INFERENCE PROCEDURES FOR THE SYMMETRY IN A CONTINGENCY TABLE

Geng, Zhi
Graduate School of Engineering Sciences, Kyushu University

https://doi.org/10.5109/13394

出版情報：Bulletin of informatics and cybernetics. 23 (1/2), pp.55-65, 1988-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 23, No. 1~2, 1988

INFERENCE PROCEDURES FOR THE SYMMETRY IN A CONTINGENCY TABLE

By

Zhi GENG*

Abstract

Several inference procedures are discussed for the partial symmetry and the complete symmetry in a two-dimensional contingency table. The inference procedures proposed in the present paper is firstly to select a model fitted to observed frequencies, and secondly to estimate the parameters of the selected model. Such procedures are also called those of the preliminary test estimation in the situation that an estimation follows a testing. We consider here the optimality of significance levels for the inference procedures based on the theory of minimax regret, and compare the optimal significance level with that based on AIC. Finally we propose a weighted estimate method for a contingency table, as a modification of the preliminary test estimation.

1. Introduction

The inference procedures proposed here carry out the selection of a suitable model and then estimate optimally parameters in the model, see Asano [2] and Kitagawa [5]. That is, a parsimonious model is at first selected in view of fitting to observed frequencies, then the parameters are estimated, depending on the selected model. The illustration for simplicity in model-building has been given in Bishop, Fienberg and Holland [3]. The overall variability of the estimates in the simpler model is smaller than that in the model with more parameters, see Altham [1]. We discuss here inference procedures for the symmetry in contingency tables. In a situation that a contingency table may be symmetric, we investigate it firstly, then may pool the frequencies if the symmetry is accepted. In this manner, we may avoid in some extent the risk of pooling asymmetrical frequencies in the table and do not lose so much profit by such a simplicity in model-building.

In section 2, two inference procedures are proposed for the partial symmetry and the complete symmetry in a contingency table. Section 3 gives the expected values and the mean square errors of the estimates based on the inference procedures. In section 4, the optimality of significance levels is discussed in detail, basing on the theory of the minimax regret. Finally, a weighted estimate procedure is proposed as a modified preliminary test estimation.

* Graduate school of Engineering Sciences, Kyushu University 39, Kasuga 816, Japan.

2. Inference Procedures for Symmetry in Contingency Tables

In an $I\times J$ two-dimensional contingency table, let p_{ij} and n_{ij} be the probability and the observed frequency in the cell (i, j), respectively. Assume that the data follow the multinomial distribution with the sample size $N=\sum n_{ij}$. We consider a partial symmetry and a complete symmetry in the contingency table as below.

2.1 Partial symmetry in the contingency table

The partial symmetry is defined as $p_{ij}=p_{ji}$ for two cells (i, j) and (j, i), $i\neq j$. The null hypothesis H_0 for the partial symmetry is that $p_{ij}=p_{ji}$, and the alternative H_1 is that $p_{ij}\neq p_{ji}$.

Because a cell (i, j) and the corresponding cell (j, i) are only considered, we can test this hypothesis H_0 by the Fisher exact test, where H_0 is equivalent to a hypothesis $p_{ij}/(p_{ij}+p_{ji})=1/2$.

The maximum likelihood estimates are $\hat{p}_{ij}^{(0)}=(n_{ij}+n_{ji})/(2N)$ and $\hat{p}_{ij}^{(1)}=n_{ij}/N$ under H_0 and H_1, respectively. The estimate $\tilde{p}_{ij}$ based on the inference procedure is equal to $\hat{p}_{ij}^{(0)}$, if $\min(n_{ij}, n_{ji})>L$, otherwise $\tilde{p}_{ij}=\hat{p}_{ij}^{(1)}$, where $L=\max\{k: \sum_{i=0}^{k}\binom{m}{i}\leqq 2^{m-1}\alpha\}$, $m=n_{ij}+n_{ji}$ and α is a significance level.

2.2 Complete symmetry in the contingency table

The complete symmetry in an $I\times I$ contingency table is defined as $p_{ij}=p_{ji}$ for $1\leqq i<j\leqq I$. We test the hypothesis H_0: $p_{ij}=p_{ji}$ for all $i<j$ against the alternative H_1: not H_0 by using McNemar's test statistic

$$\chi^2=\sum_{i<j}(n_{ij}-n_{ji})^2/(n_{ij}+n_{ji}),$$

which follows a chi-square distribution with $I(I-1)/2$ degrees of freedom under H_0. The inference procedure for the complete symmetry is similar to that in section 2.1. The estimate $\tilde{p}_{ij}$ based on the procedure is equal to $\hat{p}_{ij}^{(0)}$ if $\chi^2\leqq\chi^2_{I(I-1)/2}(\alpha)$, otherwise $\tilde{p}_{ij}=\hat{p}_{ij}^{(1)}$ for all $i\neq j$, where $\chi^2_f(\alpha)$ is the upper $100\alpha\%$ point of the chi-square distribution with f degrees of freedom.

