Multi-Way Contingency Tables in Baseball: An Application of Algebraic Statistics

Brian Minter

December 4, 2015

Abstract

Algebraic statistics is a relatively new field of mathematics that combines methods from algebraic geometry and commutative algebra to develop new techniques of statistical inference. Although there are already methods to test for independence such as Pearson's Chi-Square Test and Fisher's Exact Test, methods from algebraic statistics have shown to be useful when evaluating multi-way and sparse tables. In this paper, we seek to duplicate these results by using Markov Chain Monte Carlo Methods, in particular the Metropolis Hastings Algorithm, to sample multi-dimensional distributions of baseball data to test for independence.

1 Introduction

Algebraic statistics is a relatively new field of mathematics that combines methods from algebraic geometry and commutative algebra to develop new techniques of statistical inference. Algebraic statistics offers a reliable alternative to classic methods of statistical inference. In the past, statisticians have usually relied on classic asymptotic theory. However, when evaluating multi-way tables, algebraic statistics can prove to be a more effective tool. In 1996, Pistone and Wynn [6] published their work on using Gröbner bases to study the relations in factorial designs of experiments. Then in 1998, Diaconis and Sturmfels [5] published their work on constructing Markov chain algorithms for sampling from discrete exponential families. Together, these two papers are credited for introducing algebraic statistics. Since then, the field has been rapidly growing and expanding in different directions.

In this paper, we will first discuss general two-way contingency tables. We will introduce its corresponding notation as well as Pearson's Chi-Square Test for testing independence. Then we will discuss Markov bases and then how to use them to apply them to perform the Metropolis Hastings Algorithm. We will also discuss multi-way tables and then specifically $2 \times 2 \times 2$ contingency tables. Finally we will discuss the results of our research where we attempted to use algebraic statistical methods to emulate traditional statistical methods.

2 Theory

2.1 General Two-Way Contingency Tables

In the following section, we follow the notation from *Lectures on Algebraic Statistics* [5]. A *contingency table* is a table of counts obtined by cross-classifying observed cases according to multiple discrete criteria. Consider the contingency table below from the 2000 Major League Baseball Season that shows hit type cross referenced with the number of outs in the inning when the hit occurred.

2000 MLB Regular Season

		Number of Outs			
		0	1	2	Total
Hit Type	Single	10593	10080	9027	29700
	Double	3170	2917	2184	8271
	Triple	338	319	295	952
	Home Run	2121	1776	1796	5693
	Total	16222	15092	13302	44616

In this case, the two variables are the the number of outs and the hit type. One may wonder whether or not the number of outs in the inning is associated with the type of hit that occurs. In other words, are our two variables independent of each other? In the following sections, we will discuss statistical tests of independence, specifically Pearson's Chi-Square Test of Independence, that will help us answer this question.

Consider the general two-way contingency table between two random variables X and Y where the sample space of X is $[r] = \{1, 2, \ldots, r\}$ and the sample space of Y is $[c] = \{1, 2, \ldots, c\}$. Thus there are $r \times c$ counts in our contingency table. Then the *joint probability* of each count is

$$p_{ij} = P(X = i, Y = j) \text{ where } i \in [r] \text{ and } j \in [c].$$

It follows that the *marginal probabilities* of our contingency table are

$$p_{i+} = \sum_{j=1}^{c} p_{ij} = P(X = i), \text{ where } i \in [r],$$

$$p_{+j} = \sum_{i=1}^{r} p_{ij} = P(Y = j), \text{ where } j \in [c].$$

X and Y are *independent*, denoted $X \perp\!\!\!\perp Y$ if and only if $p_{ij} = p_{i+}p_{+j}$ for all $i \in [r]$ and $j \in [c]$.

Proposition 2.1.1. Let $p = (p_{ij})$ be the $r \times c$ matrix that corresponds to our contingency table between X and Y. Then it follows that X and Y are independent if and only if $\text{rank}(p) = 1$.

Proof. ($\Longrightarrow$): Suppose X and Y are independent. If X and Y are independent, then p is the product of the vector of column sums and the vector of row sums. Thus

$$p = \begin{bmatrix} p_{1+} \\ p_{2+} \\ \vdots \\ p_{r+} \end{bmatrix} \begin{bmatrix} p_{+1} & p_{+2} & \cdots & p_{+c} \end{bmatrix}$$

Consequently $\text{rank}(p) = 1$.

