On Properties of Estimators of Testing Homogeneity in r×2 Contingency Tables of Small Sample Size

H. Yassaee, Tehran University of Technology, Iran, and
University of North Carolina at Chapel Hill

Abstract

The problem of testing homogeneity in r×2 contingency tables can be converted to the problem of testing homogeneity of r independent binomial distributions. To estimate the common probability p, of success, we use likelihood function, statistics X^2_M, X^2_p, $I(p; \hat{p})$, and X^2_L, the first of which is to be maximized with respect to p and the rest to be minimized with respect to p. In this paper we study the bias and variance of estimators through means of analytical and numerical procedures based on approximate rules and computer-obtained results. We compare the value of bias and variance for each method and give a table which presents their order of magnitude.

Key words: contingency tables; maximum likelihood estimate; minimum modified X^2_N estimate; minimum X^2_p estimate; minimum discrimination information estimate; minimum logit X^2 estimate; bias; mean square error.

1. Introduction

Cochran (1952, p.325) has shown that the problem of testing homogeneity in r×2 contingency tables with cell frequencies x_{ij}'s, i=1, 2,...,r; j=1,2,...,c may be converted to the problem stated as follows: Given r independent binomial distributions from which samples of sizes $x_{1.}$, $x_{2.}$,...,$x_{r.}$ are drawn, respectively, one is interested in testing the hypothesis that these populations are homogeneous; i.e., $p_i=p$, i=1, 2,...,r which is the probability of "success". We denote $x_{i.} = \sum_{j=1}^{2} x_{ij}$ by n_i and x_{i1} by x_i throughout the paper. To estimate p, we use the likelihood function, modified X_M^2, Pearson X_P^2, information statistic $I(p:\hat{p})$, and logit X_L^2, the first of which is to be maximized with respect to p and the rest to be minimized with respect to p. These estimators are asymptotically equivalent. However, they differ for small sample sizes. Yassaee (1975, a, b, c, d) has studied properties of estimators and distribution of various statistics in r×c and r×c×t contingency tables of small sample sizes. Some other researchers have studied the comparison of estimators and statistics in various models using different approaches. For example, see Harvey (1977), Brown and Muentz (1976), Pradhan and Sathe (1976), Chapman (1976), Hommel (1978). In this paper we study analytically the bias and mean square error of estimators under consideration, and the criteria, such as biasedness and mean square error are taken into account for comparing estimators. Numerical investigations are restricted to 2×2 contingency tables.

2. Methods of Estimation

In this section we give estimators of the parameter p, obtained by the different procedures mentioned in section 1.

The unique maximum likelihood estimate $\hat{p}$ for p is given by

$$\hat{p} = \frac{\sum_{i=1}^{r} x_i}{N}, \quad N = \sum_{i=1}^{r} n_i \tag{2.1}$$

The unique minimum modified X_M^2 (Neyman X^2) estimator for p is given

$$p_M = \frac{\sum_{i=1}^{r} \hat{W}_i \hat{p}_i}{\hat{W}}, \quad \hat{W}_i = \frac{n_i}{\hat{p}_i \hat{q}_i}, \quad \hat{W} = \sum_{i=1}^{r} \hat{W}_i \tag{2.2}$$

where $\hat{p}_i = 1-\hat{q}_i = \frac{x_i}{n_i}$. We use the normal approximation to the binomial distribution to investigate properties of p_M.

