r), 1 \le i_1 \le \cdots \le i_r \le n\}$. Let $w_i(\mathbf{i})$ be the number of times i appears in $\mathbf{i}$. Fix nonnegative integers $s_1, s_2, \ldots, s_n$. Let

$$\mathscr{X} = \{\mathbf{i}: w_i(\mathbf{i}) \le s_i, 1 \le i \le n\}.$$

In the original Hardy–Weinberg example, $r = s_1 = s_2 = 2$. Taking $s_1 = s_2 = 1$ amounts to excluding $\{i, i\}$ observations. Larger r examples arise in observations on haploid populations having, for example, four sets of genes.

Data with values in $\mathscr{X}$ give rise to counts $\{N_{\mathbf{i}}\}_{\mathbf{i} \in \mathscr{X}}$. In analogy with random mating, suppose that the chance of observing $\mathbf{i}$ is $\binom{r}{i_1 \cdots i_r} \prod_1^n p_i^{w_i(\mathbf{i})}$. Now, the sufficient statistics are

$$N_i + \sum_i w_i(\mathbf{i}) N_{\mathbf{i}}.$$

The following algorithm gives a connected symmetric Markov chain on the $\{N_{\mathbf{i}}\}_{\mathbf{i}\in\mathscr{X}}$ with fixed values of N_i.

ALGORITHM. *Fix* $r, s_1, s_2, \ldots, s_n$. *For* $\mathbf{i} = (i_1, \ldots, i_r)$ *with* $w_i(\mathbf{i}) \le s_i$, $1 \le i \le n$, *let* $N_{\mathbf{i}}$ *be nonnegative integers with* $N_i = \sum_{\mathbf{i}} w_i(\mathbf{i}) N_{\mathbf{i}}$.

(i) *Choose indices* $\mathbf{i}, \mathbf{i}' \in \mathscr{X}$ *at random.*

(ii) *Form* $\mathbf{j}, \mathbf{j}'$ *from* $\mathbf{i}, \mathbf{i}'$ *by transposing randomly chosen elements of* $\mathbf{i}, \mathbf{i}'$ (*and sorting if needed*). *If* $\mathbf{j}, \mathbf{j}' \in \mathscr{X}$, *go to* (3); *else go to* (1).

(iii) *Choose* $\varepsilon = \pm 1$ *with probability* $\frac{1}{2}$. *Form new counts*

$$N_{\mathbf{i}} - \varepsilon, N_{\mathbf{i}'} - \varepsilon, N_{\mathbf{j}} + \varepsilon, N_{\mathbf{j}'} + \varepsilon.$$

If these are all nonnegative the chain moves to the new counts. If not, the chain stays at the old counts.

PROPOSITION 4.3. *The algorithm gives a symmetric, connected, aperiodic Markov chain on the set of nonnegative* $\{N_{\mathbf{i}}\}_{\mathbf{i}\in\mathscr{X}}$ *with fixed values of* N_i.

PROOF. Theorem 14.2 of Sturmfels (1996) considers the toric ideal in variables $\{x_{\mathbf{i}}\}_{i\in\mathscr{X}}$ generated by binomials

$$\mathscr{I} = \langle x_u x_v \cdots x_w - x_{u'} x_{v'} \cdots x_{w'} : \operatorname{sort}(u, v, \ldots, w) = \operatorname{sort}(u', v', \ldots, w') \rangle$$

with "sort" denoting the sorting operator for strings over the alphabet $\{1, 2, \ldots, n\}$. The theorem shows that there is a term order in $k[\mathscr{X}]$ such that a Gröbner basis for the ideal $\mathscr{I}$ is

$$\{x_{u_1 \cdots u_r} x_{v_1 \cdots v_r} - x_{w_1 w_3 \cdots w_{2r-1}} x_{w_2 w_4 \cdots w_{2r}} : w_1 w_2 w_2 \cdots w_{2r} = \operatorname{sort}(u_1 v_1 u_2 v_2 \cdots u_r v_r)\}.$$

The moves of the algorithm are a translation of these generators. Now, Theorem 3.1 proves the assertion. □

REMARKS. (i) For the Hardy–Weinberg case $r = s_1 = s_2 = 2$, the algorithm reduces to the moves of Guo and Thompson (1992). In this case, there is also available a straightforward method for sampling from the exact conditional distribution which would be the method of routine choice. Lange and Lazzeroni (1997) have found a different Monte Carlo Markov chain which seems to perform faster than the straightforward algorithm and comes with a guaranteed stopping time to say how long it should run. In all cases, the chain above would usually be modified to have a hypergeometric distribution using the Metropolis algorithm as in Lemma 2.2.