3. Expected Values and Mean Square Errors of Estimates

The estimates of parameters in the contingency table depend on the methods of testing hypothesis and the significance level in the inference procedures. Therefore the expected values and the mean square errors (MSEs) of these estimates are given as below, for both procedures using the exact test and using McNemar's test, respectively.

3.1 For the partial symmetry

The estimate $\tilde{p}_{ij}$ proposed in section 2.1 can be expressed for the partial symmetry as follows,

$$\tilde{p}_{ij}=\begin{cases}(n_{ij}+n_{ji})/(2N), & \text{if } \min(n_{ij},\ n_{ji})>L\,, \\ n_{ij}/N, & \text{otherwise}.\end{cases} \tag{3.1}$$

According to the procedure, the expected value of the estimate of p_{ij} is

$$E(\tilde{p}_{ij})=\sum_{\Omega_0} p(\boldsymbol{n})(n_{ij}+n_{ji})/(2N)+\sum_{\Omega_0} p(\boldsymbol{n})n_{ij}/N, \tag{3.2}$$

where $\Omega_0=\{\boldsymbol{n}:\text{ where } \boldsymbol{n}=[n_{11},\ n_{12},\ \cdots,\ n_{1I},\ n_{21},\ \cdots,\ n_{II}] \text{ and } \min(n_{ij},\ n_{ji})>L\}$, $\Omega_1=\{\boldsymbol{n}:\text{ where } \min(n_{ij},\ n_{ji})\leqq L\}$, and $p(\boldsymbol{n})$ is the probability of n_{ij} and n_{ji} occurring,

$$p(\boldsymbol{n})=Cp_{ij}^{n_{ij}}p_{ji}^{n_{ji}}(1-p_{ij}-p_{ji})^{(N-n_{ij}-n_{ji})},$$

where $C=N!/[(N-n_{ij}-n_{ji})!\ n_{ij}!\ n_{ji}!]$.

From (3.2), we have

$$\begin{aligned}E(\tilde{p}_{ij})=\frac{1}{2}\Big\{p_{ij}+p_{ji}+(n_{..}-1)!\sum_{k=0}^{N}\Big[&\frac{(1-p_{ij}-p_{ji})^{N-k}}{(N-k)!} \\ &\cdot\sum_{x=0}^{g(k)}\frac{k-2x}{x!(k-x)!}(p_{ij}p_{ji})^x(p_{ij}^{k-2x}-p_{ji}^{k-2x})\Big]\Big\},\end{aligned} \tag{3.3}$$

where $g(k)=\max\left\{x:\ \sum_{i=0}^{x}\binom{k}{i}\leqq 2^{k-1}\alpha\right\}$.

Similarly, the MSE of the estimate $\tilde{p}_{ij}$ in (3.1) can be shown as

$$\begin{aligned}\mathrm{MSE}(\tilde{p}_{ij})=\mathrm{MSE}\Big(\frac{n_{ij}+n_{ji}}{2N}\Big)+N!\sum_{k=0}^{N}\Big\{&\frac{(1-p_{ij}-p_{ji})^{N-k}}{(N-k)!}\sum_{x=0}^{g(k)}\frac{(p_{ij}p_{ji})^x}{x!(k-x)!} \\ &\cdot[(k-2x)p_{ij}(p_{ji}^{k-2x}-p_{ji}^{k-2x})/N-k^2(p_{ji}^{k-2x}+p_{ij}^{k-2x})/(4N^2) \\ &+x^2p_{ji}^{k-2x}/N^2+(k-x)^2p_{ij}^{k-2x}/N^2]\Big\},\end{aligned} \tag{3.4}$$

where $\mathrm{MSE}[(n_{ij}+n_{ji})/(2N)]=(p_{ij}+p_{ji})(1-p_{ij}-p_{ji})/(4N)+(p_{ij}-p_{ji})^2/4$.

If $p_{ij}=p_{ji}$, we have $E(\tilde{p}_{ij})=p_{ij}$, i.e. $\tilde{p}_{ij}$ is an unbiased estimate of p_{ij}. Moreover, when $p_{ij}=p_{ji}$, the MSE in (3.4) can be written as

$$\begin{aligned}\mathrm{MSE}(\tilde{p}_{ij})&=\mathrm{MSE}\Big(\frac{n_{ij}+n_{ji}}{2N}\Big)+N!\sum_{k=0}^{N}\Big[\frac{(1-2p_{ij})^{N-k}}{(N-k)!}\sum_{x=0}^{g(k)}\frac{p_{ij}^k(k-2x)^2}{x!(k-x)!(2N^2)}\Big] \\ &\geqq\mathrm{MSE}[(n_{ij}+n_{ji})/(2N)].\end{aligned}$$

That is, the MSE of $\tilde{p}_{ij}$ is not less than the MSE of $\hat{p}_{ij}^{(0)}$ when $p_{ij}=p_{ji}$. The MSE of $\hat{p}_{ij}^{(1)}$ is equal to $p_{ij}(1-p_{ij})/N$. Also we can show in the similar way that the MSE of $\tilde{p}_{ij}$ is not larger than the MSE of $\hat{p}_{ij}^{(1)}$ when $p_{ij}=p_{ji}$.

