Mantel-Haenszel Test Statistics for Correlated Binary Data

by

Jie Zhang and Dennis D. Boos
Department of Statistics, North Carolina State University
Raleigh, NC 27695-8203
tel: (919) 515-1918
fax: (919) 515-7591
e-mail: jzhang@stat.ncsu.edu, boos@stat.ncsu.edu

Institute of Statistics Mimeo Series No. 2274

August 1995

Mantel-Haenszel Test Statistics for Correlated Binary Data

Summary

This paper proposes two new Mantel-Haenszel test statistics for correlated binary data in 2×2 tables that are asymptotically valid in both sparse data (many strata) and large-strata limiting models. Monte Carlo experiments show that the statistics compare favorably to previously proposed test statistics, especially for 5-25 small to moderate-sized strata. Confidence intervals are also obtained and compared to those obtained from the test of Liang (1985).

1 Introduction

Multicenter randomized clinical trials are frequently used to test the efficacy of new medical regimens. When data from these trials are summarized in a series of 2×2 tables, these data may have the following two characteristics:

1. The number of centers (k) is large, but the number of patients in each center is small. This *sparse* data situation occurs for example when enrollment of large numbers of patients is not possible at individual sites.

2. Dependence exists between observations due to repeated measurement on the same person or to subsampling from clusters such as family units. Such data are often called correlated or clustered binary data.

When k is small and the number of patients in each center is large, hypothesis tests may be based on the generalized estimating equations (GEE) approach of Liang and Zeger(1986), see also Rotnitzky and Jewell(1990), but standard GEE will not be appropriate in the sparse situation 1) above because of the many strata parameters that need to be estimated. On the other hand, when the data are sparse, but the independence assumption is valid, hypothesis tests based on the Mantel-Haenszel and conditional maximum likelihood approach will be appropriate (Breslow, 1981). But as pointed out by Liang (1985), both of these methods will be invalid under the correlation situation 2) above.

Liang (1985) proposed two tests which handle the sparse correlated data situation, but their asymptotic validity depends on $k \to \infty$. More to the point, these tests are not very efficient when k is small since the variance estimates use centers as the primary sampling unit and are thus based on k degrees of freedom. Donald and Donner (1987) also gave an adjusted Mantel-Haenszel test statistic for correlated binary data. Their statistic is based on estimating the intra-class correlation. This test statistic, however, requires an assumption of common correlation across the strata, and further, the estimator of intra-class correlation is not stable in the sparse case.

In this paper two new score tests are proposed which are valid under both asymptotic situations for correlated binary data. Monte Carlo simulations compare these tests

with the standard Mantel-Haenszel test, with one of Liang's (1985) tests, and with the test of Donald and Donner (1987). The simulations show that the new tests perform very favorably in situations with 5-25 strata and small to moderate strata sample sizes. Moreover the simulations confirm the validity of a simple power approximation which makes sample size calculations straightforward. Explicit confidence intervals for the common odds ratio are also derived from one of the new test statistics and from Liang's (1985) statistic. Monte Carlo comparisons are given for these intervals as well.

2 Testing The Null Hypothesis of No Treatment Effect

2.1 Notation and Data Structure

Data from each center (or strata) can be summarized as follows:

	success	failure	total
treatment	$x_{i.}$	$n_{i.} - x_{i.}$	$n_{i.}$
control	$y_{i.}$	$m_{i.} - y_{i.}$	$m_{i.}$
total	$t_{i.}$		N_i

Here $x_{i.} = \sum_{j=1}^{n_i} x_{ij}$, $y_{i.} = \sum_{j=1}^{m_i} y_{ij}$ and $n_{i.} = \sum_{j=1}^{n_i} n_{ij}$, $m_{i.} = \sum_{j=1}^{m_i} m_{ij}$, where n_i and m_i represent the number of patients in the treatment and control groups in center i, n_{ij} and m_{ij} represent the total repeated measurements on one patient or cluster size in the treatment and control groups, respectively, in center i, and x_{ij} and y_{ij} are the associated number of successes.

We assume that all the patient results x_{ij} and y_{ij} are independent within and across the centers and that the expectations of x_{ij} and y_{ij} are equal to $n_{ij}p_{ti}$ and $m_{ij}p_{ci}$, where p_{ti} and p_{ci} are the treatment and control probabilities of success for a single binary trial in the ith center.

The common odds ratio $\Psi = \{p_{ti}/(1-p_{ti})\}/\{p_{ci}/(1-p_{ci})\}$ is the parameter of interest, and we will be testing $H_0 : \Psi = 1$.

2.2 Test Statistics

The Mantel-Haenszel statistic

$$T_{MH} = \frac{\left\{\sum_{i=1}^{k}(x_{i.} - n_{i.}t_{i.}/N_i)\right\}^2}{V_{MH}}$$

has denominator

$$V_{MH} = \sum_{i=1}^{k} \frac{n_{i.}m_{i.}t_{i.}(N_i - t_{i.})}{N_i^2(N_i - 1)}$$

derived from the hypergeometric distribution. The related Cochran (1954) statistic based on the binomial distribution merely replaces $N_i - 1$ by N_i in the denominator of V_{MH}. In either case V_{MH} is too small on average in the presence of positive within-cluster correlations.

