

ON HOMOGENEITY TEST USING ESTIMATING FUNCTION

Fujii, Yoshinori
Department of Mathematics, Miyazaki University

<https://doi.org/10.5109/13436>

出版情報 : Bulletin of informatics and cybernetics. 26 (1/2), pp.101-107, 1994-03. Research
Association of Statistical Sciences

バージョン :

権利関係 :



ON HOMOGENEITY TEST USING ESTIMATING FUNCTION

By

Yoshinori FUJII*

Abstract

This paper examines the tests of homogeneity of odds ratios in independent experiments. Breslow and Day [2] and Tarone [10] proposed the tests which incorporate the Mantel-Haenszel estimator, but are essentially based on the efficient scores from the likelihood function. To broaden the distributional assumption and also to include the Mantel-Haenszel estimating equation, we introduce estimating equations and develop a homogeneity test. The extension makes it possible for us to apply this procedure to 2×2 tables which follow a negative binomial distribution.

Key words and phrases: Mantel-Haenszel estimator, negative binomial distribution, odds ratio, score test, 2×2 tables.

1. Introduction

The analysis of the association between two factors in a series of independent 2×2 tables is often conducted under the hypothesis that the odds ratios are constant across strata. However, it is important to confirm this homogeneity of odds ratios before incorporating them into an analysis. Several homogeneity tests have been proposed. The Pearson chi-squared test was introduced by Zelen [12] and later corrected by Halperin et. al. [4]. In addition many statistical packages, such as SAS, employ the likelihood ratio chi-squared test. These two procedures are closely related to the maximum likelihood estimator which is consistent in large sample cases when the number of tables is fixed and the sample size in each table is large, but inconsistent in sparse cases, i.e. the sample size in each table is fixed and the number of tables is large. Recently, due to the rapid development of computer technology, the Pearson chi-squared test based on the conditional maximum likelihood estimator has also been used. This test is superior to the former tests since the conditional maximum likelihood estimator is dually consistent i.e. consistent in the sparse cases as well as in the large sample cases.

The Mantel-Haenszel estimator of the common odds ratio is also dually consistent (Gart [3] and Breslow [1]) and is preferred for estimation problems. It is also consistent when subject responses are correlated within tables (Liang [5]). The Mantel-Haenszel

* Department of Mathematics, Miyazaki University, Miyazaki 889-21, Japan
E-mail: e03105u@muipc.cc.miyazaki-u.ac.jp

estimator has these properties because it is given by the solution of a simple estimating equation. To save the computational time needed to compute the Pearson chi-square test based on the conditional maximum likelihood estimator, Breslow and Day ([2], p142) proposed the use of the Mantel- Haenszel estimator in place of the conditional maximum likelihood estimator. Later, Tarone [10] suggested subtracting a correction factor from the test statistic because of the inefficiency of the Mantel-Haenszel estimator. However, the Pearson chi-squared test is equivalent to the efficient score test based on the likelihood function. Therefore, we feel that it reduces the effectiveness of the Mantel-Haenszel estimator. To broaden the distributional assumption and also to include the Mantel-Haenszel estimating equation, we introduce estimating equations and develop a homogeneity test. The extension makes it possible for us to apply this procedure to 2×2 tables which follow a negative binomial distribution.

2. Proposed test

2.1. General theorem

We are interested in a univariate parameter θ . For the inference of θ suppose that the data are obtained from K independent experiments. Let x_k represent a random sample of size n_k from an underlying distribution in the k -th experiment. We use an estimating function for θ given x_k which is denoted by $g_k(x_k; \theta)$. Assume that each estimating function satisfies the assumptions below.

Assumptions

- (1) $E\{g_k(x_k; \theta)\} = 0$ for any θ .
- (2) $E\{\frac{\partial}{\partial \theta} g_k(x_k; \theta)\} < 0$ for any θ .
- (3) There exist a variance function $V_k(\theta)$ such that

$$\frac{g_k(x_k; \theta)}{V_k(\theta)^{1/2}} \rightarrow N(0, 1) \quad \text{as } n_k \rightarrow \infty$$

Let the estimator $\hat{\theta}$ for the parameter θ be defined by the solution of the equation $\sum g_k(x_k; \theta) = 0$. We first consider the statistic

$$T = \sum_{k=1}^K \frac{g_k(x_k; \hat{\theta})^2}{V_k(\hat{\theta})}$$

which is obtained by a method similar to the score test statistics. We have the following theorem for the distribution of T .