As an illustration, the numerical evaluation is shown in Table 1, where the sample size is 20 and the significance level is 17% which is the optimal significance level given in section 4. In the table, MEAN, VAR and MSE show expected values, variances and mean square errors of estimate $\tilde{p}_{ij}$, respectively. MEANk, VARk and MSEk show those of $\hat{p}_{ij}^{(k)}$. The efficiency EFFIC of the estimate $\tilde{p}_{ij}$ to the estimate $\hat{p}_{ij}^{(1)}$ is defined as MSE1/MSE.

Table 1. TE procedure of a contingency table for the partial symmetry ($N=20$, $\alpha=0.17$)

PIJ	PJI	MEAN	VAR	MSE	MEAN0	VAR0	MSE0	MEAN1	VAR1	MSE1	EFFIC
0.1	0.1	0.1000	0.0027	0.0027	0.1000	0.0020	0.0020	0.1000	0.0045	0.0045	1.6703
0.1	0.2	0.1288	0.0043	0.0052	0.1500	0.0026	0.0051	0.1000	0.0045	0.0045	0.8688
0.1	0.3	0.1386	0.0073	0.0088	0.2000	0.0030	0.0130	0.1000	0.0045	0.0045	0.5125
0.1	0.4	0.1324	0.0093	0.0104	0.2500	0.0031	0.0256	0.1000	0.0045	0.0045	0.4346
0.1	0.5	0.1204	0.0093	0.0097	0.3000	0.0030	0.0430	0.1000	0.0045	0.0045	0.4623
0.1	0.6	0.1107	0.0080	0.0081	0.3500	0.0026	0.0651	0.1000	0.0045	0.0045	0.5534
0.1	0.7	0.1045	0.0063	0.0064	0.4000	0.0020	0.0920	0.1000	0.0045	0.0045	0.7067
0.1	0.8	0.1012	0.0051	0.0051	0.4500	0.0011	0.1236	0.1000	0.0045	0.0045	0.8755
0.1	0.9	0.1003	0.0047	0.0047	0.5000	0.0000	0.1600	0.1000	0.0045	0.0045	0.9536
0.2	0.1	0.1712	0.0065	0.0073	0.1500	0.0026	0.0051	0.2000	0.0080	0.0080	1.0930
0.2	0.2	0.2000	0.0048	0.0048	0.2000	0.0030	0.0030	0.2000	0.0080	0.0080	1.6692
0.2	0.3	0.2278	0.0063	0.0071	0.2500	0.0031	0.0056	0.2000	0.0080	0.0080	1.1260
0.3	0.2	0.2722	0.0079	0.0087	0.2500	0.0031	0.0056	0.3000	0.0105	0.0105	1.2046
0.3	0.3	0.3000	0.0061	0.0061	0.3000	0.0030	0.0030	0.3000	0.0105	0.0105	1.7292
0.3	0.4	0.3269	0.0075	0.0082	0.3500	0.0026	0.0051	0.3000	0.0105	0.0105	1.2837
0.4	0.3	0.3731	0.0084	0.0091	0.3500	0.0026	0.0051	0.4000	0.0120	0.0120	1.3199
0.4	0.4	0.4000	0.0065	0.0065	0.4000	0.0020	0.0020	0.4000	0.0120	0.0120	1.8459
0.4	0.5	0.4236	0.0081	0.0086	0.4500	0.0011	0.0036	0.4000	0.0120	0.0120	1.3937
0.5	0.4	0.4764	0.0085	0.0091	0.4500	0.0011	0.0036	0.5000	0.0125	0.0125	1.3751
0.5	0.5	0.5000	0.0060	0.0060	0.5000	0.0	0.0	0.5000	0.0125	0.0125	2.0939

3.2 For the complete symmetry

The estimates $\tilde{p}_{ij}$'s for the complete symmetry can be similarly expressed as in (3.1). But in this situation, we use McNemar's test rather than the exact test. Thus the regions Ω_0 and Ω_1 are changed into

$$\Omega_0=\{\boldsymbol{n}:\ \sum_{i<j}(n_{ij}-n_{ji})^2/(n_{ij}+n_{ji})<\chi^2_{I(I-1)/2}(\alpha)\}\,,$$

$$\Omega_1=\{\boldsymbol{n}:\ \sum_{i<j}(n_{ij}-n_{ji})^2/(n_{ij}+n_{ji})\geqq\chi^2_{I(I-1)/2}(\alpha)\}\,,$$

To evaluate the expected values and MSEs, we first introduce two variables

$$X\equiv\sum_{i\neq j}\frac{[n_{ij}-(n_{ij}+n_{ji})/2]^2}{(n_{ij}+n_{ji})/2}\,, \tag{3.5}$$

$$Y\equiv\sum_{i\neq j}\frac{[n_{ij}-(n_{ij}+n_{ji})/2]^2}{(p_{ij}+p_{ji})N/2}\,. \tag{3.6}$$

THEOREM 1. *Two random variables X and Y, defined by* (3.5) *and* (3.6), *have a same asymptotic distribution, when $E(X)<\infty$.*