The minimum Pearson X_p^2 estimate for p is obtained by solving the equation

$$\left\{ \sum_{i=1}^{r} n_i (\hat{q}_i - \hat{p}_i) \right\} p^2 + 2\left\{ \sum_{i=1}^{r} \hat{p}^2{}_i \, n_i p - \sum_{i=1}^{r} n_i \hat{p}_i{}^2 \right\} = 0 \tag{2.3}$$

The only root which is between 0 and 1 and minimizes X_p^2 is given by

$$p_p = \frac{-\sum_{i=1}^{r} n_i \hat{p}^2{}_i + \left\{ \left[\sum_{i=1}^{r} n_i \hat{p}^2{}_i \right] \left[\sum_{i=1}^{r} n_i \hat{q}_i{}^2 \right] \right\}^{\frac{1}{2}}}{\sum_{i=1}^{r} n_i \, (\hat{q}_i - \hat{p}_i)} \tag{2.4}$$

Should it happen that the coefficient of p^2 is zero, then we take $p_p = \frac{1}{2}$.

Hence there is only one root between 0 and 1.

Let $n_i(\hat{q}_i - \hat{p}_i) = \hat{y}_i$, $\hat{y} = \sum_{i=1}^{r} \hat{y}_i$.

Then one deduces the following inequalities

$$\frac{\sum_{i=1}^{r} \hat{y}_i \hat{p}_i}{\hat{y}} \leq p_p \leq \frac{1}{2} \qquad \text{if } \hat{y} > 0$$

$$p_p = \frac{1}{2} \qquad \text{if } \hat{y} = 0$$

$$\frac{1}{2} < p_p < \frac{\sum_{i=1}^{r} \hat{y}_i \hat{p}_i}{\hat{y}} \qquad \text{if } \hat{y} < 0 \ .$$

These inequalities are helpful to check the computed p_p values. The minimum discrimination information estimate p for which information statistic

$$I(p:\hat{P}) = \sum_{i=1}^{r} n_i p \ \ell n \frac{n_i p}{x_i} + \sum_{i=1}^{r} n_i q \ \ell n \frac{n_i q}{n_i - x_i} \ , \quad q = 1-p,$$

is a minimum, is given by

$$p^* = \frac{a}{1+a} \tag{2.5}$$

where

$$\ell n \ a = \frac{1}{N} \left(\sum_{i=1}^{r} n_i \right) \text{logit } \hat{p}_i,$$

and

$$N = \sum_{i=1}^{r} n_i \ .$$

Since $I(p:\hat{p})$ is a convex function of p, it is a minimum at p*, see Kullback (1959).

We define

$$\text{logit } X_L^2 = \sum_{i=1}^{r} \hat{w}_i \ (L_i - \hat{L}_i)^2,$$

where

$$L_i = \text{logit } p = \log \frac{p}{q}, \ \hat{L}_i = \text{logit } \hat{p},$$

$$\hat{w}_i = n_i \hat{p}_i \hat{q}_i = \frac{x_i (n_i - x_i)}{n_i}$$

the minimum logit X^2 estimator P_L is given by

$$\text{logit } p_L = \frac{\sum_{i=1}^{r} \hat{w}_i \text{ logit } \hat{p}_i}{\sum_{i=1}^{r} \hat{w}_i}$$

$$p_L = \frac{A}{1+A} \quad . \tag{2.6}$$

In form, P_L is the same as p^* given in (2.5) with $\hat{w}_i$ and $\sum_{i=1}^{r} \hat{w}_i$ substituted for n_i and N, respectively.

3. On the Bias and Variance of Estimates

It is well known that $\hat{p}$ is an unbiased estimate of p for all values of p. In order to compute different estimates of p simultaneously, we replaced 0 by $\frac{1}{2}$ in the analysis of 2×2 contingency tables. For $p = \frac{1}{2}$, $E(\hat{p}) = \frac{1}{2}$, exactly. For fixed N, $\hat{p}$ is unbiased only at $p = \frac{1}{2}$ in the range of p, $0 < p < 1$. Since $\frac{db(p)}{dp}$ is not zero, in general

$$\sigma^2(\hat{p}|p = \tfrac{1}{2}) \geq \frac{[1 + (\frac{db}{dp})|p = \frac{1}{2}]^2}{4N} \tag{3.1}$$

