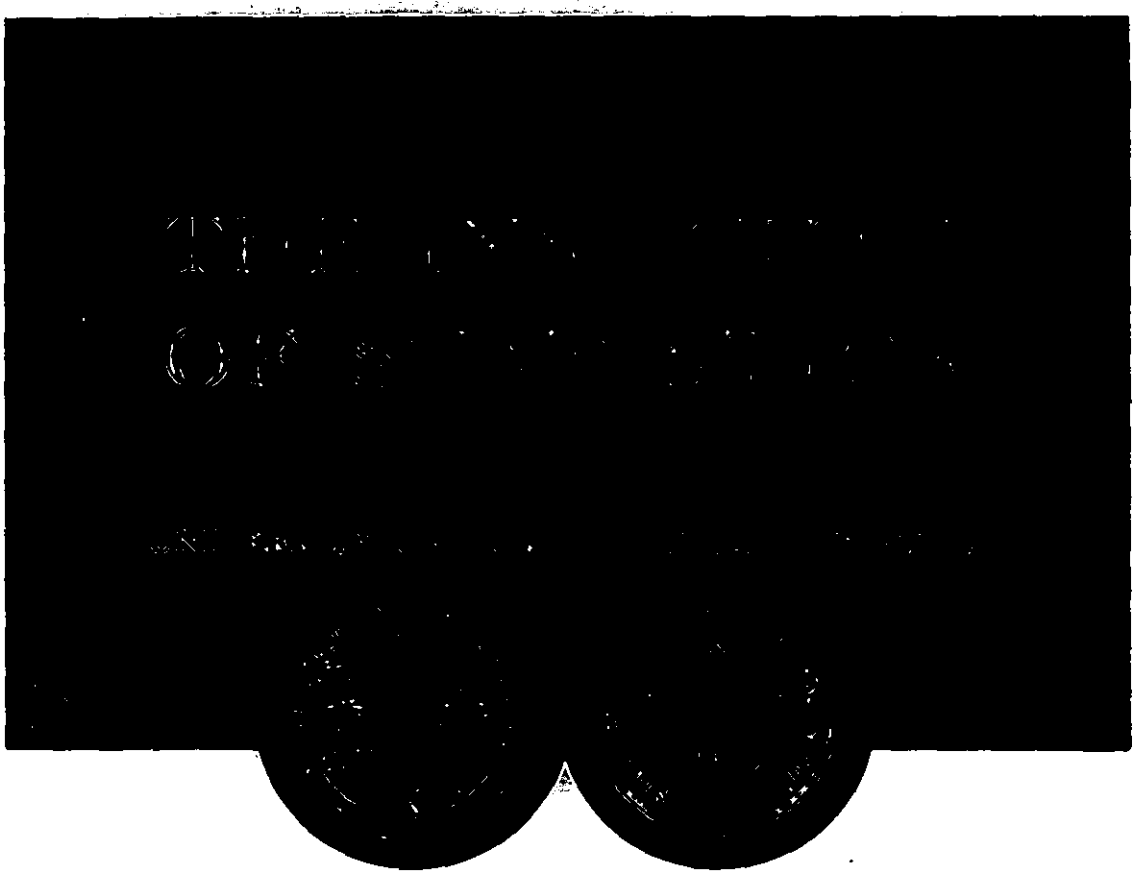


Generalized Cochran-Mantel-Haenszel Test Statistics for Correlated Categorical Data