($\Longleftarrow$): Suppose p has rank 1. Since p has rank 1, each row of p is a scalar multiple of each other row. Thus p can be written as $p = ab^T$ for $a \in \mathbb{R}^r$ and $b \in \mathbb{R}^c$. Since all entries in p are non-negative, a and b can be chosen to have non-negative entries as well. Let a_+ and b_+ be the sums of the entries in a and b, respectively. Then $p_{i+} = a_i b_+, p_{+j} = a_+ b_j$, and $a_+ b_+ = 1$. Therefore, $p_{ij} = a_i b_j = a_i b_+ a_+ b_j = p_{i+} p_{j+}$ for all i, j. [5] $\square$

From $p = (p_{ij})$ we can generate a *two-way contingency table* $u = (u_{ij})$ as a table of counts where

$$u_{ij} = \sum_{k=1}^{n} 1_{\{X^{(k)}=i, Y^{(k)}=j\}}, \quad \text{where} \quad i \in [r], j \in [c].$$

Let u_{ij} be the number of observed occurrences such that $\{X = i, Y = j\}$. Then the row total

$$u_{i+} = \sum_{j=1}^{c} u_{ij}$$

counts how often the event $\{X = i\}$ occurred. Similarly, the column total

$$u_{+j} = \sum_{i=1}^{r} u_{ij}$$

counts how often the event $\{Y = j\}$ occurred. Once we know the row sums and column sums, we calculate the *degrees of freedom* of the observed frequency vector u as $(r-1) \times (c-1)$. We calculate it as such since the last row and column can be uniquely determined by the other cells once we know the row sums and columns sums. It should be noted that according to traditional statistical analysis, the row sums and columns sums would be considered a *sufficient statistic*. For our purposes we will simply refer to the row and column sums as all we need to know to perform our statistical analysis.

Suppose we observe n events between our variables X and Y. Then there are n independent pairs

$$\binom{X^{(1)}}{Y^{(1)}}, \binom{X^{(2)}}{Y^{(2)}}, \ldots, \binom{X^{(n)}}{Y^{(n)}}$$

where each pair corresponds to one of the observed cases. Each pair is from the same sample space. In other words

$$P(X^{(k)} = i, Y^{(k)} = j) = p_{ij} \quad \text{for all } i \in [r], j \in [c], k \in [n].$$

For our table of number of outs cross-referenced with the hit type, the corresponding probability matrix is

$$\begin{bmatrix} .2374 & .2259 & .2023 \\ .0711 & .0653 & .0490 \\ .0076 & .0071 & .0066 \\ .0475 & .0398 & .0403 \end{bmatrix}$$

This matrix was derived by taking each cell entry in the table and dividing it by the total number of events. For example, there were 10593 singles hit with one out. Thus its corresponding probability is $10593/44616 = .2374$. The joint probability matrix $p = (p_{ij})$ derived from the observed cases is considered to be an unknown element of the $rc - 1$ dimensional probability simplex

$$\Delta_{rc-1} = \left\{ q \in \mathbb{R}^{rc} : q_{ij} \geq 0 \text{ for all } i, j \text{ and } \sum_{i=1}^{r} \sum_{j=1}^{c} q_{ij} = 1 \right\}.$$

Thus Δ_{rc-1} is the set of all possible probabilistic matrices with dimensions $r \times c$. Note that the dimensions of our probability simplex is $rc - 1$ rather than rc because of the restriction that the sum of the elements of each matrix in our simplex must sum to 1. Hence we have a codimension one subspace of $\mathbb{R}^{rc}$. Also because we require for each element $q_{ij} \geq 0$, we really only have the intersection of this subspace and the first octant in $\mathbb{R}^{rc}$.