(ii) The algorithm (4.2) can equivalently be used to sample randomly from the set of vector partitions of a fixed integer r with parts bounded by $s_1, \ldots, s_n$:

$$\mathscr{X} = \{(x_1, x_2, \ldots, x_n) \in \mathbb{N}^n : x_1 + \cdots + x = r, 0 \le x_1 \le s_1, \ldots, 0 \le r_n \le s_n\}.$$

By using the bijection mapping $(x_1, x_2, \ldots, x_n)$ into the weakly increasing string $(\overbrace{1\ 1\ \cdots\ 1}^{x_1}\overbrace{2\ 2\ \cdots\ 2}^{x_2}\overbrace{n\ n\ \cdots\ n}^{x_n})$.

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.

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 $\mathscr{A}$ of $\mathbb{Z}^d$. A logistic model specifies a log-linear relation of form

$$P(Y = 1|z) = \frac{e^{z\cdot\beta}}{1 + e^{z\cdot\beta}}, \qquad P(Y = 0|z) = \frac{1}{1 + e^{z\cdot\beta}},$$

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 $n_1(z)$ be the number of $i \in \{1, \ldots, N\}$ with $z_i = z$ and $Y_i = 1$. The collection $\{n(z)\}_{z\in\mathscr{A}}$ and the vector sum $\sum_{z\in A} n_1(z)z$ 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 $\mathscr{X} = \{(0, z), (1, z), z \in \mathscr{A}\}$, and let $T: \mathscr{X} \to \mathbb{Z}^{d+|\mathscr{A}|}$ be defined by

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

where there is a single 1 in the last $|\mathscr{A}|$ coordinates at the zth position. Then for given data $f: \mathscr{X} \to \mathbb{N}$, the sum $t = \sum_{x\in\mathscr{X}} f(x)T(x)$ fixes the sufficient statistics. This general problem can now be solved using the techniques of Section 3. The ideals arising are called of Lawrence type in Sturmfels (1996), Chapter 7, which contains further discussion.

EXAMPLE. Haberman (1978), Chapter 7, gives data from the 1974 social science survey on men's response to the statement "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, $0 \le i \le 12$. The data are given in Table 4. Here $n_1(i)$ is the number of "approving" and $n(i)$ is the total number in the sample with i years of education. Also shown are $p(i) = n_1(i)/n(i)$, the proportion approving. These proportions seem to decrease with years of education. It is natural to fit a logistic model of form

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

TABLE 4
Men's response to "Women should run their homes and leave men to run the country" (1974/75)[1]

i	0	1	2	3	4	5	6	7	8	9	10	11	12
$n_1(i)$	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
$p(i)$	0.66	1	1	0.66	0.5	0.65	0.74	0.64	0.60	0.50	0.42	0.38	0.32

[1] With years of education i

This falls into the framework above with $d = 2$, $\mathscr{A} = \{(1, 0), (1, 1), \ldots, (1, 12)\}$. The sufficient statistics to be preserved are

$$\{n(i)\}_{i=0}^{12}, \qquad \sum_{i=0}^{12} n_1(i), \qquad \sum_{i=1}^{12} n_1(i)i. \tag{5.3}$$

A randomization test with these statistics fixed would be appropriate in testing the linear logistic model (5.2) against the nonparametric alternative where $P(Y = 1|i)$ is allowed to take arbitrary values.