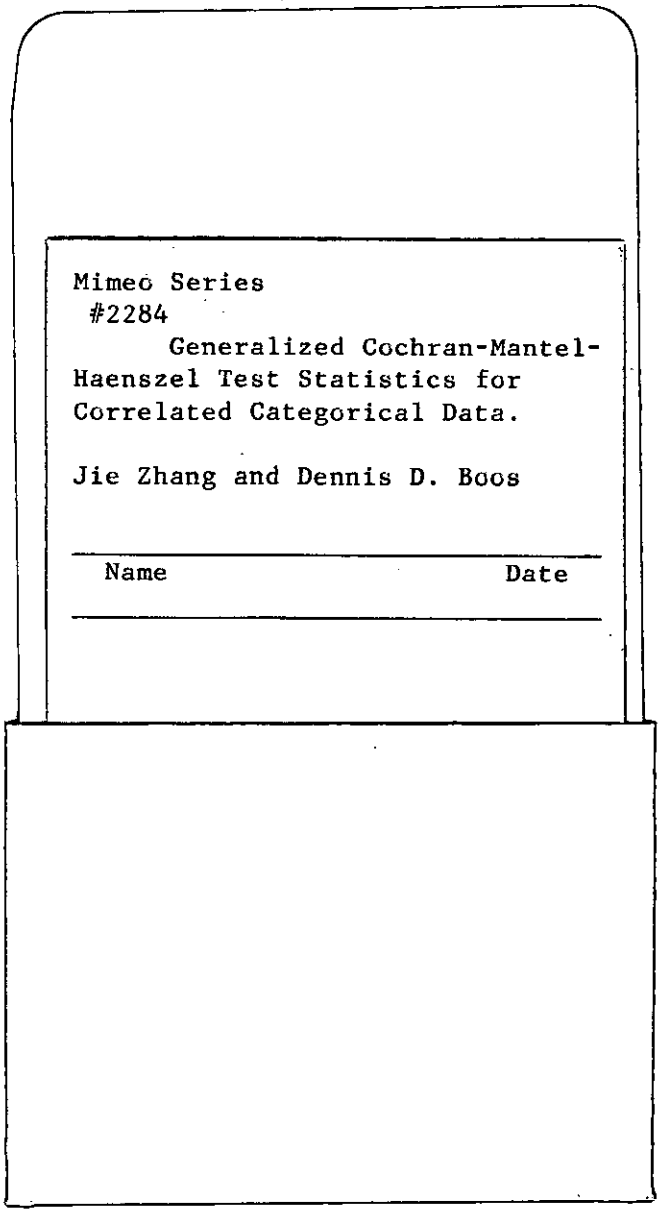
by

Jie Zhang and Dennis D. Boos
Department of Statistics
e-mail: jzhang@stat.ncsu.edu, boos@stat.ncsu.edu

April 1996

Institute of Statistics Mimeo Series No. 2284

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina



Mimeo Series
#2284

Generalized Cochran-Mantel-Haenszel Test Statistics for Correlated Categorical Data.

Jie Zhang and Dennis D. Boos

Name Date

ABSTRACT

Three new test statistics are introduced for correlated categorical data in stratified $R \times C$ tables. They are similar in form to the standard generalized Cochran-Mantel-Haenszel statistics but modified to handle correlated outcomes. Two of these statistics are asymptotically valid in both many-strata (*sparse* data) and large-strata limiting models. The third one is designed specifically for the many-strata case but is valid even with a small number of strata. This latter statistic is also appropriate when strata are assumed to be random.

Key words: correlated categorical data, generalized Cochran-Mantel-Haenszel statistics, sparse data.

1 INTRODUCTION

In multicenter clinical trials, the responses are often recorded on a discrete scale, such as stages of disease severity or levels of improvement following an intervention. In addition, the categorical responses may be correlated because of repeated or multiple measurements on each individual or subsampling from clusters such as family units. Such data are often called correlated categorical data. Moreover, the subjects in the trials are usually randomized to two or more treatment groups such as the different doses of an active ingredient, and one primary objective of the trials is to study the treatment effect.

Sometimes the data from the above type trials are sparse, i.e., the number of centers (q) is large, but the number of patients in many centers is small. This many-strata (*sparse* data) situation occurs for example when enrollment of large number of patients is not possible at each individual center. Furthermore, the *sparse* situation will become more serious in the data analysis when adjustment for other prognostic factors is necessary. Under the many-strata (*sparse* data) situation, since the total sample size increases with the number of nuisance parameters (here due to center effects and prognostic factors), the standard generalized estimating equations (GEE) approach and the standard likelihood method for correlated categorical data will fail; see Liang and Zeger (1995) for further explanation. Similarly the general weighted least squares (WLS) method (Koch, *et al.*, 1977) for analysis of correlated categorical data will be invalid in the many-strata (*sparse* data) situation.

In contrast to the above approaches, generalized Cochran-Mantel-Haenszel (CMH) statistics are robust to the *sparse* situation. However, the assumption of an underlying multiple hypergeometric distribution is not satisfied when there are correlated categorical responses. Although data can sometimes be analyzed by using CMH statistics separately at each time point, combining the results at each occasion to get an overall conclusion is difficult. Furthermore, when there are irregular time points for different subjects or the data are from subsampling from clusters, this approach is not available.

Liang (1985) proposed one score test which handles the sparse correlated binary data, but the asymptotic validity depends on the number of strata $q \to \infty$. Zhang and Boos (1995) proposed two score tests for correlated binary data that are asymptotically valid in both many-strata (*sparse* data, $q \to \infty$) and large-strata limiting models (Robins, Breslow

and Greenland 1986). In this paper, we extend these approaches to correlated categorical data and propose three new tests. In Section 2, the three test statistics are described, and Monte Carlo studies are presented. Power calculations based on the test statistics are given in Section 3. Section 4 contains a real example followed by a summary discussion in Section 5.