The *independence model* for X and Y, which is a subset of Δ_{rc-1} is the set

$$\mathcal{M}_{X \perp\!\!\!\perp Y} = \{p \in \Delta_{rc-1} : \text{ rank}(p) = 1\}.$$

It follows that for all $p = (p_{ij})$ such that $p \in \mathcal{M}_{X \perp\!\!\!\perp Y}$,

$$p_{ij}p_{kl} - p_{il}p_{kj} = 0 \quad \text{for all} \quad 1 \leq i < k \leq r \text{ and } 1 \leq j < l \leq c.$$

In algebraic geometry, the solution set to this equation is known as the *Segre variety*. See Appendix A for more information on Segre varieties.

Similar to Δ_{rc-1}, we denote the set of all possible two-way contingency tables with sample size n by

$$\mathcal{T}(n) = \left\{ u \in \mathbb{N}^{rc} : \sum_{i=1}^{r} \sum_{j=1}^{c} u_{ij} = n \right\}$$

2.1.1 Pearson's Chi-Square Test of Independence

We now reference [1] to discuss Pearson's Chi-Square Test of Independence. Consider the following two hypotheses:

$$H_0 : p \in \mathcal{M}_{X \perp\!\!\!\perp Y} \qquad \text{and} \qquad H_1 : p \notin \mathcal{M}_{X \perp\!\!\!\perp Y}.$$

Then H_0 is true if our two variables X and Y are independent, hence $p_{ij} = p_{i+}p_{+j}$. Assuming the null hypothesis is true, the expected number of the joint event $\{X = i, Y = j\}$ is $np_{i+}p_{+j}$ where n is our sample size is called the *expected frequency*. It follows that the corresponding observed proportions are

$$\hat{p}_{i+} = \frac{u_{i+}}{n} \qquad \text{and} \qquad \hat{p}_{+j} = \frac{u_{+j}}{n}$$

where u_{i+} and u_{+j} are the observed row totals and column totals respectively. We can now estimate our expected counts $np_{i+}p_{+j}$ as $\hat{u}_{ij} = n\hat{p}_{i+}\hat{p}_{+j}$. To compare our observed counts to the expected counts, we use the *chi-square statistic* developed by Karl Pearson in 1900 [1]. We calculate the Pearson chi-square statistic of the observed counts, $X^2(u)$, as

$$X^2(u) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(u_{ij} - \hat{u}_{ij})^2}{\hat{u}_{ij}}.$$

If the null hypothesis is true, then we can expect our observed counts to be "close" to the expected counts. Hence the chi-square statistic would be small. If X^2 is large enough, we reject the null hypothesis. To determine our criterion to reject the null hypothesis, we use the *chi-square test*. We calculate the probability, or the *p-value*, of observing the counts in our contingency table, or something more extreme, given the null hypothesis is true. A small p-value means it's unlikely to observe the counts given. Hence a small p-value provides support to reject the null hypothesis. Typically, statisticians have accepted that a p-value below 0.05 or 0.01 provides adequate evidence against the null hypothesis. A large p-value doesn't provide support for the null hypothesis. Rather, we say the chi-square test was inconclusive.

Consider the contingency table given earlier of hit types cross classified by the number of outs in the inning when the hit occurred. Using the methods previously discussed, we calculate the chi-square statistic as 72.796 with 6 degrees of freedom and the p-value is less than 0.001. Since the p-value is less than 0.05, we find evidence to reject the null hypothesis. Thus the hit type and how many outs in the inning are not independent.

2.1.2 Walks on Fibers

We now reference [2] to discuss configuration matrices and fibers. Given the row sums and column sums for an $r \times c$ matrix x, let t be the column vector of the row and column sums. Thus

$$t = (x_{1+}, \ldots, x_{r+}, x_{+1}, \ldots, x_{+c}).$$

Also let u be the frequency vector of the $r \times c$ case. Hence

$$u = (x_{11}, x_{12}, \ldots, x_{1c}, x_{21}, \ldots, x_{rc}).$$

It follows that

$$t = Au$$

where A is the $(r + c) \times rc$ *configuration* matrix. Note that for a 2×2 matrix,

$$\begin{pmatrix} u_{1+} \\ u_{2+} \\ u_{+1} \\ u_{+2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ u_{21} \\ u_{22} \end{pmatrix}.$$