For the data of Table 4, the maximum likelihood estimates of α and β in the model (5.2) are $\hat{\alpha} = 2.0545$, $\hat{\beta} = -0.2305$. The chi-squared statistic for goodness-of-fit is $\sum_{i=1}^{12}(n(i)\hat{p}(i) - n_1(i))^2/n(i)\hat{p}(i) = 11.951$. The classical asymptotics calibrate this value with the chi-square (11) 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 16, 968 basis elements. The walk was run, titled to the hypergeometric distribution as in Lemma 2.1. Following 50,000 burn-in steps, a chi-square value was computed every 50 steps for the next 100,000 steps. 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-squared (11) distribution as shown in Figure 6.

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 4.

REMARKS. (i) The random walk was used above as a goodness-of-fit test. In Diaconis and Rabinowitz (1997) it is used to set confidence intervals and compute UMVU estimates. Briefly, the walk can be used to set confidence intervals for γ in the model

$$P\{Y = 1|i\} = \exp(\alpha + i\beta + i^2\gamma)/(1 + \exp(\alpha + i\beta + i^2\gamma)).$$

by using the distribution of $\sum i^2 n_1(i)$ under the walk.

The UMVU estimate of $P\{Y = 1|i\}$ is

$$E\left\{\frac{n_1(i)}{n(i)}\middle| n(j), 1 \le j \le n, \sum_j j n_1(j)\right\}.$$

The expectation can be carried out using the walk.

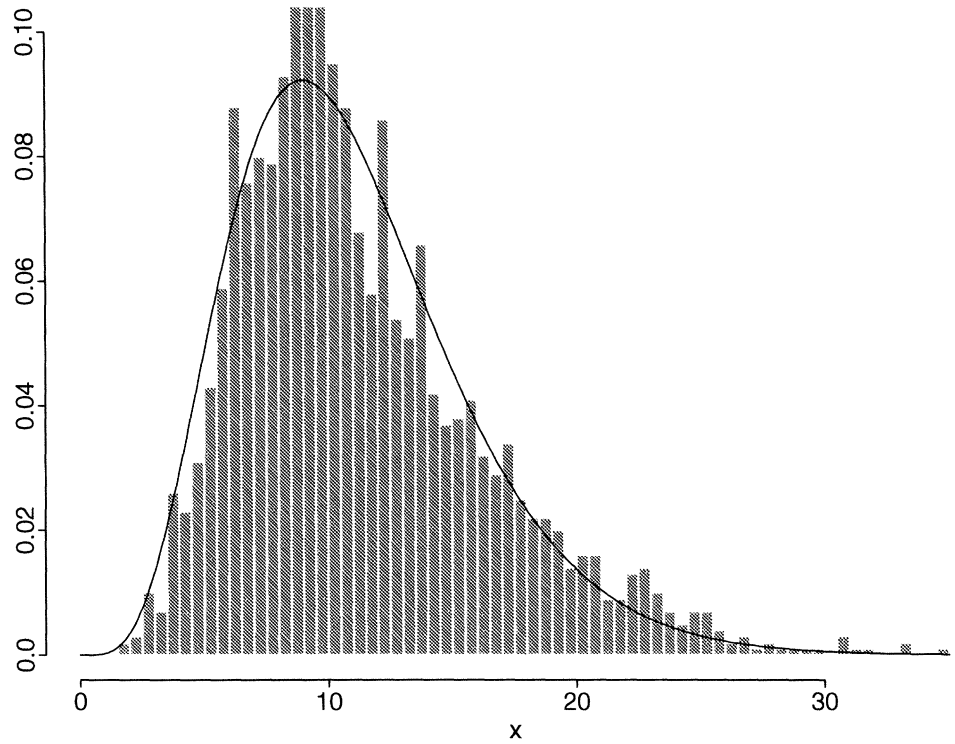


FIG. 6. *Histogram of random walk values of* χ^2 *versus a chi-square* (11).

(ii) A detailed algebraic study of the class of ideals arising from logistic regression is carried out in Diaconis, Graham and Sturmfels (1996). We give a combinatorial description of the basic moves and show that each minimal generating set is automatically a universal Gröbner basis. It is also shown that for the model (5.2) with $1 \leq i \leq n$, the maximum degree of a move is $n - 1$.