For r=2 we can write p_M given in (2.2) in the form

$$p_M = \frac{n_1 n_2 [x_1^2 x_2 + x_2^2 x_1 - N x_1 x_2]}{n_1^2 x_2^2 + n_2^2 x_1^2 - n_1^2 n_2 x_2 - n_2^2 n_1 x_1} \tag{3.2}$$

which is a function of two independent random variables x_1 and x_2 and is a rational function of independent random variables. By the use of Laplace, Mellin, or characteristic function, one can obtain the probability density function and the moments of such a function. But the algebra is somewhat lengthy, and we will not consider the problem as such here. We refer the

reader to Prasad (1970, pp 614-625).

We approximate the binomial distribution with probability of "success" p by a normal distribution with mean and variance the same as that of the binomial distribution. Numerically, if $\frac{1}{2}$ is not added to x_i when it is zero, or x_i is taken as $n-\frac{1}{2}$ when $x_i=n_i$, and the normality assumption is valid, one should have

$$E(p_M) = p ,$$

on the condition that $W_i = \frac{n_i}{pq}$ is used for $\hat{W}_i = \frac{n_i}{\hat{p}_i \hat{q}_i}$.

Let

$$\hat{y}_i = \frac{\hat{W}_i}{W} , \quad y_i = \frac{W_i}{W}$$

where

$$W_i = \frac{n_i}{pq} , \quad \hat{W}_i = \frac{n_i}{\hat{p}_i \hat{q}_i} .$$

Then

$$p_M = \sum_{i=1} \hat{y}_i \hat{p}_i .$$

Weights $\hat{y}_i$'s produce an asymptotically efficient estimate of p_M when the n_i's are sufficiently large. Under the assumption of normality, $\hat{y}_i$ and $\hat{p}_i$ are independent random variables. Therefore,

$$E[p_M] = E\{[p_M|\hat{y}_i]\} - E(\hat{p}_i) = p.$$

Let $Z_i = \frac{\hat{W}_i}{W_i}$.

Then

$$P_M = \frac{\sum_{i=1}^{r} w_i \frac{p_i}{Z_i}}{\sum_{i=1}^{r} \frac{w_i}{Z_i}}$$

$$\sigma^2_{p_M} = E_{Z_i}\left\{ \sigma^2_{p_M} | Z_i \right\}$$

Now

$$\sigma^2(P_M|Z_i) = \frac{\sum_{i=1}^{r}\left\{\frac{W_i}{Z^2_i}\right\}}{\left[\sum_{i=1}^{r}\frac{W_i}{Z_i}\right]^2}$$

Hence,

$$\sigma^2_{p_M} = E_{Z_i}\left\{\frac{\sum_{i=1}^{r}\frac{W_i}{Z^2_i}}{\left[\sum_{i=1}^{r}\frac{W_i}{Z_i}\right]^2}\right\}$$

To obtain this expectation, we applied a theorem on expectation of a function of random variables as stated in Welch (1938, pp 330-362). To get more details on what follows, one may write to the author for a long preliminary report.

One may estimate $\hat{\sigma}^2_{p_M}$ by $\frac{1}{W}\left[1+2\sum_{i=1}^{r}\frac{1}{n_i}\hat{p}_i\hat{q}_i\right]$.

Application of theorem mentioned earlier gives

$$E\left(\frac{1}{\hat{W}}\right) = \frac{1}{W}\left\{1-2\sum_{i=1}^{r}\frac{1}{n_i}p_iq_i + O\left(\sum_{i=1}^{r}\frac{1}{n_i}\right)\right\}$$

$$= \frac{1}{W}\left\{1-2\left(\sum_{i=1}^{r}\frac{1}{n_i}\right)pq + O\left(\sum_{i=1}^{r}\frac{1}{n^2_i}\right)\right\}.$$