2 TESTING THE NULL HYPOTHESIS OF NO TREATMENT EFFECT

2.1 Data Structure and Basic Questions

The data structure for the hth stratum is shown in Table 1, where each row is one subject's data. Thus x_{hijk} denotes the number of times the kth individual in the ith treatment level of the hth stratum received a response of level j. R is the number of treatment levels, C is the number of response categories, n_{hik} is the number of repeated measurements (or cluster size) of the kth individual, and n_{hi} is the number of subjects in the ith treatment level of the hth stratum.

Table 1. Data Structure in the hth Stratum

Treatment	Response Variable Categories (j)						
Levels (i)	1	2	·	j	·	C	Total
1	x_{h111}	x_{h121}	·	x_{h1j1}	·	x_{h1C1}	n_{h11}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	x_{hi11}	x_{hi21}	·	x_{hij1}	·	x_{hiC1}	n_{hi1}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	x_{hi1k}	x_{hi2k}	·	x_{hijk}	·	x_{hiCk}	n_{hik}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	$x_{hi1n_{hi}}$	$x_{hi2n_{hi}}$	·	$x_{hijn_{hi}}$	·	$x_{hiCn_{hi}}$	$n_{hin_{hi}}$
R	x_{hR11}	x_{hR21}	·	x_{hRj1}	·	x_{hRC1}	n_{hR1}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

For illustration, we list in Table 2 part of the real data from Section 4. In this example, $R = 3$ (placebo, low dose, and high dose), $C = 3$ (1 = "no improvement," 2 = "some improvement," and 3 = "marked improvement"). For example, row 1 represents the results for patient 1 from the placebo group of stratum 2. That patient was scored three times as "no improvement" and once as "some improvement" during the four follow-up visits.

Table 2. Data Structure for the Example in Section 4

Patients	Treatment	Score 1	Score 2	Score 3	Number of visits
1	Placebo	$x_{2111} = 3$	$x_{2121} = 1$	$x_{2131} = 0$	$n_{211} = 4$
⋮	⋮	⋮	⋮	⋮	⋮
6	low dose	$x_{2211} = 3$	$x_{2221} = 1$	$x_{2231} = 0$	$n_{221} = 4$
⋮	⋮	⋮	⋮	⋮	⋮
12	high dose	$x_{2311} = 0$	$x_{2321} = 4$	$x_{2331} = 0$	$n_{231} = 4$
⋮	⋮	⋮	⋮	⋮	⋮

If let $\pi_{hi*} = (\pi_{hi1}, \pi_{hi2}, \ldots, \pi_{hiC})'$, where π_{hij} is the probability that a single multinomial response is in the jth category for the ith treatment level and the hth stratum, then a single row of Table 1 $\boldsymbol{x}_{hi*k} = (x_{hi1k}, x_{hi2k}, \ldots, x_{hiCk})'$ has a correlated multinomial distribution with parameters π_{hi*} , n_{hik}, and covariance matrix $\boldsymbol{\Sigma}_{hi}$.

Let $\boldsymbol{x}_{hi*\cdot} = (x_{hi1\cdot}, x_{hi2\cdot}, \ldots, x_{hiC\cdot})'$ denote the sum of $\boldsymbol{x}_{hi*k}$ over k, then data in the hth stratum can be summarized as in Table 3. Further define $\boldsymbol{x}_h = (\boldsymbol{x}_{h1*\cdot}', \boldsymbol{x}_{h2*\cdot}', \ldots, \boldsymbol{x}_{hR*\cdot}')'$, and $\boldsymbol{m}_h = N_h(\boldsymbol{p}_{h\cdot*} \otimes \boldsymbol{p}_{h*\cdot})$ with $\boldsymbol{p}_{h*\cdot} = (p_{h1\cdot}, p_{h2\cdot}, \ldots, p_{hR\cdot})'$, and $\boldsymbol{p}_{h\cdot*} = (p_{h\cdot1}, p_{h\cdot2}, \ldots, p_{h\cdot C})'$. Here $p_{hi\cdot} = n_{hi\cdot}/N_h$, $p_{h\cdot j} = t_{hj}/N_h$, and $\otimes$ denotes the Kronecker product multiplication, the matrix on the left of $\otimes$ being multiplied by each element in the matrix on the right.

Table 3. Data Structure in the hth Stratum

Treatment levels (i)	Response Variable Categories (j) 1	2	·	j	·	C	Total
1	$x_{h11\cdot}$	$x_{h12\cdot}$	·	$x_{h1j\cdot}$	·	$x_{h1C\cdot}$	$n_{h1\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
R	$x_{hR1\cdot}$	$x_{hR2\cdot}$	·	$x_{hRj\cdot}$	·	$x_{hRC\cdot}$	$n_{hR\cdot}$
Total	t_{h1}	t_{h2}	·	t_{hj}	·	t_{hC}	N_h

In addition, we assume that the $\{\boldsymbol{x}_{hi*k}\}$ are independent from each other within and across the strata, and the expectation of $\boldsymbol{x}_{hi*k}$ is equal to $n_{hik}\pi_{hi*}$.