(iii) Very similar developments can be made for vector-valued covariates and outcome variables taking more than two values. All these problems fit into the general class of Section 1.

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 6.1. In Section 6.2 we prove that appropriate Markov chains can be found with Gröbner bases having small degree. This uses a result of Stanley (1980) and also the connection between Gröbner bases and triangulations of the convex polytope $\text{conv}\{T(x)\colon x \in \mathscr{X}\}$ developed in [Sturmfels (1991)].

6.1. *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 postion j. This is only a partial summary, since $n!$ numbers are compressed into n^2 numbers. A sequence of further summaries was described in Diaconis (1989). These arise from a decomposition

$$L(S_n) = V_0 \oplus V_1 \oplus V_2 \oplus \cdots \oplus V_k.$$

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 $V_0 \oplus V_1$. 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 to 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	39	3421	34	4321	27	
	875		279		914		194	2262

Thus 137 people ranked (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			
position	875	279	914	194
	746	433	742	341
	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	657		462		381		268		48		4

Here V_0 is the one-dimensional space of constant functions. V_1 is a nine-dimensional space of "first-order functions" spanned by $\pi \mapsto \delta_{i\pi(j)}$ 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 $\pi \mapsto \delta_{\{i,i'\},\{\pi(j),\pi(j')\}}$ and orthogonal to $V_0 \oplus V_1$. The space V_3 contains "ordered second order functions" and V_4 is a one-dimensional space recording the differences between even and odd permutations. Further details are in Diaconis (1988, 1989) or Marden (1995).

Below each subspace is shown the length of the projection of the original data f. The first two subspaces V_0 and V_1 pick up much of the total length. 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 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 $\mathscr{X} = S_4$, 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: $\mathscr{X} \to \mathbb{N}$, then the 4×4 matrix

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

is the first-order summary reported above. We identify f with the monomial $\prod_\pi x_\pi^{f(\pi)}$ in the variables $x_\pi = [\pi_1 \pi_2 \pi_2 \pi_4]$, $\pi \in \mathscr{X}$. The permutation group was ordered using lex order ($1234 > 1243 > \cdots > 4321$). Then grevlex order was used on monomials in $k[\mathscr{X}]$. 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 quadratic 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 numbre of Gröbner basis elements (1,050 relations of degree 2 and 56,860 of degree 3). Remark (v) of Section 2.1 shows how to use the degree bound developed below to carry out a walk on larger permutation groups.

6.2. *Toric ideals for permutation data.* We write x_π for the indeterminate asociated with $\pi \in \mathscr{X} = S_n$ and $t_{i,j}$ for the indeterminate associated with the entries in the permutation matrix. The ring homomorphism φ_T of Section 3 here becomes

$$\begin{aligned} \varphi\colon k[\mathscr{X}] &\to k\big[t_{ij},\ 1 \le i, j \le n\big], \\ x_\pi &\mapsto \prod_{i=1}^{n} t_{i,\pi(i)}. \end{aligned} \tag{6.1}$$

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

THEOREM 6.1. *Let* $\succ$ *be any of the* $(n!)!$ *graded reverse lexicographic term orders on* $k[\mathscr{X}]$. *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 lexicograhic 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 convex polytopes 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 $(\mathscr{I})$ is generated by square-free monomials.

Stanely (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 removing 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 simplices are identified with square-free initial ideals. This completes the proof of (6.2).

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

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

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

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

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

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

REMARKS. (i) 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.

(ii) Stanley's result, used to prove (6.2), has the following direct combinatorial interpretation: let t be any $n \times n$ matrix with nonnegative 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 $\{\mathscr{X}^{f_i} - \mathscr{X}^{g_i}\}$.

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 second-order effects determined by the projection onto V_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 $\mathbb{Z}_2^d$ used for panel studies or item analysis); see Diaconis (1988), Chapter 7.