Thus Liang (1985) proposed

$$T_L = \frac{\left\{\sum_{i=1}^{k}(x_{i.} - n_{i.}t_{i.}/N_i)\right\}^2}{V_L}$$

with variance estimate

$$V_L = \sum_{i=1}^{k}(x_{i.} - n_{i.}t_{i.}/N_i)^2.$$

Note that T_L has the form of a squared t-statistic $(\sum Z_i)^2/\sum Z_i^2$ except that the null hypothesis assumption $E(Z_i) = 0$ is used in the variance estimate instead of the sample average $\bar{Z}$. Under weak regularity conditions the central limit theorem yields an asymptotic chi-squared distribution with one degree of freedom for T_L under H_0 as $k \to \infty$. If k is fairly small, one might use critical values from the square of a t distribution with k degrees of freedom. However, the fact that the Z_i typically have different variances often makes the chi-squared approximation better. In the simulations of Table 1, the chi-squared approximation was quite adequate and even conservative in places.

Our two new test statistics have the same numerator as those of the Mantel-Haenszel test statistics above but different variance estimators in the denominator. The first one is

$$T_P = \frac{\left\{\sum_{i=1}^{k}(x_{i.} - n_{i.}t_{i.}/N_i)\right\}^2}{V_P},$$

where

$$V_P = \sum_{i=1}^{k}\left\{(1-\lambda_i)^2 \sum_{j=1}^{n_i} \frac{(x_{ij} - n_{ij}\hat{p}_i)^2}{(1 - n_{ij}/N_i)} + \lambda_i^2 \sum_{j=1}^{m_i} \frac{(y_{ij} - m_{ij}\hat{p}_i)^2}{(1 - m_{ij}/N_i)}\right\},$$

with $\lambda_i = n_{i.}/N_i$, $\hat{p}_i = (x_{i.} + y_{i.})/N_i$.

The motivation behind V_P is as follows. To be consistent for correlated data, each of the site components of the variance estimate need to have the form of an empirical variance because we are not modeling the variance as a function of the mean. In addition, in the sparse data case where we are relying on laws of large numbers as $k \to \infty$, it is crucial that the ith component be approximately unbiased in order for the sum of variance estimates over sites to be consistent.

First note that the null expectation of the numerator of T_{MH} is given by

$$\sum_{i=1}^{k} Var(x_{i.} - n_{i.}t_{i.}/N_i) = \sum_{i=1}^{k}\left\{(1-\lambda_i)^2 \sum_{j=1}^{n_i} Var(x_{ij}) + \lambda_i^2 \sum_{j=1}^{m_i} Var(y_{ij})\right\}.$$

If we knew the value of $p_i = p_{ti} = p_{ci}$ under H_0, then $\sum_{j=1}^{n_i}(x_{ij} - n_{ij}p_i)^2$ would be an unbiased estimate of $\sum_{j=1}^{n_i} Var(x_{ij})$. The variance estimate V_P replaces p_i by the pooled estimate $\hat{p}_i = (x_{i.} + y_{i.})/N_i$ and divides by $(1 - n_{ij}/N_i)$ to adjust for this replacement. This adjustment works exactly when $Var(x_{ij}) = n_{ij}c_i$ and $Var(y_{ij}) = m_{ij}c_i$ for positive constants c_i. We summarize this result in the following theorem.

Theorem 1. If the x_{ij} and y_{ij} are all independent with means $E(x_{ij}) = n_{ij}p_i$ and $E(y_{ij}) = m_{ij}p_i$ and variances $Var(x_{ij}) = n_{ij}c_i$ and $Var(y_{ij}) = m_{ij}c_i$ for positive constants $c_1, \ldots, c_k$, then

$$E(V_P) = E\left\{\sum_{i=1}^{k}(x_{i.} - n_{i.}t_{i.}/N_i)\right\}^2.$$

Note that beta-binomial data does not satisfy the assumption of Theorem 1 that $Var(x_{ij}) = n_{ij}c_i$ and $Var(y_{ij}) = m_{ij}c_i$ except when $n_{ij} = m_{ij}$. Nevertheless, simulations

for beta-binomial data in Table 1 show that T_P works well even for unequal cluster sizes.

Our second test statistic has denominator

$$V_U = \sum_{i=1}^{k} \left\{ \frac{(1-\lambda_i)^2}{\delta_{ti}} \sum_{j=1}^{n_i} \frac{(x_{ij} - n_{ij}\hat{p}_{ti})^2}{(1-2n_{ij}/n_{i.})} + \frac{\lambda_i^2}{\delta_{ci}} \sum_{j=1}^{m_i} \frac{(y_{ij} - m_{ij}\hat{p}_{ci})^2}{(1-2m_{ij}/m_{i.})} \right\},$$

with $\hat{p}_{ti} = x_{i.}/n_{i.}, \hat{p}_{ci} = y_{i.}/m_{i.}$. and

$$\delta_{ti} = 1 + \sum_{j=1}^{n_i} \frac{(n_{ij}/n_{i.})^2}{(1-2n_{ij}/n_{i.})}, \quad \delta_{ci} = 1 + \sum_{j=1}^{m_i} \frac{(m_{ij}/m_{i.})^2}{(1-2m_{ij}/m_{i.})}.$$