Hence for 2×2 matrices, the configuration matrix is the one above. Similarly, for the counts given earlier from type of hits cross-referenced with the number of outs during the 2000 MLB season, the corresponding matrices in the equation $t = Au$ is

$$\begin{pmatrix} 29700 \\ 8271 \\ 952 \\ 5693 \\ 1622 \\ 15091 \\ 13302 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 10593 \\ 10080 \\ 9027 \\ 3170 \\ 2917 \\ 2184 \\ 338 \\ 319 \\ 295 \\ 2121 \\ 1776 \\ 1796 \end{pmatrix}.$$

For every configuration matrix A, $\text{rank}(A) = r + c - 1$. It follows that by the Rank Theorem, that $\dim(\ker(A)) = rc - (r + c - 1)$. Thus $\dim(\ker(A)) = (r - 1)(c - 1)$. Given the row sums and column sums, we define the *conditional sample space*, denoted $\mathcal{F}_t$ as

$$\mathcal{F}_t = \{x \in \mathbb{Z}^{rc} : x \geq 0, t = Ax\}.$$

Our conditional sample space will also be referred to as the t-fiber, or the fiber according to t. The t-fiber is the set of all vectors in $\mathbb{Z}^{rc}$ where each vector has nonnegative integer cell counts and multiplying its corresponding configuration matrix yields t. For example, let t be the matrix of row sums followed by the column sums according to our table given previously of hit type cross-referenced with the number of outs. Then

$$t = \begin{bmatrix} 29700 \\ 8271 \\ 952 \\ 5693 \\ 1622 \\ 15091 \\ 13302 \end{bmatrix}$$

Then let s be the matrix

$$s = \begin{bmatrix} 10593 \\ 10080 \\ 9027 \\ 3170 \\ 2917 \\ 2184 \\ 338 \\ 319 \\ 295 \\ 2121 \\ 1776 \\ 1796 \end{bmatrix}.$$

Then $s \in \mathcal{F}_t$ because

$$\begin{bmatrix} 29700 \\ 8271 \\ 952 \\ 5693 \\ 1622 \\ 15091 \\ 13302 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 10593 \\ 10080 \\ 9027 \\ 3170 \\ 2917 \\ 2184 \\ 338 \\ 319 \\ 295 \\ 2121 \\ 1776 \\ 1796 \end{bmatrix}$$

Since we want to consider exact tests, we want to evaluate the *hypergeometric distribution* over $\mathcal{F}_t$. Once we have a test statistic $\phi(x)$ and the row sums and column sums of a contingency table, we want to evaluate its distribution, where x is distributed according to the hypergeometric distribution over $\mathcal{F}_t$. We can use this distribution to estimate the likelihood of observing a matrix such as t, given that H_0 holds. This determines our p-value and allows us to decide whether to reject the null hypothesis or not.

If we are given a test statistic ϕ such that the larger ϕ is, the more likely we are to reject the null hypothesis H_0, then we can measure the deviation from H_0 with its *p-value*. If we are given the *level of significance*, denoted α, we reject H_0 if $p \leq \alpha$. Let x_0 be the observed contingency table. Then we calculate the p-value as

$$p = P(\phi(x) \geq \phi(x_0) \mid H_0) = \sum_{x \in \mathcal{F}_t, \phi(x) \geq (x_0)} p(x \mid t = Ax_0, H_0)$$

which is the probability of observing $\phi(x)$ greater than or equal to $\phi(x_0)$. We will calculate the p-value by a Markov Chain Monte Carlo method where we randomly sample from x and use the hypergeometric distribution over $\mathcal{F}_t$ as our stationary distribution.

2.2 Markov Bases

When the conditional sample space is large, a Markov Chain Monte Carlo method provides an effective way to sample from the probability distribution. Consider the independence model of general two-way contingency tables discussed previously. The *t-fiber*, denoted $\mathcal{F}_t$ is the set of $r \times c$ contingency tables with fixed row and column sums according to a contingency table t. Thus

$$\mathcal{F}_t = \{u \geq 0 : u_{i+}, i \in [r], u_{+j}, j \in [c] \text{ are fixed according to } t\}.$$