The overall null hypothesis of no treatment effect can be described as:

$$H_0\text{: } \pi_{h1*} = \pi_{h2*} = \cdots = \pi_{hR*} \text{ , for } h = 1, 2, \ldots, q.$$

Three alternative hypotheses of practical interest are briefly described in the following. A detailed discussion can be found in Landis *et al.* (1978) and Kuritz *et al.* (1988).

1. Linear Trend in Mean Responses: When the responses and treatment levels are both on ordinal scales, a consistent positive (or negative) association between the response variables and the treatment levels in each stratum is especially interesting.
2. Mean Responses Differ: In situations involving ordinal responses, we are interested in the extent to which measures of location, reflected often as average responses, differ across the treatment levels.
3. General Association: In the most general case, we are interested in the extent to which H_0 can be rejected in favor of the distribution of the response variables differing in nonspecific patterns across levels of the row factor adjusted for the strata. Here the levels of both the response variables and the treatments are treated as nominal scale regardless of any possible underlying ordinal categories.

2.2 Test Statistics

The standard generalized Cochran-Mantel-Haenszel statistic (Landis *et al.*, 1978) is defined as:

$$T_{CMH} = \boldsymbol{G}' \boldsymbol{V}_{CMH}^{-1} \boldsymbol{G},$$

where $\boldsymbol{G} = \sum_{h=1}^{q} \boldsymbol{G}_h = \sum_{h=1}^{q} \boldsymbol{B}_h(\boldsymbol{x}_h - \boldsymbol{m}_h)$ and $\boldsymbol{V}_{CMH} = \sum_{h=1}^{q} \boldsymbol{B}_h \boldsymbol{V}_{Gh} \boldsymbol{B}_h'$, with $\boldsymbol{B}_h = \boldsymbol{C}_h \otimes \boldsymbol{R}_h$ and $\boldsymbol{V}_{Gh} = N_h^2/(N_h - 1)\{(\boldsymbol{D}_{p_{h\cdot *}} - \boldsymbol{p}_{h\cdot *}\boldsymbol{p}_{h\cdot *}') \otimes (\boldsymbol{D}_{p_{h*\cdot}} - \boldsymbol{p}_{h*\cdot}\boldsymbol{p}_{h*\cdot}')\}$.

Here $\boldsymbol{D}_a$ is a diagonal matrix with elements of $\boldsymbol{a}$ on its main diagonal, and $\boldsymbol{C}_h$ and $\boldsymbol{R}_h$ are the matrices defined according to the above alternative hypotheses. Typical choices of $\boldsymbol{R}_h$ and $\boldsymbol{C}_h$ are briefly described in the following. When we are interested in the alternative hypothesis of "general association," $\boldsymbol{R}_h = [\boldsymbol{I}_{R-1}, -\boldsymbol{J}_{R-1}]$ and $\boldsymbol{C}_h = [\boldsymbol{I}_{C-1}, -\boldsymbol{J}_{C-1}]$, where $\boldsymbol{I}_{R-1}$ is an identity matrix of rank $R-1$, and $\boldsymbol{J}_{R-1}$ is an $(R-1) \times 1$ vector of ones. When we are interested in the "mean responses differ," $\boldsymbol{R}_h$ is the same as the one used for the "general association" and $\boldsymbol{C}_h = (c_{h1}, ..., c_{hC})$, where c_{hj} is an appropriate score reflecting the ordinal nature of the jth category of response for the hth stratum. When we are interested in the "linear trend in mean responses," $\boldsymbol{C}_h$ can be defined as the same as the one for the "mean responses differ" and $\boldsymbol{R}_h = (r_{h1}, ..., r_{hR})$, where r_{hi} is an appropriate score reflecting the ordinal nature of the ith level of treatment for the hth stratum. The choice of $\boldsymbol{C}_h$ and $\boldsymbol{R}_h$

will not be discussed further here, but a more detailed summary can be found in Landis *et al.* (1978).

Under the assumption of independence between observations, T_{CMH} is approximately distributed as a chi-squared distribution with degrees of freedom equal to the rank of $\boldsymbol{B}_h$ under the null H_0. However, $\boldsymbol{V}_{CMH}$ will be invalid in the presence of positive within-cluster correlations.