THEOREM 2.1. *Let D_1 and D_2 be independent random variables that follow chi-squared distributions with $K-2$ and 1 degree of freedom, respectively. The statistic T is asymptotically equivalent to $D_1 + cD_2$ in the large sample cases where*

$$c = \frac{\sum_{j=1}^K V_j(\theta) \sum_{k=1}^K E\{\frac{\partial}{\partial \theta} g_k(x_k; \theta)\}^2 / V_k(\theta)}{\left[\sum E\{\frac{\partial}{\partial \theta} g_k(x_k; \theta)\} \right]^2}$$

PROOF. Let

$$\mathbf{y} = (g_1(x_1; \hat{\theta}) V_1(\theta)^{-1/2}, g_2(x_2; \hat{\theta}) V_2(\theta)^{-1/2}, \dots, g_K(x_K; \hat{\theta}) V_K(\theta)^{-1/2})^t.$$

Note that we substitute θ not $\hat{\theta}$ in the variance function. We consider the asymptotic variance of \mathbf{y} denoted by

$$AVar(\mathbf{y}) = \mathbf{V}^{-1/2}(\mathbf{I} - \frac{1}{\beta} \mathbf{b} \mathbf{a}^t)(\mathbf{I} - \frac{1}{\beta} \mathbf{a} \mathbf{b}^t) \mathbf{V}^{-1/2}$$

where

$$\begin{aligned} \mathbf{V} &= \text{diag}\{V_1(\theta), V_2(\theta), \dots, V_K(\theta)\}, \\ \mathbf{a} &= (V_1^{1/2}(\theta), V_2^{1/2}(\theta), \dots, V_K^{1/2}(\theta))^t, \\ \mathbf{b} &= \mathbf{V}^{-1/2} (E\{\frac{\partial}{\partial \theta} g_1(x_1; \theta)\}, E\{\frac{\partial}{\partial \theta} g_2(x_2; \theta)\}, \dots, E\{\frac{\partial}{\partial \theta} g_K(x_K; \theta)\})^t, \end{aligned}$$

\mathbf{I} is the order K identity matrix and $\beta = \mathbf{a}^t \mathbf{b}$.

We choose a base in R^K whose first and second members are \mathbf{a} and \mathbf{b} . If we construct an orthogonal matrix Γ based on the base by Gram-Schmidt process, then

$$\begin{aligned} \Gamma^t AVar(\mathbf{y}) \Gamma &= \Gamma^t \mathbf{V}^{-1/2} (\mathbf{I} - \frac{1}{\beta} \mathbf{b} \mathbf{a}^t) (\mathbf{I} - \frac{1}{\beta} \mathbf{a} \mathbf{b}^t) \mathbf{V}^{-1/2} \Gamma \\ &= \text{diag}(0, c, 1, 1, \dots, 1). \end{aligned}$$

So we can divide $\sum g_k(x_k; \hat{\theta})^2 / V_k(\theta)$ by two random variables which are asymptotically equivalent to cD_2 and D_1 . Since $\hat{\theta}$ is a consistent estimator for θ , $V_k(\hat{\theta})/V_k(\theta)$ converges in probability to 1 when $n_k \rightarrow \infty$ for all k . We can show that the statistic T is asymptotically equivalent to $\sum g_k(x_k; \hat{\theta})^2 / V_k(\theta)$. \square

Cauchy-Schwarz's inequality derives that c is not less than unity and c equals to 1 if and only if $V_k(\theta) = \alpha E\{\frac{\partial}{\partial \theta} g_k(x_k; \theta)\}$, for some constant α . If we use the derivative of log likelihood for an estimating function, the statistic T , which is equivalent to efficient score test, is asymptotically equivalent to a chi-squared random variable with $K-1$ degree of freedom. Generally, the statistic T is asymptotically larger than the chi-squared random variable with $K-1$ degree of freedom. We next consider a correction of T which is asymptotically equivalent to the chi-squared random variable.

THEOREM 2.2. Let

$$T^* = \sum_{k=1}^K \frac{g_k(x_k; \hat{\theta})^2}{V_k(\hat{\theta})} - \frac{\left[\sum g_k(x_k; \hat{\theta}) E\{\frac{\partial}{\partial \theta} g_k(x_k; \theta)\} / V_k(\hat{\theta}) \right]^2}{\sum E\{\frac{\partial}{\partial \theta} g_k(x_k; \theta)\}^2 / V_k(\hat{\theta})}.$$

Then T^* follows an asymptotic chi-squared distribution with $K-1$ degrees of freedom.