PROOF. Let $\hat{p}_{ij}$ be the maximum likelihood estimate n_{ij}/N. Therefore, by the property of the MLEs, $\hat{p}_{ij}-p_{ij}=O(1/\sqrt{N})$. Using the Taylor's theorem, we get the expansion of X at p_{ij}'s,

$$X=\sum_{i<j}\frac{N(p_{ij}-p_{ji})^2}{(p_{ij}+p_{ji})}+\sum_{i<j}\frac{N(p_{ij}-p_{ji})(p_{ij}+3p_{ji})}{(p_{ij}+p_{ji})^2}(\hat{p}_{ij}-p_{ij})$$

$$+\sum_{i<j}\frac{N(p_{ji}-p_{ij})(p_{ji}+3p_{ij})}{(p_{ij}+p_{ji})^2}(\hat{p}_{ji}-p_{ji})+\sum_{i<j}\frac{4N}{(p_{ij}+p_{ji})^3}$$

$$\cdot[p_{ji}(\hat{p}_{ij}-p_{ij})-p_{ij}(\hat{p}_{ji}-p_{ji})]^2+O(1/\sqrt{N}).$$

As $N\to\infty$, the expected value $E(X)$ tends to infinite except that $(p_{ij}-p_{ji})^2<O(1/N)$.

Similarly, we expand Y and calculate the difference between X and Y under the condition $(p_{ij}-p_{ji})^2<O(1/N)$,

$$X-Y=\sum_{i<j}\frac{N(p_{ij}-p_{ji})^2}{(p_{ij}+p_{ji})^2}(p_{ij}-\hat{p}_{ij})+\sum_{i<j}\frac{N(p_{ij}-p_{ji})^2}{(p_{ij}+p_{ji})^2}(p_{ji}-\hat{p}_{ji})$$

$$+\sum_{i<j}\frac{N(p_{ij}-p_{ji})}{(p_{ij}+p_{ji})^3}[p_{ij}-\hat{p}_{ij}+p_{ji}-\hat{p}_{ji}]$$

$$\cdot[p_{ji}^2-p_{ij}^2+p_{ji}(3\hat{p}_{ij}-\hat{p}_{ji})-p_{ij}(3\hat{p}_{ji}-\hat{p}_{ij})]=O(1/\sqrt{N}),$$

that is, the difference tends to zero as $N\to\infty$. Therefore, the result follows directly from Rao [6]. □

THEOREM 2. *When $E(Y)<\infty$, the random variable Y, defined by* (3.6), *follows asymptotically the noncentral chi-square distribution with $I(I-1)/2$ degrees of freedom and the noncentrality parameter*

$$\delta=\sum_{i<j}N(p_{ij}-p_{ji})^2/(p_{ij}+p_{ji}). \tag{3.7}$$

PROOF. Let $p_{ij}=p_{ji}+\mu_{ij}$ and $y_{ij}=(n_{ij}-n_{ji})/[(p_{ij}+p_{ji})N]^{1/2}$. It is known that y_{ij} has the asymptotic normal distribution with the mean μ_{ij} and the variance σ_{ij}^2, where

$$\mu_{ij}=(p_{ij}-p_{ji})[N/(p_{ij}+p_{ji})]^{1/2},$$

$$\sigma_{ij}^2=1-(p_{ij}-p_{ji})^2/(p_{ij}+p_{ji}).$$

Since $(p_{ij}-p_{ji})^2=O(1/N)$ for $E(Y)=\sum[N(p_{ij}-p_{ji})^2/(p_{ij}+p_{ji})+1/2]<\infty$, we have $\sigma_{ij}^2\to 1$. Thus Y, the sum of y_{ij}^2 over all $i<j$, follows asymptotically the noncentral chi-square distribution with $I(I-1)/2$ degrees of freedom and the noncentrality parameter δ given in (3.7). □

THEOREM 3. *Let a J-dimensional vector W follow the multinormal distribution $N(\theta, I)$, a $J\times J$ matrix A be positive definite and $\Phi(X)$ be an arbitrary real function. Then*

$$E[\Phi(W'W)W]=\theta E[\Phi(\chi^2_{[J+2,\,\theta'\theta]})],$$

$$E[\Phi(W'W)W'AW]=E[\Phi(\chi^2_{[J+2,\,\theta'\theta]})]\operatorname{tr}A+E[\Phi(\chi^2_{[J+4,\,\theta'\theta]})]\theta'A\theta,$$

$$E\Big[\Phi\Big(\sum_{i=1}^{J}w_i^2\Big)w_j^2\Big]=E[\Phi)\chi^2_{[J+2,\,\theta'\theta]})]+\theta_j^2E[\Phi(\chi^2_{[J+4,\,\theta'\theta]})],$$

where $\chi^2_{[K,\delta]}$ denotes the noncentral chi-square distribution with K degrees of freedom and the noncentrality parameter δ.