The key difference between V_P and V_U is that V_P uses the pooled estimator $\hat{p}_i$ whereas V_U uses the unpooled estimators $\hat{p}_{ti}$ and $\hat{p}_{ci}$. V_U seems more complicated than V_P, but it obtains the desired unbiasedness without any assumptions on the form of the variances of x_{ij} and y_{ij}. The adjustments to achieve this unbiasedness follow easily upon noting that

$$E(x_{ij} - n_{ij}\hat{p}_{ti})^2 = Var(x_{ij})(1-2n_{ij}/n_{i.}) + \frac{n_{ij}^2}{n_{i.}^2} \sum_{j=1}^{n_i} Var(x_{ij}). \tag{1}$$

Theorem 2. If the x_{ij} and y_{ij} are all independent with means $E(x_{ij}) = n_{ij}p_{ti}$ and $E(y_{ij}) = m_{ij}p_{ci}$, then

$$E(V_U) = E\left\{ \sum_{i=1}^{k} (x_{i.} - n_{i.}t_{i.}/N_i) \right\}^2.$$

Although V_U is unbiased in general, we have found that V_P is usually preferable because the pooled estimate $\hat{p}_i$ makes V_P more stable than V_U under H_0. T_P has several additional properties as well. When $n_{i.} = m_{i.} = 1$, i.e., in the case of matched pairs, T_P reduces to Liang's T_L. In the case where $n_{ij} = m_{ij} = n_0$, the data can be placed in k separate $2 \times (n_0 + 1)$ contingency tables. If SAS PROC FREQ is used with column scores equal to 0, 1, 2, ..., n_0, and $n_i = m_i$, then the second Cochran-Mantel-Haenszel (CMH) statistic is exactly equal to T_P. This analysis might arise if one decided to analyze just the first n_0 episodes of each patient observed over time. This also shows that when $n_0 = 1$ (pure binary data), T_P reduces to the usual Mantel-Haenszel statistic T_{MH}.

We conclude this section with several theorems on the asymptotic distribution of T_P and T_U. Inherent in the assumptions are that the total number of patients on treatment and control both go to ∞: $\sum_{i=1}^{k} n_i \to \infty$ and $\sum_{i=1}^{k} m_i \to \infty$.

Theorem 3. If the x_{ij} and y_{ij} are all independent with means $E(x_{ij}) = n_{ij}p_i$ and $E(y_{ij}) = m_{ij}p_i$, the sample sizes n_{ij} and m_{ij} are bounded by $N_0 < \infty$, the variances $Var(x_{ij})$ and $Var(y_{ij})$ are bounded below by $c_0 > 0$, and $\sum_{i=1}^{k} \left\{(1-\lambda_i)^2 n_i + \lambda_i^2 m_i\right\} \to \infty$, then

$$\frac{\left\{\sum_{i=1}^{k}(x_{i.} - n_{i.}t_{i.}/N_i)\right\}^2}{\sum_{i=1}^{k}\left\{(1-\lambda_i)^2 \sum_{j=1}^{n_i}(x_{ij} - n_{ij}p_i)^2 + \lambda_i^2 \sum_{j=1}^{m_i}(y_{ij} - m_{ij}p_i)^2\right\}} \xrightarrow{d} \chi_1^2.$$

To get the null asymptotic chi-squared result for T_P and T_U, we need only to show that the ratios of V_P and V_U to the denominator in Theorem 3 converge to 1 in probability. This results in the following theorem.

Theorem 4. Assume that either i) $k \to \infty$ with $1 \le n_i \le c_1 < \infty$ and $1 \le m_i \le c_1 < \infty$ or ii) k remains bounded with $\min(n_i, m_i) \to \infty$ for $i = 1, \ldots, k$. Then under the assumptions of Theorems 1 and 3, $T_P \xrightarrow{d} \chi_1^2$, and under the assumptions of Theorems 2 and 3, $T_U \xrightarrow{d} \chi_1^2$.

Proofs of the above theorems are outlined in the Appendix.

2.3 Power Calculations

Following Wittes and Wallenstein (1987,(2.3)), the power of any of the Mantel-Haenszel tests may be approximated by

$$\Phi\left(\frac{\sum_{i=1}^{k}(n_{i.}m_{i.}/N_i)\Delta_i}{\sqrt{V}} - Z(1-\alpha/2)\right), \tag{2}$$

where Φ is the standard normal distribution function, $\Delta_i = p_{ti} - p_{ci}$ is the difference of success probabilities for the ith center, $Z(1-\alpha/2)$ is the $1-\alpha/2$ quantile of a standard normal, and

$$V = \sum_{i=1}^{k}\left\{(1-\lambda_i)^2 \sum_{j=1}^{n_i} Var(x_{ij}) + \lambda_i^2 \sum_{j=1}^{m_i} Var(y_{ij})\right\}$$

is the variance of the square root of the numerator of the Mantel-Haenszel statistics. For simplicity consider a study with only one binary response per patient, the same alternative $\Delta_i = \Delta$, constant treatment and control probabilities $p_{ti} = p_t$ and $p_{ci} = p_c$, all sample sizes equal with n total patients in the treatment group and also in the placebo group, and $\alpha = .05$. The power approximation is just

$$\Phi\left(\frac{\sqrt{n}\Delta}{\sqrt{p_t(1-p_t)+p_c(1-p_c)}} - 1.96\right).$$

For example, at $\Delta = .5 - .4 = .1$, $n = 400$ patients in each group would give an approximate power of .815. Here of course we would be using the standard Mantel-Haenszel statistic T_{MH}.