Acknowledgments. 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, David Eisenbud, Anil Gangolli, Ron Graham, Susan Holmes, Serkan Hosten, David Johnson, Steffan Lauritzen, Bruce Levin, Jun Liu, Brad Mann, Asya Rabinowitz, Charles Stein, Mike Stillman, Rekha Thomas, Thomas Yan, Alan Zaslowzki and Günter Ziegler for their help.

REFERENCES

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

AGRESTI, A. (1992). A survey of exact inference for contingency tables. *Statist. Sci.* **7** 131–177.

ALDOUS, D. (1987). On the Markov chain stimulation method for uniform combinatorial distributions and simulated annealing. *Probab. Engrg. Infom. Sci.* **1** 33–46.

ANDREWS, D. and HERZBERG, A. (1985). *Data*. Springer, New York.

BAGLIVIO, J., OLIVIER, D. and PAGANO, M. (1988). Methods for the analysis of contingency tables with large and small cell counts. *J. Amer. Statist. Assoc.* **83** 1006–1013.

BAGLIVIO, J., OLIVIER, D. and PAGANO, M. (1992). Methods for exact goodness-of-fit tests. *J. Amer. Statist. Assoc.* **87** 464–469.

BAGLIVIO, J., OLIVIER, D. and PAGANO, M. (1993). Analysis of discrete data: rerandomization methods and complexity. Technical report, Dept. Mathematics, Boston College.

BAYER, D. and STILLMAN, M. (1989). *MACAULAY*: a computer algebra system for algebraic geometry. Available via anonymous ftp from zariski.harvard.edu.

BELISLE, C., ROMEIJN, H. and SMITH, R. (1993). Hit and run algorithms for generating multivariate distributions. *Math. Oper. Res.* **18** 255–266.

BESAG, J. and CLIFFORD, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76** 633–642.

BIRCH, B. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. Ser. B* **25** 220–233.

BISHOP, Y., FINEBERG, S. and HOLLAND, P. (1975). *Discrete Multivariate Analysis*. MIT Press.

BJÖRNER, A., LAS VERGNAS, M., STURMFELS, B., WHITE, N. and ZIEGLER, G. (1993). *Oriented Matroids*. Cambridge Univ. Press.

BOKOWSKI, J. and RICHTER-GEBERT, J. (1990). On the finding of final polynomials. *European J. Combin.* **11** 21–34.

BOKOWSKI, J. and RICHTER-GEBERT, J. (1991). On the classification of non-realizable oriented matroids. II. Preprint, T. H. Darmstadt.

BÖKENHOLT, U. (1993). Applications of Thurstonian models to ranking data. *Probability Models and Statistical Analysis for Ranking Data. Lecture Notes in Statist.* **80** 157–172. Springer, New York.

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

CHRISTENSEN, R. (1990). *Log-Linear Models*. Springer, New York.

CHUNG, F., GRAHAM, R. and YAU, S. T. (1996). On sampling with Markov chains. *Random Structures Algorithms* **9** 55–77.

COHEN, A., KEMPERMAN, J. and SACKROWITZ, H. (1994). Unbiased testing in exponential family regression. *Ann. Statist.* **22** 1931–1946.

CONTI, P. and TRAVERSO, C. (1991). Buchberger algorithm and integer programming. *Proceedings AAECC-9. Lecture Notes in Comp. Sci.* **539** 130–139. Springer, New York.

COX, D. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372.

COX, D. (1988). Some aspects of conditional and asymptotic inference. *Sankhyā Ser. A* **50** 314–337.

COX, D., LITTLE, J. and O'SHEA, D. (1992). *Ideals, Varieties, and Algorithms*. Springer, New York.

CROON, M. (1989). Latent class models for the analysis of rankings. In *New Developments in Psychological Choice Modeling* (G. De Solte, H. Feger and K. C. Klauer, eds.) 99–121. North-Holland, Amsterdam.

DARROCH, J., LAURITZEN, S. and SPEED T. (1980). Markov fields a log-linear interaction models for contingency tables. *Ann. Statist.* **8** 522–539.

DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. IMS, Hayward, CA.

DIACONIS, P. (1989). A generalization of spectral analysis with application to ranked data. *Ann. Statist.* **17** 949–979.