PROOF. The theorem is based on Stein [9] and Sclove, et al. [8]. □

Suppose the following indicator function

$$I_{[a,b)}(X)=\begin{cases}1, & a\leqq X<b,\\ 0, & \text{otherwise}.\end{cases}$$

Since X follows $\chi^2_{[I(I-1)/2,\,\delta]}$ by theorems 1 and 2 where δ is given in (3.7), the expected values and MSEs of the estimates based on McNemar's test can be given respectively as

$$E(\tilde{p}_{ij})=E\left[I_{[0,\,\chi^2_\alpha)}(X)\frac{n_{ij}+n_{ji}}{2N}\right]+E\left[I_{[\chi^2_\alpha,\,\infty)}(X)\frac{n_{ij}}{N}\right]$$

$$=p_{ij}-E\left[I_{[0,\,\chi^2_\alpha)}(X)\frac{n_{ij}-n_{ji}}{2N}\right],$$

$$\mathrm{MSE}(\tilde{p}_{ij})+\mathrm{MSE}(\tilde{p}_{ji})=\{p_{ij}(1-p_{ij})+p_{ji}(1-p_{ji})\}/N$$

$$+E\{I_{[0,\,\chi^2_\alpha)}(X)[(p_{ij}-p_{ji})(n_{ij}-n_{ji})/N-(n_{ij}-n_{ji})^2/(2N^2)]\}.$$

From Theorem 3, we obtain

$$E(\tilde{p}_{ij})=p_{ij}-(p_{ij}-p_{ji})\Pr(\chi^2_{[f+2,\,\delta]}<\chi^2_\alpha)/2,$$

$$\mathrm{MSE}(\tilde{p}_{ij})+\mathrm{MSE}(\tilde{p}_{ji})=\{p_{ij}(1-p_{ij})+p_{ji}(1-p_{ji})\}/N$$

$$-[(p_{ij}+p_{ji})/(2N)-(p_{ij}-p_{ji})^2]\Pr(\chi^2_{[f+2,\,\delta]}\leqq\chi^2_\alpha)$$

$$-(p_{ij}-p_{ji})^2\Pr(\chi^2_{[f+4,\,\delta]}\leqq\chi^2_\alpha)/2,$$

where $f=I(I-1)/2$. The numerical evaluation are shown in Table 2 with the sample size 20 and the significance level 17%, where VIJJI shows the sum of variances of $\tilde{p}_{ij}$ and $\tilde{p}_{ji}$, MSEPIJ+JI the sum of MSEs, and EFFIJJI denotes $[\mathrm{MSE}(\hat{p}^{(1)}_{ij})+\mathrm{MSE}(\hat{p}^{(1)}_{ji})]/[\mathrm{MSE}(\tilde{p}_{ij})+\mathrm{MSE}(\tilde{p}_{ji})]$.

Table 2. TE procedure of a contingency table for the complete symmetry ($N=20$, $\alpha=0.17$)

PIJ	PJI	MEAN	VIJJI	MSEPIJ+JI	MEAN0	VAR0	MSE0	MEAN1	VAR1	MSE1	EFFIJJI
0.1	0.1	0.1000	0.0070	0.0070	0.1000	0.0020	0.0020	0.1000	0.0045	0.0045	1.2884
0.1	0.2	0.1161	0.0123	0.0128	0.1500	0.0026	0.0051	0.1000	0.0045	0.0045	0.9780
0.1	0.3	0.1205	0.0190	0.0199	0.2000	0.0030	0.0130	0.1000	0.0045	0.0045	0.7550
0.1	0.4	0.1176	0.0234	0.0240	0.2500	0.0031	0.0256	0.1000	0.0045	0.0045	0.6869
0.1	0.5	0.1126	0.0244	0.0247	0.3000	0.0030	0.0430	0.1000	0.0045	0.0045	0.6873
0.1	0.6	0.1082	0.0229	0.0230	0.3500	0.0026	0.0651	0.1000	0.0045	0.0045	0.7172
0.1	0.7	0.1049	0.0198	0.0198	0.4000	0.0020	0.0920	0.1000	0.0045	0.0045	0.7563
0.1	0.8	0.1028	0.0158	0.0158	0.4500	0.0011	0.1236	0.1000	0.0045	0.0045	0.7913
0.1	0.9	0.1016	0.0111	0.0111	0.5000	0.0000	0.1600	0.1000	0.0045	0.0045	0.8099
0.2	0.1	0.1839	0.0123	0.0128	0.1500	0.0026	0.0051	0.2000	0.0080	0.0080	0.9780
0.2	0.2	0.2000	0.0120	0.0120	0.2000	0.0030	0.0030	0.2000	0.0080	0.0080	1.3366
0.2	0.3	0.2176	0.0164	0.0170	0.2500	0.0031	0.0056	0.2000	0.0080	0.0080	1.0858
0.3	0.2	0.2824	0.0164	0.0170	0.2500	0.0031	0.0056	0.3000	0.0105	0.0105	1.0858
0.3	0.3	0.3000	0.0150	0.0150	0.3000	0.0030	0.0030	0.3000	0.0105	0.0105	1.4041
0.3	0.4	0.3183	0.0185	0.0191	0.3500	0.0026	0.0051	0.3000	0.0105	0.0105	1.1751
0.4	0.3	0.3817	0.0185	0.0191	0.3500	0.0026	0.0051	0.4000	0.0120	0.0120	1.1751
0.4	0.4	0.4000	0.0159	0.0159	0.4000	0.0020	0.0020	0.4000	0.0120	0.0120	1.5055
0.4	0.5	0.4187	0.0185	0.0192	0.4500	0.0011	0.0036	0.4000	0.0120	0.0120	1.2759
0.5	0.4	0.4813	0.0185	0.0192	0.4500	0.0011	0.0036	0.5000	0.0125	0.0125	1.2759
0.5	0.5	0.5000	0.0149	0.0149	0.5000	0.0	0.0	0.5000	0.0125	0.0125	1.6749