Now for comparison consider a study also with equal sample sizes but where each patient has $n_{ij} = m_{ij} = n_0$ repeated episodes, and we make the assumption that the x_{ij} and y_{ij} have beta-binomial(ρ) distributions so that

$$Var(x_{ij}) = n_0 p_t(1-p_t)[1+(n_0-1)\rho].$$

The power approximation is

$$\Phi\left(\frac{\sqrt{n_0}\sqrt{n}\Delta}{\sqrt{[p_t(1-p_t)+p_c(1-p_c)][1+(n_0-1)\rho]}} - 1.96\right). \tag{3}$$

If $n_0 = 10$, $\rho = .3$, and $n = 150$ patients per group, then the approximate power is .820. The values $n_0 = 10$ and $\rho = .3$ are quite realistic as we have seen in several recent clinical trials. This comparison demonstrates the potential savings (300 versus 800 patients) which is possible by using repeated measurements for each patient. Of course the actual power can be a little different with unequal sample sizes and/or unequal (Δ_i, p_{ci}) values, but the comparisons are still similar.

Finally let us replace $p_t(1-p_t)+p_c(1-p_c)$ by $1/2$ in (3), set the power equal to .80, Z(.80)=.84, and solve for n:

$$n = \left(\frac{1}{\Delta}\right)^2\left(\frac{1+(n_0-1)\rho}{2n_0}\right)[.84+1.96]^2. \tag{4}$$

Graphs of the sample size n versus n_0 from (4) show that most of the advantage of taking multiple observations per patient is achieved by $n_0 = 5$ to $n_0 = 10$. Thus in

Table 1: Estimates of size for nominal $\alpha = 0.05$ tests for data from the beta-binomial(ρ) distribution.

		$\rho = 0.0$			$\rho = 0.2$			$\rho = 0.8$		
	$n_{ij}, m_{ij} =$	5	5-10	5-15	5	5-10	5-15	5	5-10	5-15
k=5	T_{MH}	.056	.054	.042	.141	.175	.261	.340	.437	.493
	T_L	.022	.022	.016	.020	.019	.033	.017	.023	.023
	T_{DD}	.046	.038	.032	.051	.050	.066	.044	.049	.043
	T_P	.057	.052	.039	.050	.050	.064	.041	.047	.036
	T_U	.058	.058	.040	.051	.051	.067	.045	.048	.040
k=15	T_{MH}	.047	.065	.032	.154	.215	.238	.325	.432	.482
	T_L	.049	.051	.037	.045	.051	.039	.040	.038	.038
	T_{DD}	.037	.050	.022	.086	.095	.104	.119	.133	.131
	T_P	.045	.060	.032	.045	.057	.049	.054	.047	.052
	T_U	.048	.060	.037	.048	.061	.051	.055	.049	.057
k=25	T_{MH}	.043	.059	.046	.157	.203	.234	.344	.411	.491
	T_L	.047	.045	.058	.047	.052	.048	.045	.040	.043
	T_{DD}	.024	.029	.024	.049	.057	.064	.081	.091	.093
	T_P	.044	.057	.044	.049	.055	.053	.047	.050	.054
	T_U	.042	.059	.043	.047	.053	.052	.049	.048	.053

planning a study one might keep the length of the study relatively short because of the limited value of having n_{ij} and $m_{ij} > 10$.

3 Monte Carlo Study

3.1 Size Comparison

A simulation study was conducted to assess the size of the two new test statistics, T_P and T_U, and to compare to the Mantel-Haenszel test statistic T_{MH}, T_L from Liang (1985), and T_{DD} from Donald and Donner (1987).

The total number of patients or clusters was fixed at 200, and the number of

centers was chosen to be 5, 15, and 25, respectively. Cluster sizes n_{ij} and m_{ij} were fixed at 5 (columns 1,4,7) and allowed to range from 5-10 (columns 2,5,8) and from 5-15 (columns 3,6,9). The control group success probability p_{ci} ranged from 0.2 to 0.8 with increments 0.12 (k=5), 0.04 (k=15), and 0.024 (k=25). The x_{ij} and y_{ij} were generated from the beta-binomial distribution with $\rho = 0$ (binomial), $\rho = 0.2$, and $\rho = 0.8$. A total of 1000 simulated data sets were run for each combination of parameters. SAS IML macros were used for all programming.

From Table 1 we can see that the new tests hold their level very well across all situations, and perform very similarly. The Mantel-Haenszel test is of course far too liberal when $\rho > 0$. Liang's T_L tends to be too conservative when k=5, but approximates the 5% level well when k=15 and k=25. Donald and Donner's T_{DD} has erratic behavior for $k = 15$ and $k = 25$.

3.2 Power Comparison

We use the same set up as for the size comparisons of Table 1, with data generated under the alternative $H_0 : \Psi = 1.5$. Because the nominal size of T_{DD} was not very stable in Table 1, it was not included here. T_{MH} results are only given when $\rho = 0$. The results are summarized in Table 2.