One can see that this estimate has asymptotically a negative bias which is approximately equal to first term $\hat{\sigma}^2_{p_M}$ neglected. If one wants to obtain an estimate with bias of order $O\left(\sum_{i=1}^{r}\frac{1}{n^2_i}\right)$, then $\sigma^2_{p_M}$ should be estimated by

$$\hat{\sigma}^2_{p_M} = \frac{1}{W}\left\{1+4\sum_{i=1}^{r}\frac{1}{n_i}\hat{p}_i\hat{q}_i\right\}.$$

The exact variance of P_M under normal distribution is given by

$$\sigma^2_{P_M} = \frac{\frac{W_1}{Z^2_1} + \frac{W_2}{Z^2_2}}{\left(\frac{W_1}{Z_1} + \frac{W_2}{Z_2}\right)^2} = \frac{1}{W}\left\{1+W_1W_2\left[\frac{\left(\frac{Z_1}{Z_2} - 1\right)^2}{\left(W_1+W_2\frac{Z_1}{Z_2}\right)^2}\right]\right\} .$$

But $\frac{Z_1}{Z_2}$ has an F-distribution with (n_1-1, n_2-1) degrees of freedom. Consequently, we get the expectation $\sigma^2_{(P_M|Z_1)}$ with respect to the F-distribution.

Finally

$$\sigma^2_{P_M} = \frac{1}{W} + K$$

where

$$K = \frac{1}{W} E_F\left\{W_1W_2 \frac{(F-1)^2}{(W_1+W_2F)^2}\right\}$$

For the case $n_1 = n_2 = n$, we have

$$E\left\{\sigma^2_{P_M}\right\} = \frac{1}{W}\left\{1 + \frac{1}{n+1} - \frac{2}{(n+1)(n+3)}\right\} .$$

One may refer to (2.4) and expand P_p in Taylor series about values $p_i=p$ to find approximate bias term for p_p or get the variance of p_p. Due to the length of the algebra we will not derive the bias and variance here.

We expand logit $\hat{p}_i = \hat{L}_i$ in a Taylor series about the true value p, to get

$$\hat{L}_i = L + \frac{\hat{p}_i-p_i}{pq} + \frac{1}{2}\frac{(p-q)(\hat{p}_i-p)^2}{p^2q^2} + \frac{1}{3}\frac{(\hat{p}_i-p)^3}{p^3q^3}(p^2 - pq + q^2) + O(n_i^{-3}).$$

Consequently, according to (2.5) we have

$$E(\text{logit } p^*) \doteq L + \frac{1}{2N}(p-q) + \frac{1}{3N}\left[\sum_i \frac{1}{n_i}\right](q-p)(1-3pq)$$

because

$$E(\hat{p}_i - p)^3 = n_i^{-2} pq(q-p) .$$

Neglecting the third term, we have

$$E(\text{logit } p^*) \doteq L + \frac{1}{2N}(p-q) .$$

According to this approximation, as p takes small values, i.e., between 0 and .5 the bias value of logit p^* is of negative sign. For $p > .50$, the value of bias is positive. For $p = \frac{1}{2}$, logit p^* is unbiased. If the n_i's are sufficiently large, the value of bias tends to zero.

Following the derivations just presented, we conclude that the biases of MDI and MLG are of different signs. Thus one may see that an effective comparison can not be achieved by using analytical methods. In the next section we further study this issue numerically.

4. Computational Details of 2×2 Contigency Tables

Let (n_1, n_2) be a set of given raw totals. Then all possible tables can be enumerated whenever the set is specified. For each set (n_1, n_2) there are $(n_1+1)(n_2+1)$ tables to be generated. For each table, $\hat{p}$, p_M, p_p, p^* and p_L are computed according to formulae (2.1), (2.2), (2.5), (2.6), and (2.7), respectively. Since some estimators are not admissible for $X_i = 0$ or $X_i = n_i$, the following rule is used:

$$x_i = \begin{cases} \frac{1}{2} & \text{if } x_i = 0 \\ x_i & \text{if } x_i = 1,2,\ldots,n_i - 1 \\ n_i - \frac{1}{2} & \text{if } x_i = n_i \end{cases}$$

To compare estimators under the same experimental conditions, this rule was applied to all estimators. To get the mean, variance, standard deviation, bias, ratio of absolute value of bias to standard deviation, and exact level of estimators, the following values were used as true values of p,

$$p = .10\ (.05),\ .50$$

i.e., .10, .15,..., .50

The mean and variance of each estimator, for each p, was computed by using the definition of mean and variance of a random variable and formula λ given in section 2.