4. The Optimal Significance Level

Let the minimax regret be the criterion for the optimal significance level. The significance levels based on the criterion are compared with those based on AIC. The regret R is now defined as

$$R \equiv \sum_{i\neq j} \mathrm{MSE}(\tilde{p}_{ij}) - \min\left\{ \sum_{i\neq j} \mathrm{MSE}(\hat{p}_{ij}^{(0)}),\ \sum_{i\neq j} \mathrm{MSE}(\hat{p}_{ij}^{(1)}) \right\}.$$

Then the optimal significance level α_{opt} is determined such that the maximum regret $\max_p(R)$ is minimized.

The maximum difference in the sums of the MSEs between $\tilde{p}_{ij}$'s and $\hat{p}_{ij}^{(0)}$'s is

$$R_0 \equiv \max_p \left\{ \sum_{i\neq j} \mathrm{MSE}(\tilde{p}_{ij}) - \sum_{i\neq j} \mathrm{MSE}(\hat{p}_{ij}^{(0)}) \right\},$$

and that between $\tilde{p}_{ij}$'s and $\hat{p}_{ij}^{(1)}$'s is

$$R_1 \equiv \max_p \left\{ \sum_{i\neq j} \mathrm{MSE}(\tilde{p}_{ij}) - \sum_{i\neq j} \mathrm{MSE}(\hat{p}_{ij}^{(1)}) \right\}.$$

For the complete symmetry of $I\times I$ contingency tables, R_0 and R_1 can be easily obtained as follows,

$$R_0 = \max_p \sum_{i<j} \left\{ \left[\frac{p_{ij}+p_{ji}}{2N} - (p_{ij}-p_{ji})^2 \right] \Pr(\chi^2_{[f+2,\,\delta]} \geqq \chi^2_\alpha) \right.$$
$$\left. + (p_{ij}-p_{ji})^2 \Pr(\chi^2_{[f+4,\,\delta]} \geqq \chi^2_\alpha)/2 \right\},$$

$$R_1 = \max_p \sum_{i<j} \left\{ \left[(p_{ij}-p_{ji})^2 - \frac{p_{ij}+p_{ji}}{2N} \right] \Pr(\chi^2_{[f+2,\,\delta]} \leqq \chi^2_\alpha) \right.$$
$$\left. - (p_{ij}-p_{ji})^2 \Pr(\chi^2_{[f+4,\,\delta]} \leqq \chi^2_\alpha)/2 \right\},$$

where $f=I(I-1)/2$. With the significance level α decreasing, R_0 decreases but R_1 increases. The optimal significance level α_{opt} can be determined so that R_0 is equated to R_1.

Sakamoto, Isikuro and Kitagawa [7] proposed an AIC method for model selection for contingency tables. The difference between AIC values under H_1 and H_0 is given as

$$\mathrm{AIC}(H_1) - \mathrm{AIC}(H_0) = -\chi^2 + 2[I(I-1)/2],$$

where $\chi^2 = 2\sum_{i\neq j} n_{ij}\log[2n_{ij}/(n_{ij}+n_{ji})]$. Thus the significance level α_{AIC} based on AIC is determined so that $\chi^2_{I(I-1)/2}(\alpha) = I(I-1)$.

To compare α_{opt} with α_{AIC}, we consider the significance levels for 2×2 contingency tables in Table 3. Table 4 gives the corresponding regrets, and it is shown that α_{AIC} shows the larger maximum regret than that based on α_{opt}.

Table 3. α_{opt} and α_{AIC}

N	α_{opt}	α_{AIC}
10	0.171	0.157
20	0.171	0.157
50	0.171	0.157
100	0.171	0.157
500	0.171	0.157

Table 4. Maximum regrets based on α_{opt} and α_{AIC}

N	α_{opt}	α_{AIC}	
	R0=R1	R0	R1
10	0.0299	0.0286	0.0326
20	0.0150	0.0143	0.0163
50	0.0060	0.0057	0.0065
100	0.0030	0.0029	0.0032
500	0.0006	0.0006	0.0006

5. Weighted Estimate for Symmetry in Contingency Tables

The inference procedures proposed above may be formulized in the following way. Let $\hat{\theta}^{(0)}$ and $\hat{\theta}^{(1)}$ be the maximum likelihood estimates under H_0 and H_1, respectively. Suppose T is a test statistic and $\Phi(T)$ is a decision function defined as

$$\Phi(T)=\begin{cases}0, & \text{if } H_0 \text{ is accepted},\\ 1, & \text{if } H_0 \text{ is rejected}.\end{cases}$$

Then the estimate $\tilde{\theta}$ based on the inference procedures can be written as

$$\tilde{\theta}=\Phi(T)\hat{\theta}^{(1)}+[1-\Phi(T)]\hat{\theta}^{(0)} \tag{5.1}$$

For the normal distribution, Huntsberger [4] tried to generalize $\Phi(T)$ to a real function and gave a weighted estimate method. According to this weighted estimate method, let us investigate estimates of probabilities in contingency tables for the symmetry.

