ALGEBRAIC THEORY OF CONDITIONAL AND SEQUENTIAL CONDITIONAL TESTS OF THREE WAY CONTINGENCY TABLES

Sakata, Toshio
Department of Human living system design, Kyushu University

Sawae, Ryuichi
Department of Applied Mathematics, Okayama University of Science

https://doi.org/10.5109/12590

出版情報：Bulletin of informatics and cybernetics. 37, pp.41-48, 2005-12. Research Association of Statistical Sciences
バージョン：
権利関係：

ALGEBRAIC THEORY OF CONDITIONAL AND SEQUENTIAL CONDITIONAL TESTS OF THREE WAY CONTINGENCY TABLES

by

Toshio Sakata and **Ryuichi Sawae**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.37

FUKUOKA, JAPAN
2005

Bulletin of Informatics and Cybernetics, Vol. 37, 2005

ALGEBRAIC THEORY OF CONDITIONAL AND SEQUENTIAL CONDITIONAL TESTS OF THREE WAY CONTINGENCY TABLES

By

Toshio Sakata* and **Ryuichi Sawae†**

Abstract

The conditional test of no three way interaction for three way tables is overviewed, based on algebraic point of views. A sequential conditional test is proposed and its performance is numerically tested.

Key Words and Phrases: Gröbner base, lattice base method, echelon form, three way contingency tables, log-linear model, sequential exact test, sequential MCMC test, power performance.

1. Introduction

In the analysis of contingency tables the conditional inference with fixed margins, called Fisher's exact test, are often used. Then a p-value is usually obtained by the χ^2 approximation of a null distribution. However the approximation might be doubtful if the sample size is not large. Some ingenious programming techniques to calculate it exactly have been developed by Mehta and Patel (1983) and others. Recently MCMC technique has been applied to the calculation. That is, a Metropolis walk with the hypergeometric distribution as the limiting distribution is generated and a p-value is estimated by the relative proportion of more extreme tables than the given one among all tables generated in Ω, where Ω denotes the set of all tables with given margins. The literature has shown that the method is useful(for example, Sakata et al. (1997)). Diaconis and Sturmfels (1998) has shown that the generation of a Metropolis walk is realized by using a Markov basis and it is obtained through the Gröb ner basis of a toric ideal in the ring of polynomial functions. For the case of two way tables the calculation of a Gröbner basis is very easy. The toric ideal is the Segre ideal generated by all 2×2 minors(see Sturmfels (1991)). In this paper we consider three way tables and then the situation drastically changes and the problem becomes a very difficult one. There are two methods, Elimination method and Lattice basis method, in calculating the Gröbner basis. Elimination method is sensitive to the size of the tables(see Diaconis and Sturmfels (1998), and Sturmfels (1991)). On the other hand the lattice basis method is comparatively robust to the size of tables. Sakata and Sawae (2000) derived a lattice basis for contingency tables through the echelon form of a defining matrix and used it for running MCMC for three way tables. On the other hand

* Department of Human living system design, Kyushu University Shiobaru 4–9–1 Minami-ku Fukuoka 815-85401 Japan. tel +81–92–642–2693 sakata@design.kyushu-u.ac.jp

† Department of Applied Mathematics, Okayama University of Science Ridai-cho 1–1 Okayama 700-0005 Japan. tel +81–86–256–9441 sawae@xmath.ous.ac.jp

Sakata and Sawae (2000), Sakata and Sawae (2001) and Sakata and Sawae (2003) developed a sequential conditional test for two way tables, based on Sasaki's creation operator of some partial differential equation systems, motivated by Saito et al. (1997), which includes an application to linear programming. They showed that the sequential test performs well for relatively small two way tables. Finally we mention some studies for three dimensional tables are still developing in relation to the Diaconis and Sturmfels's algebraic theory(for example see Aoki (2003)). The outline of our paper is as follows. In Section 2 we briefly summarize our method of generating three way contingency tables and an alternative model with three way interaction is newly proposed. The sequential conditional test is proposed in Section 3 and the performance of the test, the power and mean sample sizes, under the alternative model are studied by simulation for $3 \times 3 \times 3$ and $3 \times 3 \times 4$ tables.

2. Conditional test of no three way interaction

Here in the first subsection we briefly review algorithms of generating three way contingency tables with given margins, with a special comment of our programming and in the second subsection we propose an alternative model with three way interaction for which we test the null hypothesis.

2.1. A method of generating tables with given margins

(A) The first algorithm is based on Gröbner basis. For this we have two different algorithms (A-1) and (A-2).