Our three new test statistics have exactly the same form as the standard generalized Cochran-Mantel-Haenszel statistics above but different covariance matrix estimators in place of $\boldsymbol{V}_{CMH}$.

The first statistic is a generalization of the statistic of Liang (1985):

$$T_{EL} = \boldsymbol{G}' \boldsymbol{V}_{EL}^{-1} \boldsymbol{G},$$

where $\boldsymbol{V}_{EL} = \frac{q}{q-1} \sum_{h=1}^{q} (\boldsymbol{G}_h - \overline{\boldsymbol{G}})(\boldsymbol{G}_h - \overline{\boldsymbol{G}})'$ with $\overline{\boldsymbol{G}} = \sum_{h=1}^{q} \boldsymbol{G}_h / q$.

Note that a direct generalization of Liang's (1985) statistic would give us:

$$T_L = \boldsymbol{G}' \boldsymbol{V}_L^{-1} \boldsymbol{G} \quad \text{with} \quad \boldsymbol{V}_L = \sum_{h=1}^{q} \boldsymbol{G}_h \boldsymbol{G}_h'.$$

The reason for using T_{EL} instead of T_L is as follows. Under weak regularity conditions, both T_{EL} and T_L have asymptotic chi-squared distribution with degrees of freedom equal to the rank of $\boldsymbol{B}_h$ under H_0 as $q \to \infty$. However, when q is small to moderate, the chi-squared approximation is not adequate, and we have found from simulations that T_L is very conservative and T_{EL} is very liberal. But since T_{EL} is a monotone transformation of T_L which has the form of Hotelling's T^2 statistic, we can use the Hotelling's T^2 distribution to improve the approximation when q is small to moderate. Simulations show that the resulting true Type 1 error rates are adequate for $q = 8$ and quite good for $q \geq 16$.

One drawback of T_{EL} is that since the variance estimator uses the strata as the primary sampling units, the efficiency of T_{EL} is affected. Of course, when treating the stratum effects to be random (see, for example, Boos and Brownie 1992, and Beitler and Landis 1985), T_{EL} is a natural test statistic.

Our next two statistics are the direct extension of the two statistics proposed by Zhang and Boos (1995):

$$T_P = \boldsymbol{G}' \boldsymbol{V}_P^{-1} \boldsymbol{G} \quad \text{and} \quad T_U = \boldsymbol{G}' \boldsymbol{V}_U^{-1} \boldsymbol{G},$$

where $\boldsymbol{V}_P = \sum_{h=1}^{q} [\boldsymbol{B}_h \boldsymbol{V}_{Ph} \boldsymbol{B}_h{}']$ and $\boldsymbol{V}_U = \sum_{h=1}^{q} \{\boldsymbol{B}_h \boldsymbol{V}_{Uh} \boldsymbol{B}_h{}'\}$ with

$$\boldsymbol{V}_{Ph} = \sum_{i=1}^{R} \{\boldsymbol{A}_{hi} \{ \sum_{k=1}^{n_{hi}} [(\boldsymbol{x}_{hi*k} - n_{hik} \widehat{\boldsymbol{\pi}}_h)(\boldsymbol{x}_{hi*k} - n_{hik} \widehat{\boldsymbol{\pi}}_h)' / (1 - n_{hik}/N_h)]\} \boldsymbol{A}_{hi}{}'\},$$

and

$$\boldsymbol{V}_{Uh} = \sum_{i=1}^{R} \{\boldsymbol{A}_{hi} \{ \frac{1}{\delta_{hi}} \sum_{k=1}^{n_{hi}} [(\boldsymbol{x}_{hi*k} - n_{hik} \widehat{\boldsymbol{\pi}}_{hi})(\boldsymbol{x}_{hi*k} - n_{hik} \widehat{\boldsymbol{\pi}}_{hi})' / (1 - 2n_{hik}/n_{hi\cdot})]\} \boldsymbol{A}_{hi}{}'\}.$$

In the above, $\widehat{\boldsymbol{\pi}}_h = (t_{h1}/N_h, \ldots, t_{hC}/N_h)'$, $\widehat{\boldsymbol{\pi}}_{hi} = (x_{hi1\cdot}/n_{hi\cdot}, \ldots, x_{hiC\cdot}/n_{hi\cdot})'$,

$$\delta_{hi} = 1 + \sum_{k=1}^{n_{hi}} \left\{ (n_{hik}^2/n_{hi\cdot}^2)/(1 - 2n_{hik}/n_{hi\cdot}) \right\},$$

and $\boldsymbol{A}_{hi} = \boldsymbol{I}_C \otimes \boldsymbol{\Lambda}_{hi}$ with

