Environmental Health Perspectives
Vol. 87, pp. 103–107, 1990

Statistical Analysis of *K* 2 × 2 Tables: A Comparative Study of Estimators/Test Statistics for Association and Homogeneity

by Thomas W. O'Gorman,* Robert F. Woolson,* Michael P. Jones,* and Jon H. Lemke*

In order to control for confounding variables, epidemiologists often obtain data in the form of a 2 × 2 table. One variable is usually the disease status, while the other variable represents a dichotomous exposure variable that is suspected of being a risk factor. If a confounding variable is present, the data are often stratified into several 2 × 2 tables. The objectives of the analysis are to test for the association between the suspected risk factor and the disease and to estimate the strength of this relationship. Before estimating a common odds ratio, it is important to check whether the odds ratios are homogeneous. This paper presents the results of a Monte Carlo study that was performed to determine the size and power of a number of tests of association and homogeneity when the data are sparse. We also evaluated the performance of three estimators of the common odds ratio. For the Monte Carlo studies, equal numbers of cases and controls were used in a wide variety of sparse data situations. On the basis of these studies, we recommend the Breslow-Day test for nonsparse data, and the T_4 and T_5 statistics for sparse data to test for homogeneity. The Mantel-Haenszel test of association is recommended for sparse and nonsparse data sets. With sparse data, none of the odds ratio estimators are entirely satisfactory.

Introduction

Epidemiologists often stratify data to control for a confounding variable in order to evaluate the relationship between a suspected risk factor and disease. If K levels of the confounding variable are used, and if the risk factor and the disease are dichotomous, then the data can be arranged in a $K \times 2 \times 2$ table of observed cell counts.

A common objective is to perform a test of association between the disease and the risk factor after controlling for the confounding variable. Another common objective is to estimate the common disease, or the exposure odds ratio. Before performing a test of association, it is desirable to determine if the assumption of a common disease, the exposure odds ratio is tenable. Therefore, a test of homogeneity is often a first step in the analysis of several 2 × 2 tables.

Many tests and estimates have been proposed for multiway contingency tables (*1*). However, these tests that are based on asymptotic distribution theory may not be valid when used on tables having few observations in some of the cells. One objective of this study is to evaluate the behavior of these large sample tests when some of the counts are small, i.e., for sparse data.

In the last few years some homogeneity tests have been designed specifically for the sparse data setting where the number of strata (K) is large but the cell counts are small (*2*). A second objective is to evaluate the performance of these sparse data tests. Monte Carlo methods were used in this study since analytic comparisons are not feasible for these situations. This paper reviews a portion of the more detailed studies we have published on these topics. (*3,4*).

Description of Simulation

In both studies (*3,4*) we use Monte Carlo methods to generate cell counts for $K \times 2 \times 2$ tables. These cell counts are used to compute homogeneity tests, association tests, and odds ratio estimators. For the ith 2 × 2 table we use the notation for the cell counts presented in Table 1.

Let x_i be the binomial count from n_i independent trials with probability of success p_{1i}, and let y_i be an independent binomial count from m_i independent trials with probability of success p_{2i}. In a case-control study x_i is the number of exposed cases while y_i is the number of exposed controls. Let $t_i = x_i + y_i$.

*Division of Biostatistics, Department of Preventive Medicine, University of Iowa, Iowa City, Iowa 52242.
Address reprint requests to R. F. Woolson, Division of Biostatistics, Department of Preventive Medicine, University of Iowa, Iowa City, IA 52242.

Table 1. Notation for cell counts.

Group	Exposed	Unexposed	Total
Cases	x_i	$n_i - x_i$	n_i
Controls	y_i	$m_i - y_i$	m_i
	t_i	$N_i - t_i$	N_i

For the Monte Carlo study we specify the probability of a control having been exposed (p_{2i}) for $i = 1, \ldots, K$ and we specify the odds ratio (ψ_i) for $i = 1, \ldots, K$. Using the formula

$$p_{1i} = \frac{\psi_i \, p_{2i}}{1 - p_{2i} + \psi_i \, p_{2i}}$$

we compute the probability of a case having been exposed for $i = 1, \ldots, K$. For the ith stratum the number of exposed cases (x_i) is the number of random numbers that are less that p_{1i} out of n_i calls to the uniform [0,1) random number generator. The number of exposed controls (y_i) are obtained in a similar fashion and the remainder of the table is computed by subtraction. In our simulation studies equal numbers of cases and controls were used in all strata. Also, for most of the studies the strata were balanced so that

$$m_i = n_i = N/(2K)$$

$$\text{for } i = 1, \ldots, K, \text{ where } N = \sum_{i=1}^{K} N_i$$

Monte Carlo simulations based on 1000 tables were performed for a range of values of the odds ratio (ψ_i), the probability of exposure for a control (p_{2i}), the number of strata (K), and the total number of cases (n_+) and controls (m_+). We used $n_+ = m_+ = 64, 128, 256, 512$ with $K = 2, 4, 8, 16, 32$. The probability of exposure for a case was set at $p_{2i} = 0.05, 0.10, 0.30, 0.50$ for $i = 1, \ldots, K$.