PROOF. In the proof of THEOREM 2.1, $\sum g_k(x_k; \hat{\theta})^2/V_k(\theta)$ is divided by two random variables. That asymptotically equivalent to cD_2 is equal to the square of the second factor of $\Gamma^t \mathbf{y}$. It is expressed by the following form

$$\text{const.} \sum g_k(x_k; \hat{\theta}) E\left\{ \frac{\partial}{\partial \theta} g_k(x_k; \theta) \right\} / V_k(\theta).$$

T^* is the result of subtracting the extra part of above form from T . □

COROLLARY 2.3. *If we replace $E\{\frac{\partial}{\partial \theta} g_k(x_k; \theta)\}$ and $V_k(\hat{\theta})$ by consistent estimators respectively, T^* also follows an asymptotic chi-squared distribution with $K-1$ degree of freedom.*

2.2. Several 2×2 tables

We apply THEOREM 2.2 to test for the homogeneity of odds ratios in several 2×2 tables. Consider K 2×2 tables with a pair of independent binomial observations (x_k, y_k) with denominator (n_k, m_k) and success probability (p_{1k}, p_{2k}) for each k ($k=1, 2, \dots, K$). Put the odds ratio in the k -th table $\psi_k = p_{1k}q_{2k}/p_{2k}q_{1k}$ where $q_{ik} = 1 - p_{ik}$ for $i=1, 2$. We use $g_k(x_k, y_k; \psi) = R_k - \psi S_k$ for estimating functions where $R_k = x_k(m_k - y_k)/(n_k + m_k)$, and $S_k = y_k(n_k - x_k)/(n_k + m_k)$. The Mantel-Haenszel estimator $\hat{\psi}$ of the common odds ratio is given by the solution of an estimating equation $\sum g_k(x_k, y_k; \psi) = 0$. We can estimate consistently $E\{\frac{\partial}{\partial \theta} g_k(x_k, y_k; \psi)\}$ by $-S_k$. Robin et. al. [8], Phillips and Holland [7] and Sato [9] have discussed the consistent estimator of $V_k(\psi)$ in relation to estimation problems of the variance of the Mantel-Haenszel estimator. We use the estimator proposed by Phillips and Holland [7] as follows

$$\hat{V}_k = \frac{1}{2}(R_k + \hat{\psi} S_k)(P_k + \hat{\psi} Q_k)$$

where $P_k = (x_k + m_k - y_k)/(n_k + m_k)$, and $Q_k = (y_k + n_k - x_k)/(n_k + m_k)$. Then we propose the following test statistics

$$T^* = \sum_{k=1}^K \frac{(R_k - \hat{\psi} S_k)^2}{\hat{V}_k} - \frac{\left\{ \sum (R_k - \hat{\psi} S_k) S_k / \hat{V}_k \right\}^2}{\sum S_k^2 / \hat{V}_k}.$$

This procedure can be also applied to the non-central hypergeometric model since \hat{V}_k and $-S_k$ are also a consistent estimator for $V_k(\psi)$ and $E\{\frac{\partial}{\partial \psi} g_k(x_k, y_k; \psi)\}$, respectively.

2.3. The extended homogeneity test in negative binomial models

We next consider the extension of the homogeneity test in negative binomial models. Let x_{ki} ($i=1, 2, \dots, n_k$) and y_{kj} ($j=1, 2, \dots, m_k$) be samples of size n_k and m_k from

k-th experiment having negative binomial distributions $NB(\mu_{k1}, \delta)$ and $NB(\mu_{k2}, \delta)$, respectively. The probability function of $NB(\mu, \delta)$ is given by

$$pr(x, \mu, \delta) = \frac{1}{x!} \left(\prod_{h=0}^{x-1} (1 + h\delta) \right) \left(\frac{\mu}{1 + \delta\mu} \right)^x \left(\frac{1}{1 + \delta\mu} \right)^{1/\delta}.$$

Yanagimoto [11] proposed the extended odds ratio ψ_k in negative binomial models as follows

$$\psi_k = \frac{\mu_{k1}(1 + \delta\mu_{k2})}{\mu_{k2}(1 + \delta\mu_{k1})}.$$

For simplicity, we denote the sum of the samples of the k-th experiment by x_{k+} and y_{k+} , respectively. When $\delta = -1$ and $0 < \mu_{ki} < 1$ for $i = 1, 2$ then x_{k+} and y_{k+} have independent binomial distribution with denominators (n_k, m_k) and also success probabilities (μ_{k1}, μ_{k2}) , and ψ_k equals to the odds ratio in binomial models. We can easily extend the above homogeneity test for given δ . We use the following estimating function

$$\tilde{g}_k(x_{k+}, y_{k+}; \psi) = \tilde{R}_k - \psi \tilde{S}_k$$

where $\tilde{R}_k = x_{k+}(m_k + \delta y_{k+}) / (n_k + m_k)$ and $\tilde{S}_k = y_{k+}(n_k + \delta x_{k+}) / (n_k + m_k)$. For consistent estimators of $E\{\frac{\partial}{\partial \psi} g_k(x_{k+}, y_{k+}; \psi)\}$ and $Var\{g_k(x_{k+}, y_{k+}; \psi)\}$, we use $-\tilde{S}_k$ and

$$\begin{aligned} \tilde{V}_k &= \frac{1}{2} \{x_{k+}(m_k + \delta y_{k+}) + \hat{\psi} y_{k+}(n_k + \delta x_{k+})\} \\ &\quad \cdot \{m_{k+} + \delta(y_{k+} - x_{k+}) + \hat{\psi} y_{k+}(n_k + \delta x_{k+} - \delta y_{k+})\} / (n_k + m_k). \end{aligned}$$