DIACONIS, P. and EFRON, B. (1985). Testing for independence in a two-way table: new interpretations for the chi-square statistic. *Ann. Statist.* **13** 845–905.

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

DIACONIS, P., EISENBUD, D. and HOLMES, S. (1997). Speeding up algebraic random walks. Dept. Mathematics, Brandeis Univ. Preprint.

DIACONIS, P., EISENBUD, D. and STURMFELS, B. (1996). Lattice walks and primary decompositions. In *Proceedings of the Rota Fest* (B. Sagan, ed.). To appear.

DIACONIS, P. and FREEDMAN, D. (1987). A dozen deFinetti-style results in search of a theory. *Ann. Inst. H. Poincaré* **23** 397–423.

DIACONIS, P. and GANGOLLI, A. (1995). Rectangular arrays with fixed margins. In *Discrete Probability and Algorithms* (D. Aldous, et al., eds.). 15–41. Springer, New York.

DIACONIS, P., GRAHAM, R. and STURMFELS, B. (1996). Primitive partition identities. *Combinatorics. Paul Erdős Is Eighty* **2** 173–192.

DIACONIS, P., HOLMES, S. and NEALE, R. (1997). A nonreversible Markov chain sampling method. Technical report, Biometry, Cornell Univ.

DIACONIS, P. and RABINOWITZ, A. (1997). Conditional inference for logistic regression. Technical report, Stanford Univ.

DIACONIS, P. and SALOFF-COSTE, L. (1995a). Random walk on contingency tables with fixed row and column sums. Dept. Mathematics, Harvard Univ., Preprint.

DIACONIS, P. and SALOFF-COSTE, L. (1995b). What do we know about the Metropolis algorithm. Technical report, Dept. Mathematics, Harvard Univ.

DIACONIS, P. and SALOFF-COSTE, L. (1996a). Nash inequalities for finite Markov chains. *J. Theoret. Probab.* **9** 459–510.

DIACONIS, P. and SALOFF-COSTE, L. (1996b). Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.* **6** 695–750.

DIACONIS, P. and STROOCK, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.* **1** 36–61.

DYER, R., KANNAN, R. and MOUNT, J. (1995). Sampling contingency tables. *Random Structures Algorithms*. To appear.

EFRON, B. and HINKLEY, D. (1978). Assessing the accuracy of the MLE: observed versus expected Fisher information (with discussion). *Biometrika* **65** 457–487.

FARRELL, R. (1971). The necessity that a conditional procedure be almost everywhere admissible. *Z. Wahrsch. Verw. Gebiete* **19** 57–66.

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

FISHER, R. (1950). The significance of deviations from expectation in a Poisson series. *Biometrics* **6** 17–24.

FISHER, R., THORNTON, H. and MACKENZIE, N. (1922). The accuracy of the plating method of estimating the density of bacterial populations. *Ann. Appl. Biology* **9** 325–359.

FULTON, W. (1993). *Introduction to Toric Varieties*. Princeton Univ. Press.

GANGOLLI, A. (1991). Convergence bounds for Markov chains and applications to sampling. Ph.D. thesis, Dept. Computer Science, Stanford Univ.

GLONEK, G. (1987). Some aspects of log linear models. Ph.D. thesis, School of Math. Sciences, Flinders Univ. South Australia.

GOODMAN, L. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *J. Amer. Statist. Assoc.* **65** 226–256.

GUO, S. and THOMPSON, E. (1992). Performing the exact test for Hardy–Weinberg proportion for multiple alleles. *Biometrics* **48** 361–372.

HABERMAN, S. (1978). *Analysis of Qualitative Data* **1**, **2**. Academic Press, Orlando, FL.

HAMMERSLY, J. and HANDSCOMB, D. (1964). *Monte Carlo Methods*. Wiley, New York.

HARRIS, J. (1992). *Algebraic Geometry: A First Course*. Springer, New York.

HERNEK, D. (1997). Random generation and counting of rectangular arrays with fixed margins. Dept. Mathematics, Preprint, UCLA.