For studies of the odds ratio estimator and tests of association we used a constant odds ratio of $\psi_i = 1, 2, 4, 8, 16$ for $i = 1, \ldots, K$. For studies of the power of the tests of homogeneity the ψ_i were generated according to several distributions. These distributions included the lognormal, exponential, two-point, and uniform. For more details concerning the distributions used see Jones et al. (*4*).

In these sparse data simulations it is not uncommon to generate a zero for a cell count. These zero counts sometimes required special handling. If either $t_i = 0$ or $N_i - t_i = 0$ then the ith table has a zero contribution to the test statistics for homogeneity and association and to the odds ratio estimates. See O'Gorman et al. (*3*) and Jones et al. (*4*) for details on these situations.

Description of Tests

Tests of Homogeneity

The likelihood ratio test of homogeneity (LRTH) and the Pearson test of homogeneity (PH) can be computed from the maximum likelihood cell estimates (1) of the cell probabilities. These estimates are obtained from the iterative proportional fitting algorithm (*5*).

Breslow and Day (*6*) proposed the statistic

$$\text{BD} = \sum_{i=1}^{K} \frac{[x_i - e_i(\hat{\psi}_{\text{MH}})]^2}{\text{Var}(x_i \mid \hat{\psi}_{\text{MH}})}$$

where $\hat{\psi}_{\text{MH}}$ is the Mantel-Haenszel (*7*) estimator of the common odds ratio; $e_i(\hat{\psi}_{\text{MH}})$ is the expected value of x_i given $\hat{\psi}_{MH}$ and is computed as the solution to the quadratic equation $e_i(m_i - t_i + e_i) = \hat{\psi}_{MH}(n_i - e_i)(t_i - e_i)$; and the variance estimator is given by $\text{Var}(x_i \mid \hat{\psi}_{\text{MH}}) = \{1/e_i + 1/(n_i - e_i) + 1/(t_i - e_i) + 1/(m_i - t_i + e_i)\}^{-1}$.

Tarone (*8*) recommended the test statistic

$$\text{MBD} = \sum_{i=1}^{K} \frac{[x_i - e_i(\hat{\psi}_{\text{MH}})]^2}{\text{Var}(x_i \mid \hat{\psi}_{\text{MH}})} - \frac{\left[\sum_{i=1}^{K} x_i - \sum_{i=1}^{K} e_i(\hat{\psi}_{\text{MH}})\right]^2}{\sum_{i=1}^{K} \text{Var}(x_i \mid \hat{\psi}_{\text{MH}})}$$

which differs from BD only by the correction term.

Given the table margins n_i, m_i, and t_i, $i = 1, \ldots, K$, the conditional likelihood is $\Pi h_i(x_i \mid \psi_i)$ where $h_i(x_i \mid \psi_i)$ is the noncentral hypergeometric density. For future reference let us define the exact conditional moments $E_c(x_i^r \mid \psi) = \Sigma u^r h_i(u \mid \psi)$ for $r \geq 1$, where the summation is over u. From the conditional likelihood we derive the conditional score test

$$\text{CS} = \sum_{i=1}^{K} \frac{\{x_i - E_c[x_i \mid \hat{\psi}_c]\}^2}{\text{Var}_c[x_i \mid \hat{\psi}_c]}$$

where $\hat{\psi}_c$ is the maximum conditional likelihood estimator of ψ (*11*) and $\text{Var}_c[x_i \mid \hat{\psi}_c] = E_c[x_i^2 \mid \hat{\psi}_c] - E_c^2[x_i \mid \hat{\psi}_c]$.