$$\begin{aligned}
\boldsymbol{\Lambda}_{h1} &= (\lambda_{h1}^*, \ldots, \lambda_{h(i-1)}, \lambda_{hi}, \lambda_{h(i+1)}, \ldots, \lambda_{hR})', \\
\cdots &= \cdots, \\
\boldsymbol{\Lambda}_{hi} &= (\lambda_{h1}, \ldots, \lambda_{h(i-1)}, \lambda_{hi}^*, \lambda_{h(i+1)}, \ldots, \lambda_{hR})', \\
\cdots &= \cdots, \\
\boldsymbol{\Lambda}_{hR} &= (\lambda_{h1}, \ldots, \lambda_{h(i-1)}, \lambda_{hi}, \lambda_{h(i+1)}, \ldots, \lambda_{hR}^*)',
\end{aligned}$$

here $\lambda_{hi} = -n_{hi\cdot}/N_h$ and $\lambda_{hi}^* = 1 - n_{hi\cdot}/N_h$.

The motivation behind V_P and V_U is as follows. In order to estimate $Var(\boldsymbol{G})$ consistently for correlated data, the variance estimators for each stratum 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 many-strata (*sparse* data) case where we are relying on laws of large numbers as $q \to \infty$, it is crucial that the hth component be unbiased or approximately unbiased in order for the sum of variance estimators over the strata to be consistent.

Note that the variance of $\boldsymbol{G}$ can be written as

$$\sum_{h=1}^{q} Var\{B_h(\boldsymbol{x}_h - \boldsymbol{m}_h)\} = \sum_{h=1}^{q}\{B_h\{\sum_{i=1}^{R}\{A_{hi}[\sum_{k=1}^{n_{hi}} Var(\boldsymbol{x}_{hi*k})]A_{hi}'\}\}B_h'\}.$$

If we knew the value of $\pi_h = \pi_{hi}$ under H_0 for $i = 1, 2, \cdots, R$, then $\sum_{k=1}^{n_{hi}}(\boldsymbol{x}_{hi*k} - n_{hik}\pi_h)(\boldsymbol{x}_{hi*k} - n_{hik}\pi_h)'$ would be an unbiased estimator of $\sum_{k=1}^{n_{hi}} Var(\boldsymbol{x}_{hi*k})$. In practice, π_h is never known. Therefore, we need to replace π_h by an estimator and at the same time to adjust for the replacement to ensure the approximate unbiasedness of the empirical variances. In V_P, we replace π_h by the pooled estimator $\widehat{\pi}_h$ and divide by $(1 - n_{hik}/N_h)$ to adjust for this replacement. This adjustment works exactly when $Var(\boldsymbol{x}_{hi*k}) = n_{hik}\boldsymbol{\Sigma}_h$ for some positive definite matrix $\boldsymbol{\Sigma}_h$.

In V_U, we replace π_h by the unpooled estimator $\widehat{\pi}_{hi}$ and adjust with the factors δ_{hi} and $1 - 2n_{hik}/n_{hi\cdot}$. The motivation for this adjustment can be seen by noting that

$$\begin{aligned} &E(\boldsymbol{x}_{hi*k} - n_{hik}\widehat{\pi}_{hi})(\boldsymbol{x}_{hi*k} - n_{hik}\widehat{\pi}_{hi})' \\ &\quad = Var(\boldsymbol{x}_{hi*k})(1 - 2n_{hik}/n_{hi\cdot}) + \frac{n_{hik}^2}{n_{hi\cdot}^2}\sum_{k=1}^{n_{hi}} Var(\boldsymbol{x}_{hi*k}). \end{aligned} \tag{1}$$

Though the adjustment in V_U is more complicated than that in V_P, it obtains the desired unbiasedness without any assumptions on the form of the variances of $\{\boldsymbol{x}_{hi*k}\}$.

We summarize the above results in the following theorems.

Theorem 1 If the $\boldsymbol{x}_{hi*k}$'s are all independent of each other with mean $E(\boldsymbol{x}_{hi*k}) = n_{hik}\ \pi_h$ and covariance matrix $Var(\boldsymbol{x}_{hi*k}) = n_{hik}\boldsymbol{\Sigma}_h$ for $h = 1, \ldots, q$, $i = 1, \ldots, R$, $k = 1, \ldots, n_{hi}$, and some positive definite matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_q$, then $E(\boldsymbol{V}_P) = Var(\boldsymbol{G})$.

Theorem 2 If the $\{\boldsymbol{x}_{hi*k}\}$ are all independent of each other with mean $E(\boldsymbol{x}_{hi*k}) = n_{hik}\pi_{hi*}$, then $E(\boldsymbol{V}_U) = Var(\boldsymbol{G})$.