\tilde{V}_k has been given by Yanagimoto [11] as an extension of Phillips and Hollands estimator in binomial models. Then we have

$$\tilde{T} = \sum_{k=1}^K \frac{\tilde{g}_k(x_{k+}, y_{k+}; \hat{\psi})^2}{\tilde{V}_k}$$

and

$$\tilde{T}^* = \sum_{k=1}^K \frac{\tilde{g}_k(x_{k+}, y_{k+}; \hat{\psi})^2}{\tilde{V}_k} - \frac{\left\{ \sum (\tilde{R}_k - \hat{\psi} \tilde{S}_k) \tilde{S}_k / \tilde{V}_k \right\}^2}{\sum \tilde{S}_k^2 / \tilde{V}_k}.$$

We can easily show that \tilde{T}^* equals to T^* when $\delta = -1$.

3. Example

As an illustration of this procedure we use the data from the case-control study of esophageal cancer given by Breslow and Day([2],p145). The data are summarized in six 2×2 tables. We obtain the Mantel-Haenszel estimate, On the other hand, the Pearson chi-squared test based on the maximum likelihood estimator and the Mantel-Haenszel estimator are 9.320 and 9.323, respectively. Tarone's correction of the Pearson chi-squared test is 9.299. For this data, the proposed test statistic is smaller than the

other three test statistics. Several sparse cells bring the result because the Pearson chi-squared tests don't perform well when the data are sparse. (Liang and Self [6]) Also the proposed tests may be conservative because estimator of variance is not so accurate in the sparse table. As an improvement, we may use the conditional expectation of the square of $g_k(x_k, y_k; \psi)$ conditioned all marginals substituting $\hat{\psi}$ for ψ . We need the further study about the distribution of T^* in sparse cases and the estimator of variance of $g_k(x_k, y_k; \psi)$.

Acknowledgements

The author would like to thank Professor T. Yanagawa who suggested that we apply this procedure to a negative binomial model. This research was supported in part by the Grant-in-Aid for Encouragement of Young Scientist no. A-05740139 from the Ministry of Education, Science and Culture.

References

- [1] Breslow, N. E.: *Odds ratio estimators when the data are sparse*. Biometrika, **68** (1981), 73–84.
- [2] Breslow, N. E. and Day, N. E.: *Statistical Methods in Cancer Research, 1: The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer. (1980).
- [3] Gart, J. J.: *On the combination of the relative risks*. Biometrics, **18** (1962), 594–600.
- [4] Halperin, M., Ware, J. H. , Byar, D. P., Mantel, N. , Brown, C. C. , Koziol, J. , Gail, M. and Green, S. B.: *Testing for interaction in an $I \times J \times K$ contingency table*. Biometrika, **64** (1977), 271–5.
- [5] Liang, K. -Y.: *Odds ratio inference with dependent data*. Biometrika, **72** (1985), 678–682.
- [6] Liang, K. -Y. and Self, S. G.: *Test for homogeneity of odds ratio when the data sparse*. Biometrika, **72** (1985), 353–358.
- [7] Phillips, A. and Holland, P. W.: *Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate*. Biometrics, **43** (1987), 425–431.
- [8] Robin, J. M., Breslow, N. E., and Greenland, S.: *Estimators of the Mantel-Haenszel variance consistent in both sparse-data and Large-strata limiting models*. Biometrics, **42** (1986), 623–635.
- [9] Sato, T.: *Confidence limits to the common odds ratio based on the asymptotic distribution of the Mantel-Haenszel estimator*. Biometrics, **46**, 71–80.
- [10] Tarone, R. E.: *On heterogeneity tests based on efficient scores*. Biometrika, **72** (1985), 91–5.

- [11] Yanagimoto, T.: *The Mantel-Haenszel statistics for the extended odds ratio in the negative binomial distribution*. Journal of Japan Statist. Soc. , **22** (1992), 7–17.
- [12] Zelen, M.: *The analysis of several 2×2 contingency tables*. Biometrika, **58** (1971), 129–37.

Received November 26, 1993

Revised February 5, 1994

Communicated by T. Yanagawa