Liang and Self (*2*) derived two tests for the sparse data situation. The first statistic is a normal approximation to the conditional score test, CS and is denoted by T_4 where:

$$T_4 = \frac{\text{CS} - K^*}{[\text{Var}^{\text{A}}(\text{CS} \mid \hat{\psi}_c) - B^2(\hat{\psi}_c)V(\hat{\psi}_c)]^{1/2}}$$

where K^* is the number of tables with nonzero margins; $V(\psi) = \psi_2/[\Sigma \, \text{Var}_c(x_i \mid \psi)]$;

$$B(\psi) = \sum_{i=1}^{K} \frac{E_c(x_i^3 \mid \psi) - 3E_c(x_i^2 \mid \psi)\, E_c(x_i \mid \psi) + 2E_c^3(x_i \mid \psi)}{\hat{\psi}_c \, \text{Var}(x_i \mid \psi)}$$

$$\text{Var}^{\text{A}}(\text{CS} \mid \psi) = \sum_{i=1}^{K} \frac{E_c\{[x_i - E_c(x_i \mid \psi)]^4 \mid \psi\}}{\text{Var}_c^2(x_i \mid \psi)} - K^*$$

The second score test statistic is a normal approximation for a mixture model and is given by

$$T_5 = S[I_{11} - I_{12}^2 I_{22}^{-1}]^{-1/2}$$

where

$$S = \Sigma \{[x_i - E_c(x_i \mid \hat{\psi}_c)]^2 - \mathrm{Var}_c(x_i \mid \hat{\psi}_c)\},$$
$$I_{11} = \Sigma E_c\{[(x_i - E_c(x_i \mid \hat{\psi}_c)]^2 - \mathrm{Var}_c(x_i \mid \hat{\psi}_c)]^2 \mid \hat{\psi}_c\},$$
$$I_{12} = \Sigma E_c\{[x_i - E_c(x_i \mid \hat{\psi}_c]^3 \mid \hat{\psi}_c\},$$
$$I_{22} = \Sigma \mathrm{Var}_c(x_i \mid \hat{\psi}_c).$$

The LRTH, PH, MBD, and CS test statistics are asymptotically χ^2_{k-1} random variables under Ho: $\psi_i = \psi$ for $i = 1, \ldots, K$. In the sparse data setting, as $K \to \infty$ the tests statistics T_4 and T_5 are both asymptotically normal. While the asymptotic null distribution of BD is not chi-square, (Fuji and Yanagawa, personal communication), we compare it to the chi-square tabular values in our simulation studies.

Tests of Association

Four tests of association for each $K \times 2 \times 2$ table are compared by O'Gorman et al. (*3*). These are the likelihood ratio (LRA) statistic, the Mantel-Haenszel (MH) statistic, the Pearson (PA) statistic, and the weighted least squares (WLS) statistic. The LRA and PA statistics are computed from the maximum likelihood estimates of the cell counts. These estimates are obtained by using the iterative proportional fitting algorithm (*5*). Bishop, Fienberg, and Holland (*1*) show that these statistics are equal to the goodness-of-fit statistic for the no association model minus the goodness-of-fit for the association model.

The Mantel-Haenszel (*7*) test statistic for association is

$$\chi^2_{\mathrm{MH}} = \frac{\left\{\sum_{i=1}^{K} x_i - \sum_{i=1}^{K} \frac{n_i t_i}{N_i}\right\}^2}{\sum_{i=1}^{K} \frac{n_i m_i t_i (N_i - t_i)}{N_i^2 (N_i - 1)}}$$

The weighted least squares test statistic for association according to Wolf (*10*) is

$$\chi^2_{\mathrm{WLS}} = \frac{\left\{\sum_{i=1}^{K} W_i \log\left[\frac{x_i(m_i - y_i)}{y_i(n_i - x_i)}\right]\right\}^2}{\sum_{i=1}^{K} W_i}$$

where

$$W_i = \left\{\frac{1}{x_i} + \frac{1}{m_i - y_i} + \frac{1}{y_i} + \frac{1}{n_i - x_i}\right\}^{-1}$$

If any of the cells are zero they are replaced by 0.5 in our simulations before calculating χ^2_{WLS}.

Odd Ratio Estimators

The three odds ratio estimators that are used in this study are the Mantel-Haenszel (*1*) estimator, the weighted least squares estimator (*9*), and the conditional maximum likelihood estimator (*10*).

The Mantel-Haenszel estimator is defined by

$$\hat{\psi}_{\mathrm{WLS}} = \frac{\sum_{i=1}^{K} x_i(m_i - y_i)/N_i}{\sum_{i=1}^{K} y_i(n_i - x_i)/N_i}$$

The weighted least squares estimator is defined by

$$\hat{\psi}_{\mathrm{WLS}} = \exp\left\{\frac{\sum_{i=1}^{K} W_i \log \frac{x_i(m_i - y_i)}{y_i(n_i - x_i)}}{\sum_{i=1}^{K} W_i}\right\}$$

where W_i is defined above for the weighted least squares test for association.