Although V_U is unbiased in general, we have found that V_P is usually preferable because the pooled estimate $\widehat{\pi}_h$ makes V_P more stable then V_U under H_0. Also, the adjustment factor $(1 - 2n_{hik}/n_{hi\cdot})$ for V_U may be negative or zero when n_{hi} is less than 3, though it seldom happens in practice.

Both T_P and T_U have asymptotic chi-squared distributions with degrees of freedom equal to the rank of B_h as long as the total number of subjects goes to ∞. Furthermore, since the covariance matrix estimators of T_P and T_U use the individual subject as the primary

sampling unit, the power of T_P and T_U will be considerably better than that of T_{EL}, especially for the case of small or medium q.

We conclude this section with the following theorem on the asymptotic distributions of T_P and T_U. Inherent in the assumptions are that the total number of subjects $n = \sum_{h=1}^{q} \sum_{i=1}^{R} n_{hi} \to \infty$.

Theorem 3 If the $\{\boldsymbol{x}_{hi*k}\}$ are all independent, the cluster sizes $\{n_{hik}\}$ are bounded by $N_0 < \infty$, the elements of $\boldsymbol{C}_h$ and $\boldsymbol{R}_h$ are bounded in absolute value by some constant, and $Var(\boldsymbol{G})/n \to \boldsymbol{\Sigma}$ as $n = \sum_{h=1}^{q} \sum_{i=1}^{R} n_{hi} \to \infty$, then 1) under the assumptions of Theorem 1, $T_P = \boldsymbol{G}'\boldsymbol{V}_P^{-1}\boldsymbol{G} \xrightarrow{d} \chi^2_{df}$, and 2) under the assumptions of Theorem 2, $T_U = \boldsymbol{G}'\boldsymbol{V}_U^{-1}\boldsymbol{G} \xrightarrow{d} \chi^2_{df}$, where df is the rank of $\boldsymbol{B}_h$.

Proofs of Theorems 1-3 are outlined in the Appendix.

2.3 Monte Carlo Study

In this section we describe a simulation study conducted to study both the size and the power of the three new test statistics, T_{EL}, T_P, and T_U, and to compare to the standard generalized Cochran-Mantel-Haenszel test statistic T_{CMH}. In the following, $F(df, q - df)$ will denote a F distribution with degrees of freedom df and $q - df$, where df is the rank of $\boldsymbol{B}_h$.

2.3.1 Size of the Tests

The total number of subjects in the simulations was fixed at 384, the numbers of response categories (C) and treatment levels (R) were both set equal to 3, and the number of strata (q) was chosen to be 8, 16, and 32, respectively. The number of repeated measurements or cluster size n_{hik} was fixed at 4 in some runs and allowed to range from 4 to 8 in others. $\boldsymbol{x}_{hi*k}$ was generated from the Dirichlet-Multinomial distribution (n_{hik}, $\pi_{hi} = \pi_h$, ρ) with ρ = 0 (multinomial distribution), ρ = 0.2, and ρ = 0.8. Here, ρ is the intra-class correlation coefficient defined in Brier (1980). To make the simulations more realistic, when q = 16, we specified π_h similar to the estimated proportions for each response category in the control group from a real clinical trial, and we chose the sample size in each stratum nearly the same as that in the real trial. For q = 8, we combined some strata, and for q = 32, we divided some strata. The real trial will be discussed in Section 4. The parameters used in the simulations are summarized in the Appendix in Tables A1, A2, and A3 for q = 8, q = 16,

and $q = 32$, respectively. A total of 1000 simulated data sets were run for each combination of parameters. SAS IML was used for all programming. Since the results are quite similar for $q = 16$ and $q = 32$, we only give results for $q = 8$ and $q = 32$ in Tables 4 and 5.

The Dirichlet-Multinomial deviates $\boldsymbol{x}_{hi*k} = (x_{hi1k}, x_{hi2k}, ..., x_{hiCk})'$ are generated in the following two steps. In the first step, we generate independent gamma random variables G_{hijk} with the shape parameter $\pi_{hij}(1-\rho)/\rho$ $(j = 1, 2, ..., C)$. Then, $(p_{hi1k}, p_{hi2k}, ..., p_{hiCk})$ with $p_{hijk} = G_{hijk}/\sum_{j=1}^{C} G_{hijk}$ has a Dirichlet distribution. In the second step, we generate $\boldsymbol{x}_{hi*k} = (x_{hi1k}, x_{hi2k}, ..., x_{hiCk})'$ from a multinomial distribution with parameters $(p_{hi1k}, p_{hi2k}, ..., p_{hiCk})$ and n_{hik}.