(A-1) This is the method using directly a Gröbner basis for three way tables on the line of Diaconis and Sturmfels. The reduced Gröbner basis gives a Markov basis for the connecting random walk on the set of all three way tables with the same margins as a given table. This becomes a basis of MCMC-test. However, as the algorithm 5.7 on the page 44 in Sturmfels (1991) shows, the Gröbner basis is also used for the enumeration of all tables. Takemura and Aoki(Takemura and Aoki (2002)) determined the minimal Markov basis with success up to $4 \times 4 \times 4$.

(A-2) This is the method using the creation operator. This method calculates dynamically the characteristic function of the set of tables satisfying the marginal conditions, by applying the creation operator for a A-hypergeometric partial differential system.

In any case when the sizes, K, L and M, are greater than 4 the calculation requires tremendous time.
For $(A-1)$ the creation operator for three way contingency tables is at present not explicitly given algorithmically, unlike two way table case realized by Sasaki's creation operator.

(B) Here we summarize a simple method to be adopted in this paper, which is a method to generate two way tables several times by iteratively, or, directly manipulating

the free variables of the cell counts. For $M = 4$, the former one is described as follows.

(B-1) (1) First take a table as the M piles of $K \times L$ two way contingency tables.
(2) All admissible top $K \times L$ two way tables are searched and stocked in the disk file.
(3) The top tables are again read from the disk file and for each inputted top table the second top $K \times L$ admissible tables are searched.
(4) For each pair of the admissible top and the second table the third admissible tables are searched.
(5) For the admissible triple of the upper three tables the bottom table is uniquely determined.
(6) Whenever an inadmissible table occurs we go back one step.

Note that each two way table is determined row by row. So we only need the memory for the partition pattern of integers.

(B-2) For such small tables as $3 \times 3 \times 3$ and $3 \times 3 \times 4$ treated in this paper the counts of each cell are treated as free variables and are made varied between 0 and a marginal constraint. This is a brute way but most speedy if possible.

For $3 \times 3 \times 3$ tables, by this brute method, we could obtain the following list of the numbers of tables with specified margins.

Table 1
The number of $3 \times 3 \times 3$ tables with equal margins
calculation by the brute method

equal margins	6	9	12
the number of tables	43687	619219	4547458
equal margins	15	18	
the number of tables	22454470	85034962	

Further for the equal margin, 21, 24 and 27 the total number of tables are 266351932, 723488194, and 1758926584, respectively. We note that for these values of margins the brute method is performable within a reasonable time.

2.2. Alternative hypothesis

In this paper no three way interaction hypothesis is tested for the alternative hypothesis given in the later. The log-linear model for a three way contingency table decomposes the mean structure of cells as follows.

$$\log \mu_{ijk} = \mu + \mu_{(x)i} + \mu_{(y)j} + \mu_{(z)k} + \mu_{(xy)ij} + \mu_{(yz)jk} + \mu_{(zx)ki} + \mu_{(xyz)ijk}$$

where μ denotes the general mean and $\mu_{(x)}, \mu_{(y)}$ and $\mu_{(z)}$ denote the main effect and $\mu_{(xy)}, \mu_{(yz)}$, and $\mu_{(zx)}$ denote the two factor interaction and $\mu_{(xyz)}$ denotes the three factor interaction respectively. In testing hypothesis H:three factor interaction does not exist, the set of two way margins $\{m_{(xy)}, m_{(yz)}, m_{(zx)}\}$ is the sufficient statistics for the multinomial model with no three factor interaction. So, the conditional distribution of tables given the margins is of parameter free. Hence we can derive the conditional null

distribution of the exact conditional test. Further, we consider the following alternative model.
Choose a vector v and decompose it as

$$(P_0 + P_x + P_y + P_z + P_{xy} + P_{yz}v + P_{zx}v + P_{xyz})v$$

where P's denote the projection matrices of the corresponding effect spaces. And from this we construct a new vector

$$(P_{(1)} + P_{(x)} + P_{(y)} + P_{(z)} + P_{(xy)} + P_{(yz)} + P_{(zx)} + \rho P_{(xyz)})v.$$

Larger ρ values correspond monotonically to the larger three way interaction. Note that the selection of v is arbitral. In our experiment v was generated randomly. For example, for a $3 \times 3 \times 3$ table, 27 dimensional vector v is randomly drawn(27 independent uniform random variables on $[0, 1]$ were generated). After that, $(P_{(1)} + P_{(x)} + P_{(y)} + P_{(z)} + P_{(xy)} + P_{(yz)} + P_{(zx)} + \rho P_{(xyz)})v$ was calculated and normalized as a multinomial probability after exponentialization.