The third estimator considered in this evaluation is the conditional maximum likelihood estimator which is defined by first considering the conditional distribution of x_i given the margins of the table, i.e. given (n_i, m_i, t_i, $N_i - t_i$). Following Gart (*10,11*) this distribution is a noncentral hypergeometric distribution for each stratum, that is,

$$Pr[X_i = x_i \mid n_i, m_i, t_i] = \frac{\binom{n_i}{x_i}\binom{m_i}{t_i - x_i}\psi^{x_i}}{\sum_{u=\max(0, t_i - m_i)}^{\min(t_i, n_i)} \binom{n_i}{u}\binom{m_i}{t_i - u}\psi^{u}}$$

for $i = 1,2, \ldots, K$. The conditional maximum likelihood estimator, $\hat{\psi}_{\mathrm{MCLE}}$ is defined as the root of

$$\sum_{i=1}^{K} x_i = \sum_{i=1}^{K} E_i(X_i; \psi),$$

where $E_i(X_i; \psi)$ is the mean of X_i.

Results of Monte Carlo Study

Full details of the simulations are described elsewhere (*3,4*). Here we only describe the key findings from our studies.

Results for Tests of Homogeneity

The sizes of the tests of homogeneity are estimated from the percentage of times the hypothesis of a common odds ratio is rejected. When compared to the chi-square tabular values, the tests based on PH, BD, MBD, and CS generally maintain their nominal size in the large stratum situation, while the test based on LRTH rejects much too often. In the sparse data situation the tests based on T_4 and T_5 maintain their size, while generally the tests using PH, BD, MBD, and CS do not reject often enough.

The powers of the tests are estimated by the number

of times the test statistics lead to rejection of the hypothesis of a common odds ratios when the odds ratios were not held constant. The odds ratios were generated according to lognormal, exponential, two-point, and uniform distributions. For those tests that maintain their sizes near the 5% level, the PH, BD, MBD, and CS tests have about equal power for the large stratum setting where they are superior to T_4 and T_5. In the sparse data setting T_4 and T_5 are generally more powerful than the other statistics.

It should be noted that all of these tests of homogeneity have low power. For example, with 128 cases and 128 controls, if we generate ψ_i from a uniform [1.0, 4.0] distribution the power is less than 13% for all of the tests for $K = 2, 4, 8, 16$, and 32. Also, for many situations studied, the power is not sensitive to the number of strata (K) so long as the total sample size is kept constant. Jones et al. (*4*) gives further details and a discussion concerning the reasons for the low power.

We also studied the situation where 50% of the cases and controls were placed in one large table while the remainder were placed equally among the other tables. In these unbalanced tables we used $\psi = 1$ for the large table and $\psi > 1$ for the other tables.

For these situations the test based on T_5 was most powerful and the test based on BD was the second most powerful. In our studies T_5 performed well in both the balanced and unbalanced sparse data settings, while the BD statistic performed well in the large stratum settings.

Results for Tests of Association

The MH test maintained its size for both large stratum and small stratum situations. The LRA test held its size for large stratum but tended to be anti-conservative in the small stratum setting. The PA and WLS tests maintained their size for the large stratum case but were much too conservative with sparse data. The powers of these tests were estimated by the proportion that led to rejection of the hypothesis of a common odds ratio of 1.0 when the common odds ratio exceeded 1.0. The power of the LRA and MH test were approximately equal and were not related to the number of strata used. Because of their conservative sizes, the powers of the PA and WLS tests were considerably below the powers of the LRA and MH tests with sparse data.

Results for Odds Ratio Estimators

The median and the interquartile ranges of the three odds ratio estimators were also estimated in the Monte Carlo study. When $\psi_i = 1.0$ for $i = 1, \ldots, K$ all three estimators have median values near 1.0. For sparse data the interquartile range of $\hat{\psi}_{\mathrm{WLS}}$ is less than that of the other two estimators. For $\psi_i = 4.0$ for $i = 1, \ldots, K$, the median values of $\hat{\psi}_{\mathrm{MH}}$ and $\hat{\psi}_{\mathrm{MCLE}}$ are near 4.0 for nonsparse data and for sparse data. For $\psi_i = 4.0$ the median values of $\hat{\psi}_{\mathrm{WLS}}$ are near 4.0 for nonsparse data but are much below 4.0 for sparse data.