Let A be the matrix discussed in section 2.1. The kernel of A, denoted by $\ker(A)$, is the set of vectors such that $Ax = 0$ where $x \in \ker(A)$. Then the *integer kernel* is set of vectors in $\ker(A)$ defined by

$$\ker_{\mathbb{Z}}(A) = \{x : Ax = 0, x \in \mathbb{Z}^{r \times c}\}.$$

A *move* is an element of $\ker_{\mathbb{Z}}(A)$. If $s \in \mathcal{F}_t$ and $x \in \ker_{\mathbb{Z}}(A)$ then

$$A(s + x) = As + Ax = As + 0 = t$$

Thus $s + x \in \mathcal{F}_t$. Now consider the following integer matrix $z(i_1, i_2; j_1, j_2) = \{z_{ij}\}$ where

$$z_{ij} = \begin{cases} +1, (i,j) = (i_1, j_1), (i_2, j_2), \\ -1, (i,j) = (i_1, j_2), (i_2, j_1), \\ 0, \text{ otherwise.} \end{cases}$$

Adding $z(i_1, i_2; j_1, j_2)$ to a contingency table doesn't change its row sums and column sums. Consider adding the move $z(i_1, i_2; j_1, j_2)$ to the observed count matrix t for our hit type versus number of outs table. Then

$$\begin{bmatrix} 10592 & 10081 & 9027 \\ 2171 & 2916 & 2184 \\ 338 & 319 & 295 \\ 2121 & 1776 & 1796 \end{bmatrix} + \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 10593 & 10080 & 9027 \\ 2170 & 2915 & 2184 \\ 338 & 319 & 295 \\ 2121 & 1776 & 1796 \end{bmatrix}$$

Since the row sums and column sums don't change, $z(i_1, i_2; j_1, j_2)$ is in $\ker_{\mathbb{Z}}(t)$. Thus $z(i_1, i_2; j_1, j_2)$ is a move. In fact, we call $z(i_1, i_2; j_1, j_2)$ a *basic move* of the independence model for 2-way contingency tables. It follows that $x + z$ is in the fiber of t as long as all the elements in $x + z$ are still nonnegative. If we set $i_2 = r$ and $j_2 = c$, for all z then the set

$$\{z(i_1, r; j_1, c) \mid 1 \leq i_r \leq r - 1, 1 \leq j_1 \leq c - 1\}$$

forms a *lattice basis* of $\ker_{\mathbb{Z}}(A)$ since every move in $\ker_{\mathbb{Z}}(A)$ can be written as a combination of $z(i_1, r; , j_1, c)$. Let $\mathcal{B} \subset \ker_{\mathbb{Z}}(A)$ be a set of moves for the configuration matrix A. Then $\mathcal{B}$ is a *Markov basis* if and only if for all fibers $\mathcal{F}_t$ and all elements $x, y \in \mathcal{F}_t$, where $x \neq y$, there exists moves $z_1, \ldots, z_k \in \mathcal{B}$ and $\varepsilon_1, \ldots, \varepsilon_k \in \{1, -1\}$, where

$$y = x + \sum_{i=1}^{k} \varepsilon_i z_i, \text{ and } x + \sum_{i=1}^{l} \varepsilon_i z_i \in \mathcal{F}_t, \text{ where } l = k - 1.$$

In other words, $\mathcal{B}$ is a Markov basis if every element $x \in \mathcal{F}_t$ can be connected to $y \in \mathcal{F}_t$ by adding or subtracting elements of $\mathcal{B}$. Also, each move from x to y stays within $\mathcal{F}_t$. Therefore, if we're given a Markov basis, we can move throughout an entire fiber.

2.3 Metropolis Hastings Algorithm

Once we have a Markov basis, we can construct an irreducible and symmetric Markov chain over $\mathcal{F}_{x_0}$ where x_0 is the observed frequency vector. We can apply the Metropolis-Hastings algorithm to this Markov chain. The steps to the algorithm are below:

Input: A contingency table $u \in \mathcal{T}(n)$ and a Markov basis $\mathcal{B}$ for the model $\mathcal{M}_A$.
Output: A sequence of chi-square statistic values $(X^2(v_t))_{t=1}^{\infty}$ for tables v_t in the fiber $\mathcal{F}(u)$.
Step 1: Initialize $v_1 = u$.
Step 2: For $t = 1, 2, \ldots$ repeat the following steps:
(i) Select uniformly at random a move $u_t \in \mathcal{B}$.
(ii) If $\min(v_t + u_t) < 0$, then set $v_{t+1} = v_t$, else set

$$v_{t+1} = \begin{cases} v_t + u_t \\ v_t \end{cases} \quad \text{with probability} \quad \begin{cases} q \\ 1 - q \end{cases}$$

where

$$q = \min\left\{1, \frac{P(U = v_t + u_t | AU = Au)}{P(U = v_t | AU = Au)}\right\}.$$

In other words, we start at our observed matrix. Then for $t = 1, 2, \ldots$ we randomly select a move, u_t from our Markov bases. If $v_t + u_t$ has a negative entry then we don't make the move and we go back to step 2. If not then we will add u_t with probability q where q is the minimum of 1 and the ratio of the probability of observing the matrix $v_t + u_t$ divided by the probability of observing the matrix v_t. Note that the probability

$$P(U = u | AU = Au) = \frac{1/(\prod_{i \in \mathcal{R}} u_i!)}{\sum_{v \in \mathcal{F}(u)} 1/(\prod_{i\mathcal{R}} v_i!)}.$$

It follows that

$$\frac{P(U = v_t + u_t | AU = Au)}{P(U = v_t | AU = Au)} = \frac{(\prod_{i \in \mathcal{R}} (u_i + v_i)!)}{(\prod_{i \in \mathcal{R}} u_i!)}.$$

Theorem 2.4.1 The outputs of the Metropolis-Hastings algorithm is an aperiodic, reversible and irreducible Markov chain that has stationary distribution equal to the conditional distribution if $X^2(U)$ given $AU = Au$ where U is the expected frequency matrix and u is the observed frequency matrix.

Corollary 2.4.2 With probability one, the output of the Metropolis-Hastings algorithm satisfies

$$\lim_{M \to \infty} \frac{1}{M} \sum_{t=1}^{M} 1_{\{X^2(v_t) \geq X^2(u)\}} = P(X^2(U) \geq X^2(u) | AU = Au)$$

where $v_t \in \mathcal{F}_t$. Thus by performing the Metropolis-Hastings we can estimate the p-value of our given matrix.

Consider the table given in the first section. In [3], they perform the Metropolis-Hastings Algorithm with 100,000 steps. When we perform the Metropolis-Hastings algorithm, we take 1 million steps, and obtain a p-value of 0.002283. Regardless of whether our level of significance is 0.01 or 0.05, based on the p-value we reject the null hypothesis, and conclude that the type of hit that occurs is not independent of the number of outs when the hit occurs. In other words, the number of outs in an inning affects the likelihood of each hit type.

2.4 Multi-Way Tables

Now consider if we have more than two random variables. Let $X_1, \ldots, X_m$ be random variables with X_l having values in $[r_l]$. Let $\mathcal{R} = \prod_{i=1}^{m}[r_i]$ be the set of all combinations between our random variables. Then let the joint probabilities be defined by

$$p_i = p_{i_1 \ldots i_m} = P(X_1 = i_1, \ldots, X_m = i_m), i = (i_1, \ldots, i_m) \in \mathcal{R}.$$

We can then create a joint probability table determined by our joint probabilities as $p = (p_i | i \in \mathcal{R})$ which is an element of the $\#\mathcal{R} - 1$ dimensional probability simplex $\Delta_{\mathcal{R}-1}$.

Suppose we observe n events that are elements of our sample space. Then we can represent each case as

$$\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_m^{(1)} \end{pmatrix}, \begin{pmatrix} X_1^{(2)} \\ \vdots \\ X_m^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} X_1^{(n)} \\ \vdots \\ X_m^{(n)} \end{pmatrix}.$$

It follows that the number of cases of each element of our sample space can be summed as

$$U_i = \sum_{k=1}^{n} 1_{\{X_1^{(k)} = i_1, \ldots, X_m^{(k)} = i_m\}}, \text{ where } i = (i_1, \ldots, i_m) \in \mathcal{R}.$$