3. Sequential conditional test

3.1. General view

Sakata and Sawae (2000) newly developed a theory of a sequential conditional test of two way contingency tables, which has not appeared in the literature. In medical statistics it is very important to obtain a definite test result as early as possible. So the sequential test of contingency tables seems to have a potential importance. In Sakata and Sawae (2000) it was first shown that the theory of creation operators, the inverse operators of partial differential operators, is available to the sequential conditional test. In fact it originated from the theory of the hypergeometric A-systems of partial differential equations and the mathematical foundations were developed in Gelfand et.al. (1989), Saito et al. (1997), Sasaki (1991), Sturmfels and Takayama (1998) and so on. It is noted that creation operators are obtained by using a Gröbner basis in the non-commutative Weyl algebra of partial differential operators. The paper of Saito et al. (1997) dealt with an application to integer programming, which gave us a strong motivation. Again it is fortunate that for two way tables the creation operators are also easy to obtain(see Sasaki (1991) and for the general case see Saito et al. (1997)) but the available algorithm for three way tables is not known up to now and some efforts are now undergoing which discover the neat algorithms calculating the operator.

3.2. Experimental results

In this section we show the experimental results of our sequential test.

3.2.1. $3 \times 3 \times 3$ tables

For $3 \times 3 \times 3$ tables at each step of the sequential test 9 samples are drawn and then the conditional test is performed. If the calculated p-value is less than 0.05(0.025) the test rejects the hypothesis H and stop. Otherwise we proceed to the next step and takes another 9 samples and performs the conditional test and so on.
If the rejection does not occur within 10 steps we stop the sequential test and accept

the null hypothesis.
At each step for the data table drawn the conditional probability of the table, p_0, is calculated. For all tables with the same margins with the data table calculate the conditional probability and if it is smaller than p_0 add it to the p-value. The total sum of these conditional probabilities become the p-value of the data table.
For calculation of the probability of each table under null distribution the technique in the paper of Morgan and Blumenstein (1991) was used avoiding the overflow. If the calculated p value is less than 0.05(0.025) we consider that the data table rejects the null hypothesis. By 1000 times simulations, the proportions of rejection of the sampled tables are reported as the power in the tables 2 and 3. The mean sample lengths are also reported.

Table 2
Performance of the sequential test for 3×3×3 **tables**

Critical p**-value=**0.025 **case**					
ρ	0.0	0.4	0.8	1.2	1.6
power	0.068	0.123	0.279	0.531	0.80
mean sample size	87.4	86.1	82.0	74.0	62.9

Table 3
Performance of the sequential test for 3×3×3 **tables**

Critical p**-value=**0.05 **case**					
ρ	0.0	0.4	0.8	1.2	1.6
power	0.139	0.216	0.381	0.650	0.881
mean sample size	84.9	82.2	77.2	66.7	55.6

From the tables 2 and 3 we see the sequential test using a level 0.05(0.025) at each step has the overall significance level 0.139(0.68). And we see the test has a good power performance and the sequential test seems to be usable. Note that the number of repetition of simulation is 1000 for each $\rho = 0.0$ to 1.6 due to the limitation of simulation time. One key possible problem might be how to determine the critical p-value to make the test with an exact size α. However, this is easily done if we perform more detailed simulations for various critical p-values and choose the critical p-value for which the simulated size is nearest to α. This is, in fact, possible in reasonable time. However, if we want to take more samples or perform a lager size of three way tables, we will come to a situation that the number of tables becomes too big to manipulate. In such case we had better switch to the method of estimating p-value by using almost uniform sampling by MCMC on the set of all contingency tables with given margins. Here we used the minimal Markov basis in the paper of Takemura and Aoki (2002) to generate a Markov chain. For $3 \times 3 \times 3$ tables the number of the minimal basis is 81. The simulation results of the MCMC based sequential test are given below. Note that the length of the MCMC chains was 10000.

Table 4
Performance of the sequential test for $3 \times 3 \times 3$ **tables by MCMC**

Critical p**-value=**0.025 **case**					
ρ	0.0	0.4	0.8	1.2	1.6
power	0.08	0.12	0.36	0.72	0.94
mean sample size	131	129	119	101	79

Table 5
Performance of the sequential test for $3 \times 3 \times 3$ tables by MCMC

Critical p-value=0.05 case					
ρ	0.0	0.4	0.8	1.2	1.6
power	0.13	0.23	0.50	0.81	0.98
mean sample size	127	122	112	91	71

As a conclusion the sequential conditional test seems to be useful for three way contingency tables of the size $3 \times 3 \times 3$ even for PC based computational environment, if we choose the exact test or MCMC test depending on situations. However a more detailed simulation is needed.