For nonsparse data the variability of the three odds ratio estimators are approximately equal. For sparse data the interquartile range of $\hat{\psi}_{\mathrm{WLS}}$ is less than that of $\hat{\psi}_{\mathrm{MH}}$ and $\hat{\psi}_{\mathrm{MCLE}}$.

Summary

We compared the performance of three combined odds ratios estimators and four tests of association using Monte Carlo techniques (*3,4*). For these Monte Carlo studies a constant odds ratio is used with an equal number of cases and controls. In addition, a wide range of odds ratios, probabilities of exposure, numbers of cases, and strata are used. For each of the K 2×2 tables, 1000 simulations were generated for each configuration of the parameters studied. The Mantel-Haenszel (*7*), the weighted least squares (*9*), and maximum conditional likelihood (*10*) estimators of the odds ratio were computed. In addition, the likelihood ratio (*1*), Mantel-Haenszel, Pearson, and weighted least squares tests of association are studied. These studies indicate that the interquartile range of the weighted least squares estimator is usually less that of the other estimators; although in many situations the median of this least squares estimator is far from the population odds ratio. With sparse data the Mantel-Haenszel test for association maintains its size. For the range of parameters studied here, the degree of stratification does not greatly affect the power of the likelihood ratio and the Mantel-Haenszel test statistics.

In addition to studying measures of association and tests for association, we also examine several tests for homogeneity. We conclude that the Breslow Day statistic (*6*) is a reasonable statistic for use in nonsparse data settings when taking into account both the size and power of the test. In balanced sparse data settings the T_4 statistic of Liang and Self (*2*) performs the best when all tables, regardless of sample size, have odds ratios generated from the same distribution. In sparse data settings characterized by a large table with an odds ratio of 1 and many small tables of odds ratios greater than 1, the T_5 statistic of Liang and Self (*2*) performs the best. One result of these investigations is that virtually all of the homogeneity tests have generally low power in the presence of sparse data.

Recommendations

The Breslow-Day test of homogeneity is recommended for nonsparse data. For sparse data the T_4 and T_5 statistics are the most powerful tests of homogeneity and are recommended. The choice between T_4 and T_5 should be based on considerations found in (*4*). For tests of association the Mantel-Haenszel test is recommended. The three estimators studied here cannot be recommended for sparse data, although the Mantel-Haenszel performs reasonably well. A modified version of $\hat{\psi}_{\mathrm{MH}}$ studied by Hauck et al. (*12*) may be preferred in extreme sparse data settings.

This work supported in part by grant #CA39065 from the National Cancer Institute, U.S. Public Health Service.

REFERENCES

1. Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. Discrete Multivariate Analysis. Massachusetts Institute of Technology Press, Cambridge, MA, 1975.
2. Liang, K. Y., and Self, S. G. Tests for homogeneity of odds ratio when the data are sparse. Biometrika 72(2): 353–8 (1985).
3. O'Gorman, T. W., Woolson, R. F., Jones, M. P., and Lemke, J. H. A Monte Carlo study of three odds ratio estimators and four tests of association in several 2 × 2 tables when the data are sparse. Commun. Stat. 17(3): 813–835 (1988).
4. Jones, M. P., O'Gorman, T. W., Lemke, J. H., and Woolson, R. F. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. Biometrics 45: 171–181 (1989).
5. Deming, W. E., and Stephan, F. F. On a least squares adjustment of a sample frequency table when the expected marginal totals are known. Ann. Math. Statist. 11: 427–444 (1940).
6. Breslow, N. E., and Day, N. E. Statistical Methods in Cancer Research, 1. The Analysis of Case-Control Studies. International Agency for Research on Cancer, Lyon, 1980.
7. Mantel, N., and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22: 719–748 (1959).
8. Tarone, R. E. On heterogeneity tests based on efficient scores. Biometrika 72: 91–95 (1985).
9. Woolf, B. On estimating the relation between blood group and disease. Annu. Eugenics 19: 251–253 (1955).
10. Gart, J. J. The comparison of proportions: a review of significance tests, confidence intervals and adjustments for stratification. Rev. Int. Stat. Inst. 39: 148–169 (1971).
11. Gart, J. J. Point and interval estimation of the common odds ratio in the combination of 2 × 2 tables with fixed marginals. Biometrics 7: 471–475 (1970).
12. Hauck, W. W., Anderson, S., and Leahy, F. J. Finite-sample properties of some old and some new estimators of a common odds ratio from multiple 2 × 2 tables. J. Am. Stat. Assoc. 77(377): 145–152 (1982).