In an $I\times I$ contingency table, the weighted estimate is

$$\tilde{p}_{ij}=\Phi n_{ij}/N+(1-\Phi)(n_{ij}+n_{ji})/(2N), \qquad \text{for all } i\neq j, \tag{5.2}$$

where Φ is a parameter to be determined so that $\text{MSE}(\tilde{p}_{ij})$ is minimized. Since

$$\text{MSE}(\tilde{p}_{ij})=(\Phi+1)^2[(N-1)p_{ij}^2+p_{ij}]/(4N)+(\Phi-1)^2[(N-1)p_{ji}^2+p_{ji}]$$
$$/(4N)-p_{ij}^2\Phi+[(1-\Phi^2)(N-1)/(2N)+\Phi-1]p_{ij}p_{ji},$$

taking the partial derivative of $\text{MSE}(\tilde{p}_{ij})$ with respect to Φ, we have

$$\frac{\partial\text{MSE}(\tilde{p}_{ij})}{\partial\Phi}=p_{ij}(\Phi+1)[(N-1)p_{ij}+1]/(2N)-p_{ji}(1-\Phi)$$
$$\cdot[(N-1)p_{ji}+1]/(2N)-p_{ij}^2+p_{ij}p_{ji}[1-(N-1)\Phi/N].$$

Finally setting it to zero, we obtain that

$$\Phi=1-2p_{ij}(1+p_{ji}-p_{ij})/[(N-1)(p_{ij}-p_{ji})^2+p_{ij}+p_{ji}]. \tag{5.3}$$

Since Φ is a function of unknown parameters p_{ij} and p_{ji}, substituting n_{st}/N for p_{st}, we obtain

$$\hat{\Phi}=1-2n_{ij}(N+n_{ji}-n_{ij})/[(N-1)(n_{ij}-n_{ji})^2+N(n_{ij}+n_{ji})].$$

Replacing Φ in (5.2) with the above estimate $\hat{\Phi}$, we obtain

$$\tilde{p}_{ij}=\frac{n_{ij}}{N}+\frac{n_{ij}[1-(n_{ij}-n_{ji})/N](n_{ji}-n_{ij})}{(N-1)(n_{ij}-n_{ji})^2+N(n_{ij}+n_{ji})}$$

The expected value and MSE of the estimate $\tilde{p}_{ij}$ are obtained in the exact formulas

$$E(\tilde{p}_{ij})=p_{ij}+\sum_{k=0}^{N}\left\{\frac{N!}{(N-k)!}(1-p_{ij}-p_{ji})^{N-k}\left[\sum_{x=0}^{k}\frac{p_{ij}^{x}p_{ji}^{k-x}}{x!(k-x)!}\right.\right.$$
$$\left.\left.\cdot\frac{x[1-(2x-k)/N](k-2x)}{(N-1)(2x-k)^2+kN}\right]\right\},$$

$$\mathrm{MSE}(\tilde{p}_{ij})=p_{ij}(1-p_{ij})/N+2\sum_{k=0}^{N}\left\{\frac{N!}{(N-k)!}(1-p_{ij}-p_{ji})^{N-k}\right.$$
$$\cdot\sum_{x=0}^{k}\frac{x[1-(2x-k)/N](k-2x)p_{ij}^{x}p_{ji}^{k-x}}{[(N-1)(2x-k)^2+kN]x!(k-x)!}$$
$$\left.\cdot\left[\frac{x}{N}-p_{ij}+\frac{x[1-(2x-k)/N](k-2x)}{2[(N-1)(2x-k)^2+kN]}\right]\right\}.$$

In Table 5, we give the numerical evaluation for the weighted estimates with the sample size 20.