HOLMES, R. and JONES, L. (1996). On uniform generation of two-way tables with fixed margins and the conditional volume test of Diaconis and Efron. *Ann. Statist.* **24** 64–68.

HOLMES, S. (1995). Examples for Stein's method. Preprint, Dept. Statistics, Stanford Univ.

JENSEN, J. (1991). Uniform saddlepoint approximations and log-convex densities. *J. Roy. Statist. Soc. Ser. B* 157–172.

KIEFER, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72** 789–827.

KOLASSA, J. and TANNER, M. (1994). Approximate conditional inference in exponential families via the Gibbs sample. *J. Amer. Statist. Assoc.* **89** 697–702.

KOLASSA, J. and TANNER, M. (1996). Approximate Monte Carlo conditional inference. Dept. Statistics, Northwestern Univ. Preprint.

KONG, F. (1993). Edgeworth expansions for conditional distributions in logistic regression models. Technical report, Dept. Statistics, Columbia Univ.

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. Statistics, Columbia Univ.

LANGE, K. and LAZZERONI, L. (1997). Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. *Ann. Statist.* To appear.

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

LAURITZEN, S. (1996). *Graphical Models*. Oxford Univ. Press.

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

LEVIN, B. (1992). On calculations involving the maximum cell frequency. *Comm. Statist.*

LEVIN, B. (1992). Tests of odds ratio homogeneity with improved power in sparse fourfold tables. *Comm. Statist. Theory Methods* **21** 1469–1500.

MARDEN, J. (1995). *Analyzing and Modeling Rank Data*. Chapman and Hall, London.

MAYR, E. and MEYER, A. (1982). The complexity of the word problem for commutative semigroups and polynomial ideals. *Adv. in Math.* **46** 305–329.

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

MCCULLOUGH, P. (1986). The conditional distribution of goodness-to-fit statistics for discrete data. *J. Amer. Statist. Assoc.* **81** 104–107.

MEHTA, C. and PATEL, N. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Amer. Statist. Assoc.* **78** 427–434.

NEYMAN, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans.* **236** 333–380.

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. *J. Amer. Statist. Assoc.* **65** 1617–1631.

PROPP, J. and WILSON, D. (1986). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9** 232–252.

REID, N. (1995). The roles of conditioning in inference. *Statist. Sci.* **10** 138–199.

SAVAGE, L. (1976). On rereading R. A. Fisher (with discussion). *Ann. Statist.* **4** 441–450.

SCHRIJVER, A. (1986). *Theory of Linear and Integer Programming*. Wiley, New York.

SINCLAIR, A. (1993). *Algorithms for Random Generation and Counting: A Markov Chain Approach*. Birkhäuser, Boston.

SKOVGAARD, I. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Probab.* **24** 875–887.

SNEE (1974). Graphical display of two-way contingency tables. *Amer. Statist.* **38** 9–12.

STANLEY, R. (1980). Decompositoin of rational convex polytopes. *Ann. Discrete Math.* **6** 333–342.

STEIN, C. (1986). *Approximate Computation of Expectations*. IMS, Hayward, CA.

STURMFELS, B. (1991). Gröbner bases of toric varieties. *Tôhoko Math. J.* **43** 249–261.

STURMFELS, B. (1992). Asymptotic analysis of toric ideals. *Mem. Fac. Sci. Kyushu Univ. Ser. A* **46** 217–228.

STURMFELS, B. (1996). *Gröbner Bases and Convex Polytopes*. Amer. Math. Soc., Providence, RI.

THOMAS, R. (1995). A geometric Buchberger algorithm for integer programming. *Math. Oper. Res.* **20** 864–884.

VIRAG, B. (1997). Random walks on finite convex sets of lattice points. Technical report, Dept. Statistics, Univ. California, Berkeley.

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

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

YATES, F. (1984). Tests of significance for 2×2 contingency tables. *J. Roy. Statist. Soc. Ser. A* **147** 426–463.

DEPARTMENT OF MATHEMATICS
WHITE HALL
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853
E-MAIL: ims@math.cornell.edu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720