We now study the bias of estimators according to the increasing size of the sample for different true values of p. We computed the bias of estimators for all possible values of p, p = .10(.05), .50. The bias for p >.50 is easily obtained as the negative of bias at 1-p. To make the presentation of an overall conclusion for the values of biases self-explanatory, we preferred to give details on what we have observed, rather than drawing diagrams for them. If $(n_1,n_2) = (5,5)$, the bias of $\hat{p}$ is an increasing function of the true p for p <.50 and it is zero at p =.375 and p =.50. We note that the bias is due to the fact that 0 or n_i cells are replaced by $\frac{1}{2}$ and $n_i - \frac{1}{2}$, respectively, in the estimation procedure, otherwise $\hat{p}$ is unbiased. For MLE, $|B| < .034$ where B = b(p) = bias. The bias of P_M is a decreasing function of p for $p \leq .15$ and an increasing function of p for $.15 \leq p \leq .50$. It is zero at p =.50. Again, part of the bias is due to substituting $\frac{1}{2}$ for $x_i = 0$ and $n_i - \frac{1}{2}$ for $x_i = n_i$. Referring to (2.2) and the discussion in section 3, p_M would have been unbiased if the populations were normal. The bias of P_p is an increasing function of p for p <.45 and it decreases to zero at p =.50. Referring to formula (2.3), and

knowing that $|\frac{y_i}{y}| < 1$, we see that $|B|$ of p_p should be smaller than the corresponding $|B|$ of p_M. The maximum $|B|$ of p_p is equal to .021. The bias of p* behaves like that of p_M except that the bias is almost zero at .40. In other words, p* is an unbiased estimate of p in a small neighborhood of p =.50. $|B|$ of p* is less than $|B|$ of p_M. Finally, the bias of p_L behaves closely as p_p which is reasonable due to the relationship between X^2 and X_L^2 already given in Yassaee (1975).

For the case (n_1,n_2) = (10,5), (10,10), (10,15), (15,15), (20,20), and (20,25) we conclude that one cannot claim that p_L is always less biased than p_p, but other estimators can be arranged in terms of $|B|$ in ascending order

$$p - p^* - p_M$$

As n_1 and n_2 increase, the direction of the bias of $\hat{p}$ becomes the same as those of p_M and p*, which are always negative and increase to zero.

It is interesting to note that the bias of p* and that of p_L are of different sign for most of the cases except for small values of p, as we have already concluded approximately in sections (3.4) and (3.5) for bias of logit p* and logit p_L.

5. Mean Square Error of Estimators (MSE)

We briefly report on the MSE of estimators and compare estimators in this regard for various true values of p

$$(n_1,n_2) = (5,5)$$

For $p \leq .15$, the MSE of p_M is smaller than for those of others. For $.15 \leq p < .30$, the MSE of p* is the smallest and for $p \geq .30$ the MSE of p_L is the samllest.

For true values of p, the ascending order of estimators in terms of the magnitude of their MSE is given as follows:

p	Ascending order of MSE
$p \leq .15$	$p_M - p^* - \hat{p} - p_p - p_L$
$.15 < p < .30$	$p^* - p_M - \hat{p} - p_p - p_L$
$.30 \leq p \leq .50$	$p_L - p_p - \hat{p} - p^* - p_M$

for $p > .50$ the order is the same as that of $1 - p$ shown here.