Table 5. Expected values and MSEs for weighted estimates

PIJ	PJI	MEAN	VAR	MSE	MEAN0	VAR0	MSE0	MEAN1	VAR1	MSE1	EFFIC
0.1	0.1	0.0926	0.0033	0.0034	0.1000	0.0020	0.0020	0.1000	0.0045	0.0045	1.3361
0.1	0.2	0.1066	0.0041	0.0041	0.1500	0.0026	0.0051	0.1000	0.0045	0.0045	1.0889
0.1	0.3	0.1134	0.0050	0.0052	0.2000	0.0030	0.0130	0.1000	0.0045	0.0045	0.8732
0.1	0.4	0.1155	0.0056	0.0058	0.2500	0.0031	0.0256	0.1000	0.0045	0.0045	0.7724
0.1	0.5	0.1153	0.0059	0.0061	0.3000	0.0030	0.0430	0.1000	0.0045	0.0045	0.7364
0.1	0.6	0.1144	0.0059	0.0062	0.3500	0.0026	0.0651	0.1000	0.0045	0.0045	0.7317
0.1	0.7	0.1133	0.0059	0.0061	0.4000	0.0020	0.0920	0.1000	0.0045	0.0045	0.7406
0.1	0.8	0.1124	0.0058	0.0060	0.4500	0.0011	0.1236	0.1000	0.0045	0.0045	0.7542
0.1	0.9	0.1116	0.0057	0.0059	0.5000	0.0000	0.1600	0.1000	0.0045	0.0045	0.7681
0.2	0.1	0.1775	0.0061	0.0066	0.1500	0.0026	0.0051	0.2000	0.0080	0.0080	1.2122
0.2	0.2	0.1946	0.0056	0.0056	0.2000	0.0030	0.0030	0.2000	0.0080	0.0080	1.4224
0.2	0.3	0.2099	0.0063	0.0064	0.2500	0.0031	0.0056	0.2000	0.0080	0.0080	1.2540
0.3	0.2	0.2797	0.0076	0.0080	0.2500	0.0031	0.0056	0.3000	0.0105	0.0105	1.3131
0.3	0.3	0.2964	0.0068	0.0069	0.3000	0.0030	0.0030	0.3000	0.0105	0.0105	1.5317
0.3	0.4	0.3123	0.0073	0.0075	0.3500	0.0026	0.0051	0.3000	0.0105	0.0105	1.4053
0.4	0.3	0.3817	0.0081	0.0084	0.3500	0.0026	0.0051	0.4000	0.0120	0.0120	1.4279
0.4	0.4	0.3982	0.0071	0.0071	0.4000	0.0020	0.0020	0.4000	0.0120	0.0120	1.6992
0.4	0.5	0.4145	0.0073	0.0075	0.4500	0.0011	0.0036	0.4000	0.0120	0.0120	1.6014
0.5	0.4	0.4836	0.0075	0.0078	0.4500	0.0011	0.0036	0.5000	0.0125	0.0125	1.6032
0.5	0.5	0.5000	0.0062	0.0062	0.5000	0.0	0.0	0.5000	0.0125	0.0125	2.0018

To compare these methods each other, the maximum regrets are given for 2×2 contingency tables with several sample sizes in Table 6.

Table 6. The maximum regrets

N	α_{opt}	α_{AIC}		weighted esti.	
	R0=R1	R0	R1	R0	R1
10	0.0299	0.0286	0.0326	0.0268	0.0124
20	0.0150	0.0143	0.0163	0.0124	0.0064
50	0.0060	0.0057	0.0065	0.0048	0.0026

The weighted estimate method brings about the least maximum regret, although it gives no information on the model.

6. Conclusion

We have proposed several inference procedures for the symmetry in contingency tables, and have discussed the optimal significance level. Moreover, a weighted estimate method has been suggested which generalizes the decision function to a real function. In this paper, the maximum likelihood estimates are only considered as the basic estimates, i.e. $\hat{p}_{ij}^{(0)}$'s and $\hat{p}_{ij}^{(1)}$'s are used as the basic estimates under H_0 and H_1, respectively. Also, we may use other approaches with some advantages to find the basic estimates, e.g., the kernel smoothing technique and Bayesian methods may be preferable in sparse tables. Moreover, it may be possible to apply these procedures to the analyses of independence and loglinear models for contingency tables.

Acknowledgements

The author would like to express his thanks to Professor Ch. Asano for his valuable comments and suggestions.

References

[1] ALTHAM, P.M.E.: *Improving the Precision of Estimation by Fitting a Model*, J. Royal Statist. Soc., **B46** (1984), 118-119.
[2] ASANO, Ch.: *Estimations after preliminary tests of significance and their applications to biometrical researches*, Bull. Math. Statist., **9** (1960), 1-23.
[3] BISHOP, Y., FIENBERG, S.E. and HOLLAND, P.W.: *Discrete Multivariate Analysis*, MIT Press, Cambridge, (1975).
[4] HUNTSBERGER, D.V.: *A Generalization of a Preliminary testing procedure for pooling data*, Ann. Math. Statist., **26** (1955), 734-743.
[5] KITAGAWA, T.: *Successive Process of Statistical Inferences* (in Japanese), Kyoritsu Shuppan, (1984).
[6] RAO, C.R.: *Linear Statistical Inference and Its Applications*, John Wiley, New York, (1965).
[7] SAKAMOTO, Y., ISHIGURO, M. and KITAGAWA, G.: *Statistics Based on Amount of Information* (in Japanese), Kyoritsu Shuppan, (1982).
[8] SCLOVE, S.L., MORRIS, C. and RADHAKRISHNAN, R.: *Nonoptimality of Preliminary-Test Estimators for the Mean of a Multivariate Normal Distribution*, Ann. Math. Statist., **43** (1972), 1481-1490.

[9] STEIN, C.: *An Approach to the Recovery of Interblock Informations in Balance Incomplete Block Designs*, Research Papers in Statistics: Festschrift for J. Neyman (DAVID, F.N. ed.) Wiley, New York (1966), 351-366.

Received October 2, 1987
Communicated by Ch. Asano