Table 2 shows that when $\rho = 0$, the power of T_P and T_U is almost equal to the power of the Mantel-Haenszel test. When k is small, the power of T_P and T_U is much better than T_L. With the increase in k, the difference in power between T_L and T_P and T_U decreases. However, when k=25, the power of T_P and T_U is still considerably better than the power of T_L.

To check the approximate power expression (2), we calculated from (2) the powers .833 (k=5), .844 (k=15), and .839 (k=25) for the situation of column 1 of Table 2 ($n_{ij} = m_{ij} = 5$, $\rho = 0.0$). Similarly for column 5 of Table 2 ($n_{ij} = m_{ij} = 5 - 10$, $\rho = 0.2$), we obtained powers of .638 (k=5), .645 (k=15), and .636 (k=25). These and similar calculations convince us that (2)-(4) can be confidently used in planning studies with T_P and T_U.

Table 2: Estimates of power when $\alpha = .05$, $\Psi = 1.5$, and the data are beta-binomial(ρ).

		$\rho = 0.0$			$\rho = 0.2$			$\rho = 0.8$		
	$n_{ij}, m_{ij} =$	5	5-10	5-15	5	5-10	5-15	5	5-10	5-15
k=5	T_{MH}	.827	.936	.976						
	T_L	.378	.465	.582	.199	.202	.205	.083	.104	.090
	T_P	.820	.932	.976	.569	.619	.659	.276	.296	.291
	T_U	.825	.936	.977	.570	.624	.660	.280	.304	.291
k=15	T_{MH}	.834	.924	.974						
	T_L	.758	.876	.937	.521	.553	.585	.267	.237	.237
	T_P	.828	.923	.973	.606	.648	.675	.311	.286	.301
	T_U	.836	.926	.973	.603	.647	.684	.317	.301	.271
k=25	T_{MH}	.869	.945	.963						
	T_L	.821	.917	.953	.551	.587	.580	.246	.280	.270
	T_P	.861	.935	.963	.587	.616	.626	.277	.305	.293
	T_U	.865	.941	.963	.596	.618	.630	.281	.309	.300

4 Confidence Intervals for the Common Odds Ratio Ψ

4.1 Two Proposed Confidence Intervals

To derive confidence intervals for the common odds ratio Ψ, we need to generalize T_L and T_U to test $H_0 : \Psi = \Psi_0$. Then we will invert the resulting tests to get confidence intervals for Ψ.

The generalization of Liang's statistic T_L is

$$T_{L,\Psi_0} = \frac{(\sum_{i=1}^{k} u_i)^2}{\sum_{i=1}^{k} u_i^2},$$

where

$$u_i = \frac{x_{i.}(m_{i.} - y_{i.})}{N_i} - \Psi_0 \frac{(n_{i.} - x_{i.})y_{i.}}{N_i}.$$

T_{L,Ψ_0} converges to a χ_1^2 distribution as $k \to \infty$ under $H_0 : \Psi = \Psi_0$.

The generalization of T_U is

$$T_{U,\Psi_0} = \frac{(\sum_{i=1}^{k} u_i)^2}{V_{U,\Psi_0}},$$

where

$$\begin{aligned} V_{U,\Psi_0} &= \sum_{i=1}^{k} N_i^{-2} \left\{ m_i^2 A_i + \Psi_0^2 n_{i.}^2 B_i + \Psi_0^2 n_{i.}^2 B_i + (1-\Psi_0)^2 (A_i B_i + x_{i.}^2 B_i + A_i y_{i.}^2) \right. \\ &\qquad \left. -(1-\Psi_0) m_{i.} A_i y_{i.} + \Psi_0 (1-\Psi_0) n_{i.} B_i x_{i.} \right\}, \end{aligned}$$

and

$$A_i = \frac{1}{\delta_{ti}} \sum_{j=1}^{n_i} \frac{(x_{ij} - n_{ij}\hat{p}_{ti})^2}{(1 - 2n_{ij}/n_{i.})}, \quad B_i = \frac{1}{\delta_{ci}} \sum_{j=1}^{m_i} \frac{(y_{ij} - m_{ij}\hat{p}_{ci})^2}{(1 - 2m_{ij}/m_{i.})}.$$

We have not generalized T_P because the pooled proportion estimators $\hat{p}_i$ do not make sense unless $\Psi_0 = 1$. Similar to the results for V_U, V_{U,Ψ_0} is unbiased for $E \sum_{i=1}^{k} u_i^2$ under $H_0 : \Psi = \Psi_0$, and T_{U,Ψ_0} converges to a χ_1^2 distribution for either $k \to \infty$ or for the number of patients per strata $\to \infty$.

Table 3: Coverage and length for 95% confidence intervals for data from the beta-binomial(ρ) distribution with $\Psi = 1.5$.