Based on the counts obtained from our events, we can create an m-way table $U = (U_i)$ in $\mathbb{N}^{\mathcal{R}}$. Let the set of all possible contingency tables with sample size n in $\mathbb{N}^{\mathcal{R}}$ be

$$\mathcal{T}(n) = \left\{ u \in \mathbb{N}^{\mathcal{R}} : \sum_{i \in \mathcal{R}} u_i = n \right\}.$$

2.5 Independence Models of 2×2×2 Contingency Tables

Consider a $2 \times 2 \times 2$ contingency table derived from n observed cases between discrete random variables X, Y, and Z. Let $x = (x_{ijk})$ be the matrix of counts corresponding to the n observed cases where $i, j, k \in \{1, 2\}$ and let t be the vector of i sums, j sums, and k sums. If we order

t according to j, then $t = Ax$ is written as

$$\begin{bmatrix} x_{\{1,2\}}(1,1) \\ x_{\{1,2\}}(2,1) \\ x_{\{2,3\}}(1,1) \\ x_{\{2,3\}}(1,2) \\ x_{\{1,2\}}(1,2) \\ x_{\{1,2\}}(2,2) \\ x_{\{2,3\}}(2,1) \\ x_{\{2,3\}}(2,2) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x(1,1,1) \\ x(1,1,2) \\ x(2,1,1) \\ x(2,1,2) \\ x(1,2,1) \\ x(1,2,2) \\ x(2,2,1) \\ x(2,2,2) \end{bmatrix}.$$

Let x_1 be the upper left 4×4 matrix of x and let x_2 be the bottom right 4×4 matrix of x. Similarly, let t_1 be the 4-dimensional subvector containing the first four cells of t and let t_2 be the 4-dimensional subvector containing the last four cells of t. Then $x \in \mathcal{F}_t$ if and only if $x_1 \in \mathcal{F}_{x_1}$ and $x_2 \in \mathcal{F}_{t_2}$. We can now apply methods of the general 2×2 independence model. It follows that the independence model of 3-way tables consisting of all binary random variables can be represented by the equation

$$p_{111}p_{122}p_{212}p_{221} - p_{112}p_{121}p_{121}p_{222} = 0.$$

In other words, the independence model is the set of all tables whose entries satisfy this quartic equation.

3 Appendix

3.1 Appendix A: Segre Varieties

To discuss Segre Varieties, we reference [4]. In algebraic geometry, an *algebraic variety* is a set of solutions to a system of polynomials. The *Segre maps*

$$\sigma : \mathbb{P}^n \times \mathbb{P}^m \to \mathbb{P}^{(n+1)(m+1)-1}$$

are defined by assigning $([X], [Y])$ to the point in $\mathbb{P}^{(n+1)(m+1)-1}$ whose coordinates are the pairwise products of the coordinates of $[X]$ and $[Y]$. In other words,

$$\sigma : ([X_0, \ldots, X_n], [Y_0, \ldots, Y_m]) \mapsto [\ldots, X_i Y_j, \ldots]$$

where the coordinates in the target space range over all pairwise products of coordinates X_i and Y_j. The image of the Segre map is an algebraic variety, called a *Segre variety*. If we let Z_{ij} be the coordinates of the target space, we can see that it is the common zero locus of the quadratic polynomials $Z_{ij} \cdot Z_{kl} - Z_{il} \cdot Z_{kj}$. The Segre variety is also called a *determinantal variety* because it is the the zero locus of the 2×2 minors of the matrix (Z_{ij}). It follows that given an $m \times n$ matrix U, then (U_{ij}) will have rank 1 if and only if $U = V^T W$ where $V = (V_1, \ldots, V_m)$ and $W = (W_1, \ldots, W_n)$ are vectors.

References

[1] A. Agresti, Categorical Data Analysis, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York, (1990).

[2] Aoki, S., Hara, H., Takemura, A.: Markov Bases in Algebraic Statistics. Springer (2012)

[3] Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. Ann. Statist., 363-397 (1998)

[4] Harris J.: Algebraic Geometry, A First Course. Springer (1992).

[5] Drton, M., Sturmfels, B., Sullivant, S.: Lectures on Algebraic Statistics. Birkhauser, Basel (2009)

[6] Pistone, G. & Wynn, H. P. (1996). Generalised confounding with Gröbner bases.