3.2.2. $3 \times 3 \times 4$ tables

Below is the performance of the sequential test for $3 \times 3 \times 4$ table. Here the sample size of each step is set as 9 and the maximum step is set as 10. The simulation is repeated 1000 times. Note that the number of the minimal Markov basis is 450 for $3 \times 3 \times 4$ table and they are used in this simulation.

Table 6
Performance of the sequential test for $3 \times 3 \times 4$ tables by MCMC

Critical p-value=0.025 case					
ρ	0.0	0.4	0.8	1.2	1.6
power	0.083	0.142	0.355	0.592	0.814
Mean sample size	86.3	83.8	75.8	65.8	54.7

Table 7
Performance of the sequential test for $3 \times 3 \times 4$ tables by MCMC

Critical p-value=0.05 case					
ρ	0.0	0.4	0.8	1.2	1.6
power	0.159	0.266	0.474	0.710	0.879
Mean sample size	82.3	77.9	77.9	57.5	48.4

The results show that the proposed sequential test works well for relatively small size of three way tables. However, after this work we made a program to perform the sequential test for arbitral size of three way tables of height 4. Further we are now developing the exact p-value calculation by the method (A-2) using the creation operator for three way tables. We may be able to report the new results for possibly larger size of tables in the future.

References

Agresti, A. (1992). A survey of exact inference for contingency tables,*Statistical Science*, **7**, 131–153.

Aoki, S. (2003). The list of indispensable moves of the unique minimal Markov basis for $3 \times 4 \times K$ and $4 \times 4 \times 4$ contingency tables with fixed two-dimensional marginals, *Mathematical Engineering Technical Reports, 2003-38, Depr. of Mathematical Informatics, the university of Tokyo.*

Asututi, E. and Yanagawa, T. (2000). Testing Nonlinear Trend in Proportions under Binomial and Extra-Binomial Variability, *Bulletin of Information and Cybernetics*, **32**, 157–168.

Diaconis, P and Sturmfels, B. (1998). Algebraic Algorithms for Sampling Conditional Distributions. *Ann. of Statistics*, **26**, 363–397.

Gelfand,I.M., Zelevinski, A.V. and Kapranov, M.M. (1989). Hypergeometric functions and toral manifolds, *Functional Anal. and its appli.*, **23**, 94–106.

Morgan, W. and Blumenstein, B. (1991). Exact Conditional Tests for Hierarchical Models in Multidimensional Contingency Tables., *Applied Statistics*, **40**, 435–442.

Mehta, C.R. and Patel, N.R. (1983). A network algorithm for performing Fishier's exact test in $r \times c$ contingency tables,*J.A.S.A.*, **78**, 427–434.

Saito, M., Sturmfels, B. and Takayama, N. (1997). a Hypergeometric Polynomials and Integer Programming, *Composition Mathematica*, **115**, 185–204.

Sakata, T., Nomakuchi, K. and Hayashi, T. (1997). Generalized shift test method and Metropolis Walk, *Bulletin of Information and Cybernetics*, **19**, 1–13.

Sakata, T. and Sawae, R. (2000). A new development of the conditional inference of contingency tables, *Proceed. of the 7-th Japan-China Symposium on Statistics, Tokyo*, 271–274.

Sakata, T. and Sawae, R. (2000). Echelon form and conditional test for three way contingency tables, *Proceed. of the 10-th Japan and Korea Joint Conference of Statistics*, 333–338.

Sakata, T. and Sawae, R. (2001). Applications of Gröbner Basis to Analysis of Contingency Tables and Integer Programming, *The proceeding C.D. of ISPR-16, Praha.*

Sakata, T. and Sawae, R. (2003). A study of the sequential conditional test for contingency tables,*J. Japan Society of Computational Statistics*, **15**, 169–174.

Sasaki, T. (1991). Contiguity relations of Aomoto-Gelfand's hypergeometric functions and applications to Apell's system F_3 and Goursat's system ${}_3F_2$,*SIAM J. of Math. Anal.*, **22**, 821–846.

Sturmfels, B. (1991). Gröbner Bases and Convex Polytopes, *University Lect. Ser., A.M.S.*

Strumfels, B. and Takayama, N. (1998). Gröbner Basis and Hypergeometric Functions, Gröbner Bases and Applications, *B.Buchberger and F.Winkler (eds.), Camb. Univ. Press, London Math. Society Lecture Notes Series* **251**, 246–258.

Takemura, A. and Aoki, S. (2002). Some characterizations of minimal Markov basis for sampling from discrete conditional distributions, *Technical Report of Dept. Math. Engineering and Information, Physics school of engineering, Tokyo University.*

Zelterman, D., Chan, I.S., and Mielke, P.W. (1995). Exact tests of significance in higher dimensional tables, *The American Statistician*, **49**, 357–361.

Received October 27, 2003
Revised April 21, 2004