			$\rho = 0.0$			$\rho = 0.2$			$\rho = 0.8$		
		$n_{ij}, m_{ij} =$	5	5-10	5-15	5	5-10	5-15	5	5-10	5-15
k=5	C_L	Coverage	.977	.977	.977	.984	.980	.975	.964	.976	.981
		Mean Length	1.93	1.48	1.25	5.15	2.88	3.95	11.6	12.9	10.6
	C_U	Coverage	.969	.964	.968	.977	.968	.961	.947	.957	.951
		Mean Length	0.95	0.77	0.69	1.42	1.29	1.26	3.32	4.34	3.80
k=15	C_L	Coverage	.944	.955	.952	.959	.951	.959	.936	.947	.964
		Mean Length	0.94	0.80	0.69	1.32	1.26	1.21	2.31	2.39	2.50
	C_U	Coverage	.948	.962	.956	.963	.957	.959	.955	.956	.963
		Mean Length	0.94	0.78	0.68	1.35	1.29	1.23	3.01	3.73	3.40
k=25	C_L	Coverage	.957	.949	.956	.959	.962	.954	.961	.959	.960
		Mean Length	0.90	0.77	0.68	1.24	1.16	1.17	2.17	2.25	2.29
	C_U	Coverage	.962	.962	.959	.969	.969	.959	.964	.962	.964
		Mean Length	0.94	0.78	0.71	1.36	1.27	1.26	3.16	3.51	4.72

Thus our confidence intervals are the sets $C_L = \{\Psi : T_{L,\Psi} < \chi^2_1(1-\alpha)\}$ and $C_U = \{\Psi : T_{U,\Psi} < \chi^2_1(1-\alpha)\}$, where $\chi^2_1(1-\alpha)$ is the $(1-\alpha)$ quantile of the χ^2_1 distribution. Some algebra reveals that each of these sets has the form

$$\{\Psi : \frac{a_1\Psi^2 + b_1\Psi + c_1}{a_2\Psi^2 + b_2\Psi + c_2} < \chi^2_1(1-\alpha)\},$$

where the constants $a_1, b_1, c_1, a_2, b_2, c_2$ are given in the Appendix. If $a_1 - a_2\chi^2_1(1-\alpha) > 0$, then the sets are indeed intervals. In practice C_L almost always gives closed intervals as long as $k \geq 15$, and C_U generally gives closed intervals (only 17 out of 27000 data sets used in constructing Table 3 failed to result in closed intervals).

4.2 Monte Carlo Study

We use the same setup as in Table 2 with true odds ratio equal to 1.5. The results are summarized in Table 3.

The simulation results show that both C_L and C_U approximate the nominal 95% converage probability well. When k=5, C_U has much smaller shorter length than C_L. Note also that about 10% of the C_L sets were not closed intervals for $k = 5$ and thus were not included in the Table 3 results.

When k=15, both C_L and C_U have almost the same length. When k=25, C_L has a shorter length than C_U. Keep in mind that for $k = 25$ there are only 8 patients on average in each of the strata. Thus, in contrast to the testing results where T_P and T_U could be recommended in all situations studied, here it appears that the confidence interval from Liang's method can be preferable in the many small strata situation.

5 Example

In a double-blind placebo controlled clinical trial of a new heartburn treatment, patients were randomized to the treatment or control group after a one week run-in phase to determine eligibility. For two weeks these patients then kept diaries of their heartburn episodes and whether the treatment they self-administered was successful.

Table 4 displays the number of patients in each group for the first 17 investigators (sites). The cluster sizes n_{ij} and m_{ij} in the study had a mean close to 10 with a standard deviation close to 5. Instead of using the actual data (which is not available for publication) we generated cluster sizes from the actual distribution of clusters and then generated the x_{ij} and y_{ij} from a beta-binomial distribution with $\rho = .3$ and average $p_{ti} = .5$ and $p_{ci} = .4$ with the odds ratio Ψ held constant at $\Psi = 1.5$. Table 4 displays the number of successes $x_{i.}$ and $y_{i.}$ but not the individual x_{ij} and y_{ij} (because of space limitations). Thus T_{MH} and T_L can be computed from Table 4, but the full data set is required for computing T_P and T_U. We will e-mail the full data set upon request.

The standard Mantel-Haenszel statistic $T_{MH} = 37.53$ is overly optimistic due to the correlation within individual patients. Liang's statistic is $T_L = 8.53$ with p-value .0035, whereas $T_P = 8.64$ with p-value .0033, and $T_U = 8.45$ with p-value .0037. Thus, there is little difference here among the modified Mantel-Haenszel statistics although Table 2 suggests that on average T_P and T_L will be more powerful in this type of

Table 4: Summary Data from Heartburn Clinical Trial

	Treatment Group				Placebo Group			
Site	n_i	$x_{i.}$	$n_{i.}$	$\hat{p}_{ti}$	m_i	$y_{i.}$	$m_{i.}$	$\hat{p}_{ci}$
1	6	23	55	.42	6	21	65	.32
2	16	80	169	.47	16	54	153	.35
3	12	89	143	.62	15	86	152	.57
4	7	36	76	.47	8	18	62	.29
5	15	81	144	.56	12	51	132	.39
6	12	39	104	.38	13	68	176	.39
7	12	45	117	.38	12	50	100	.50
8	7	35	66	.53	6	12	45	.27
9	6	35	65	.54	6	42	68	.62
10	27	122	281	.43	26	111	262	.42
11	21	114	194	.59	23	93	232	.40
12	23	93	204	.46	23	70	226	.31
13	6	40	62	.65	6	20	70	.29
14	21	86	186	.46	24	90	237	.38
15	24	134	230	.58	24	128	244	.52
16	25	119	253	.47	27	95	221	.43
17	15	80	159	.50	16	63	150	.42

situation.