We now present a table which summarizes the ascending order of estimators for the MSE according as $(n_1, n_2) = (5,10), (10,10), (15,10), (15,15)$.

Table of the Ascending Order of Estimators

(n_1, n_2)	p	Ascending order of MSE
5,10	$p \leq .20$	$p_M - p^* - \hat{p} - p_p - p_L$
	$p = .25$	$\hat{p} - p_p - p^* - p_L - p_M$
	$.25 < p \leq .50$	$p_L - p_p - \hat{p} - p^* - p_M$
10,10	$p \leq .20$	$p_M - p^* - \hat{p} - p_p - p_L$
	$p = .25$	$\hat{p} - p_p - p^* - p_L - p_M$
	$.25 < p \leq .50$	$p_L - p_p - p - p^* - p_M$
(15,10)	$p \leq .10$	$p_M - p^* - \hat{p} - p_p - p_L$
	$p = .15$	$p^* - p_M - \hat{p} - p_p - p_L$
	$p = .20$	$\hat{p} - p^* - p_p - p_M - p_L$
	$.20 < p \leq .30$	$p_p - p_L - \hat{p} - p^* - p_M$
	$.30 < p < .50$	$p_L - p_p - \hat{p} - p^* - p_M$
(15,15)	$p \leq .10$	$p_M - p^* - \hat{p} - p_p - p_L$
	$.15 \leq p \leq .20$	$p^* - \hat{p} - p_M - p_p - p_L$
	$p = .20$	$\hat{p} - p_p - p^* - p_L - p_M$
	$p = .25$	$p_p - p_L - \hat{p} - p^* - p_M$
	$p = .30$	$p_p - p_L - \hat{p} - p^* - p_M$
	$.35 \leq p \leq .50$	$p_L - p_p - \hat{p} - p^* - p_M$

REFERENCES

Berkson, J. (1946), "Approximation of χ^2 by probits and logits", JASA 41, pp 70-74.

Brown, C. C. and Muentz, L. R. (1976), "Reduced mean square error estimation in contingency tables", JASA 71, pp 176-182.

Chapman. J. A. W. (1976), "A comparison of the χ^2, -2 log R, and multinomial probability criteria for significance tests when expected frequencies are small", JASA 71, pp 854-863.

Cochran, W. G. (1952), "The χ^2 test of goodness-of-fit", Ann. Math. Stat. 22, pp 315-345.

Harvey, A. C. (1977), "A comparison of preliminary estimators for robust regressions", JASA 72, pp 910-913.

Hommel, G. (1978), "Tail probabilities for contingency tables with small expectations", JASA 73, pp 764-766.

Kullback, S. (1959), Information theory and statistics. John Wiley & Sons, New York (Dover Publ. Co. (1968), Peter Smith (1978)).

Pradhan, M. and Sathe, Y. S. (1976), "Analytical remarks on Cramer's minimum method for finding the better of two binomial populations", JASA 71, pp 239-241.

Prasad, R. (1970), "Probability distribution of algebraic functions of independent random variables", S.I.A.M., J. of Appl. Math. 18, pp 614-625.

Welch, B. L. (1938), "The significance of the difference between two means when the population variations are unequal", Biometrika 29, pp 350-362.

Yassaee, H. (1975a), "On Monte Carlo comparison of estimators in r×c contingency tables of small size", Technical Report, Arya-Mehr University of Technology.

_____(1975b), "A comparative exact and Monte Carlo study of estimators in multidimensional contingency tables: logit model", Title: Proceedings of multivariate analysis III, North Holland Publ. Co.

_____(1975c), "On comparison of various statistics in r×c contingency tables: Test of homogeneity based on small samples", Technical Report, Arya-Mehr University of Technology.

_____(1975d), "On Monte Carlo comparison of various statistics in r×c contingency tables of small sample size: independence model", Technical Report, Arya-Mehr University of Tehran, Iran.