The 95% confidence interval for the odds ratio Ψ from inverting Liang's statistic is (1.18,1.73), whereas the interval obtained from T_U is (1.12, 1.84). Recall that Table 3 shows an advantage for Liang's intervals in terms of average length.

6 Concluding Remarks

We have introduced two new test statistics for correlated binary data which perform well for a wide range of strata sizes and number of strata. They were motivated by randomized clinical trials, but the results can also be applied to cohort and case-control studies.

The statistic T_P which uses pooled estimators in the variance estimate is our preference. It is the same as Liang's (1985) T_L in the case of matched pairs and also reduces to a SAS CMH statistic in the case where the number of episodes per patient is held constant. The statistic T_U which uses unpooled strata proportion estimators in its variance estimate performed very similar to T_P in the Monte Carlo studies. Both T_P and T_L have power advantages over T_L, especially for a small number of strata. In addition, the approximate power calculations of Section 2.3 are easy to use in the practical design of a clinical trial.

For confidence intervals for Ψ the interval based on T_U is much better than that based on T_L when the number of strata is small. The situation is reversed when the number of strata is large with very few patients per strata.

We might add that all the methods discussed in this paper are such that the binary outcomes are equally weighted within strata. Some researchers would prefer that patients be equally weighted instead. One simple approach is to use ANOVA methods on the variables $\hat{p}_{xij} = x_{ij}/n_{ij}$ and $\hat{p}_{yij} = y_{ij}/m_{ij}$. We implemented this approach using the F statistic for treatment with SAS Type III sums of squares in a model with strata, treatment, and strata by treatment interaction (see Searle, 1987) for further explanation. The simulation results are not given because of limited space, but they were very similar to those obtained by T_P and T_U in Tables 1 and 2 except that the power was

slightly smaller on average.

Acknowledgement

We would like to thank El Giefer whose broad experience with analyzing clinical trial data motivated the results of this paper.

Appendix

Proof of Theorem 1. Direct calculations give

$$E(x_{ij} - n_{ij}\hat{p}_i)^2 = Var(x_{ij})(1 - \frac{n_{ij}}{N_i}) - \frac{n_{ij}}{N_i}Var(x_{ij}) + \frac{n_{ij}^2}{N_i^2}\left\{\sum_{j=1}^{n_i} Var(x_{ij}) + \sum_{j=1}^{m_i} Var(y_{ij})\right\}.$$

When we substitute $Var(x_{ij}) = n_{ij}c_i$, the last two terms above cancel. Thus

$$E\sum_{j=1}^{n_i}\frac{(x_{ij} - n_{ij}\hat{p}_i)^2}{(1 - n_{ij}/N_i)} = \sum_{j=1}^{n_i} Var(x_{ij}).$$

Since a similar result holds for $(y_{ij} - m_{ij}\hat{p}_i)^2$, the theorem follows.

Proof of Theorem 2. From (1) we get

$$\begin{aligned} E\sum_{j=1}^{n_i}\left\{\frac{(x_{ij} - n_{ij}\hat{p}_{ti})^2}{(1 - 2n_{ij}/n_{i.})}\right\} &= \left(1 + \sum_{j=1}^{n_i}\left\{\frac{(n_{ij}/n_{i.})^2}{(1 - 2n_{ij}/n_{i.})}\right\}\right)\sum_{j=1}^{n_i} Var(x_{ij}) \\ &= \delta_{ti}\sum_{j=1}^{n_i} Var(x_{ij}). \end{aligned}$$

A similar result holds for $(y_{ij} - m_{ij}\hat{p}_i)^2$, and the theorem follows.

Proof of Theorem 3. The square root of the numerator may be written as a sum of bounded random variables for which it is easy to verify the Lindeberg condition. The key point is that the sum of the variances of these random variables converges to ∞:

$$\sum_{i=1}^{k} Var(x_{i.} - n_{i.}t_{i.}/N_i) \geq \sum_{i=1}^{k}\left\{(1 - \lambda_i)^2 n_i c_0 + \lambda_i^2 m_i c_0\right\} \to \infty.$$

Thus by the central limit theorem $\sum_{i=1}^{k}(x_{i.} - n_{i.}t_{i.}/N_i)$ is asymptotically normal with mean zero and

$$\sum_{i=1}^{k} Var(x_{i.} - n_{i.}t_{i.}/N_i).$$

Now the ratio of the denominator of the expression in Theorem 3 to the variance expression above converges in probability to 1 because the variance of that ratio is bounded by

$$\frac{\sum_{i=1}^{k}\{(1-\lambda_i)^4 n_i N_0^4 + \lambda_i^4 m_i N_0^4\}}{\left(\sum_{i=1}^{k}\left\{(1-\lambda_i)^2 n_i c_0 + \lambda_i^2 m_i c_0\right\}\right)^2} \leq \frac{(N_0^4/c_0^2)}{\sum_{i=1}^{k}\left\{(1-\lambda_i)^2 n_i + \lambda_i^2 m_i\right\}},$$

which easily is seen to converge to 0. Theorem 3 then follows from Slutsky's Theorem.

Proof of Theorem 4. For case i) it is easy to show that the variance of the ratio of V_P and V_U to $\sum_{i=1}^{k} Var(x_{i.} - n_{i.}t_{i.}/N_i)$ is $o(k)$ as $k \to \infty$. Together with the unbiasedness results for V_P and V_U we thus obtain convergence in probability to 1 of these ratios. Theorem 3 and Slutsky's Theorem then yield the asymptotic χ_1^2 results. For case ii) we write V_P as

$$\frac{\sum_{i=1}^{k}\left\{(1-\lambda_i)^2 \sum_{j=1}^{n_i}(x_{ij} - n_{ij}p_i)^2 + \lambda_i^2 \sum_{j=1}^{m_i}(y_{ij} - m_{ij}p_i)^2\right\} + R_x + R_y}{\sum_{i=1}^{k} Var(x_{i.} - n_{i.}t_{i.}/N_i)},$$

where

$$R_x = \sum_{j=1}^{n_i}\left\{\frac{\frac{n_{ij}}{N_i}(x_{ij} - n_{ij}p_i)^2 - 2(x_{ij} - n_{ij}p_i)n_{ij}(\hat{p}_i - p_i) + n_{ij}^2(\hat{p}_i - p_i)^2}{1 - n_{ij}/N_i}\right\},$$

and R_y is similar. The first part of the ratio was shown to converge in probability to 1 in the proof of Theorem 3 above. The pieces involving R_x and R_y can be shown to converge in probability to 0 and the results obtains. A similar proof works for V_U as well.

Confidence interval details. As mentioned in Section 4.1, both C_L and C_U are obtained by solving

$$\frac{a_1\Psi^2 + b_1\Psi + c_1}{a_2\Psi^2 + b_2\Psi + c_2} < \chi_1^2(1-\alpha)\}.$$

For C_L the constants are

$$a_1 = \left\{\sum_{i=1}^{k} \frac{(n_{i.} - x_{i.})y_{i.}}{N_i}\right\}^2, \; b_1 = -2\sum_{i=1}^{k} \frac{(n_{i.} - x_{i.})y_{i.}}{N_i} \sum_{i=1}^{k} \frac{x_{i.}(m_{i.} - y_{i.})}{N_i}, \; c_1 = \left\{\sum_{i=1}^{k} \frac{(m_{i.} - y_{i.})x_{i.}}{N_i}\right\}^2,$$

$$a_2 = \sum_{i=1}^{k} \left\{\frac{(n_{i.} - x_{i.})y_{i.}}{N_i}\right\}^2, \; b_2 = -2\sum_{i=1}^{k} \frac{(n_{i.} - x_{i.})y_{i.}x_{i.}(m_{i.} - y_{i.})}{N_i{}^2}, \; c_2 = \sum_{i=1}^{k} \left\{\frac{(m_{i.} - y_{i.})x_{i.}}{N_i}\right\}^2.$$

For C_U, a_1, b_1, c_1 are the same, but

$$a_2 = \sum_{i=1}^{k}(n_{i.}/N_i)^2 B_i + \sum_{i=1}^{k}(1/N_i)^2(A_iB_i + x_{i.}^2 B_i + A_i y_{i.}^2) - \sum_{i=1}^{k}(n_{i.}/(N_i)^2 B_i x_{i.},$$

$$b_2 = -2\sum_{i=1}^{k}(1/N_i)^2(A_iB_i + x_{i.}^2 B_i + A_i y_{i.}^2) - \sum_{i=1}^{k}(n_{i.}/N_i^2)B_i x_{i.},$$

$$c_2 = \sum_{i=1}^{k}(m_{i.}/N_i)^2 A_i + \sum_{i=1}^{k}(1/N_i)^2(A_iB_i + x_{i.}^2 B_i + A_i y_{i.}^2) - \sum_{i=1}^{k}(m_{i.}/N_i^2)A_i y_{i.},$$

where

$$A_i = \frac{1}{\delta_{ti}} \sum_{j=1}^{n_i} \frac{(x_{ij} - n_{ij}\hat{p}_{ti})^2}{(1 - 2n_{ij}/n_{i.})}, \quad B_i = \frac{1}{\delta_{ci}} \sum_{j=1}^{m_i} \frac{(y_{ij} - m_{ij}\hat{p}_{ci})^2}{(1 - 2m_{ij}/m_{i.})}.$$

REFERENCES

Boos, D.D. (1992). On Generalized Score Tests. *The American Statistician* 46, 327-333.

Boos, D.D. (1993). Analysis of dose-response data in the presence of extrabinomial variation. *Applied Statistics* 42, 173-183.

Breslow, N.E. (1981). Odds ratio estimators when the data are sparse. *Biometrika* 68, 73-84.

Cochran, W. G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics* 10, 417-451.

Donald, A., and Donner, A. (1987). Adjustments to the Mantel-Haenszel chi-square statistics and odds ratio variance estimator when the data are clustered. *Statistics in Medicine* 6, 491-499.

Liang, K.Y. (1985). Odds ratio inference with dependent data. *Biometrika* 72, 678-682.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.

Rotnitsky, A. and Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 77, 485-497.

Searle, S. R. (1987). *Linear Models for Unbalanced Data.* New York: John Wiley & Sons, Inc.

Wittes, J. and Wallenstein, S. (1987). The power of the Mantel-Haenszel test. *Journal of the American Statistical Association* 82, 1104-1109.