One way to explain ρ is as follows. Let $\boldsymbol{x}_{hi*k} = \sum_t \boldsymbol{z}_{hi*kt}$, where the t corresponds to the time point in the repeated measurements, and $\boldsymbol{z}_{hi*kt} = (z_{hi1kt}, z_{hi2kt}, ..., z_{hiCkt})$ has the form (0, 1, ..., 0) with only one element equals to 1 and the others equal to 0. Then the correlation between the corresponding elements in $\boldsymbol{z}_{hi*kt}$ and $\boldsymbol{z}_{hi*kt'}$ is equal to ρ; for example, $corr(z_{hi1kt}, z_{hi1kt'}) = \rho$.

From Tables 4, and 5, we can see that both T_P and T_U hold their 5% level very well across all situations, and perform very similarly. Using the Hotelling T^2 distribution, T_{EL} is a little conservative at $q = 8$ but very good by $q = 32$. T_{CMH} is of course far too liberal when $\rho > 0$. Note that the Dirichlet-Multinomial distribution does not satisfy the assumption of Theorem 1 that $Var(\boldsymbol{x}_{hi*k}) = n_{hik}\Sigma_h$ except when $n_{hik} = n_{h0}$ for different i and k. Nevertheless, simulations show that T_P works well even for unequal cluster sizes.

Table 4. Estimates of Size for Nominal $\alpha = 0.05$ Tests for Data from the Dirichlet-Multinomial (ρ) Distribution for $q = 8$

		$\rho = 0.0$		$\rho = 0.2$		$\rho = 0.8$	
Alternative	$n_{hik} =$	4	4-8	4	4-8	4	4-8
	T_{CMH}	.041	.038	.118	.206	.310	.436
Linear	T_{EL}	.039	.040	.037	.044	.052	.045
Trend	T_P	.046	.038	.046	.065	.058	.053
	T_U	.048	.039	.046	.064	.061	.054
	T_{CMH}	.042	.041	.149	.278	.420	.640
Means	T_{EL}	.035	.024	.031	.026	.027	.023
Differ	T_P	.040	.039	.052	.056	.055	.048
	T_U	.043	.044	.055	.059	.060	.053
	T_{CMH}	.048	.042	.193	.380	.589	.820
General	T_{EL}	.019	.019	.011	.017	.017	.016
Association	T_P	.045	.039	.045	.055	.051	.044
	T_U	.058	.048	.050	.062	.061	.055

T_{EL}: $(q - df)/(df(q-1))T_{EL}$ compare to a $F(df, q - df)$ distribution.

Table 5. Estimates of Size for Nominal $\alpha = 0.05$ Tests for Data from the Dirichlet-Multinomial (ρ) Distribution for $q = 32$

		$\rho = 0.0$		$\rho = 0.2$		$\rho = 0.8$	
Alternative	$n_{hik} =$	4	4-8	4	4-8	4	4-8
	T_{CMH}	.058	.046	.120	.197	.304	.470
Linear	T_{EL}	.048	.045	.054	.051	.057	.049
Trend	T_P	.056	.046	.052	.055	.057	.048
	T_U	.056	.051	.053	.056	.058	.050
	T_{CMH}	.048	.044	.174	.273	.460	.644
Means	T_{EL}	.053	.050	.051	.059	.049	.041
Differ	T_P	.049	.047	.048	.060	.051	.048
	T_U	.052	.049	.054	.066	.056	.052
	T_{CMH}	.055	.041	.215	.392	.645	.840
General	T_{EL}	.040	.047	.051	.050	.063	.046
Association	T_P	.048	.038	.047	.058	.057	.051
	T_U	.059	.049	.056	.066	.067	.059

T_{EL}: $(q - df)/(df(q-1))T_{EL}$ compare to a $F(df, q - df)$ distribution.

2.3.2 Power of the Tests

We use the same setup as that in Tables A1-A3. The alternative probabilities are defined as follows. The probabilities $\pi_{h1*} = (\pi_{h11}, \pi_{h12}, \pi_{h13})'$ in the control group are the same as those in Tables A1-A3. The probabilities $\pi_{h2*} = (\pi_{h21}, \pi_{h22}, \pi_{h23})'$ in the treatment

level 1 group are:

$$\pi_{h21} = \pi_{h11} - 0.08, \quad \pi_{h22} = \pi_{h12} + 0.05, \quad \pi_{h23} = \pi_{h13} + 0.03.$$

The probabilities $\pi_{h3*} = (\pi_{h31}, \pi_{h32}, \pi_{h33})'$ in the treatment level 2 group are:

$$\pi_{h31} = \pi_{h11} - 0.12, \quad \pi_{h32} = \pi_{h12} + 0.07, \quad \pi_{h33} = \pi_{h13} + 0.05.$$

The T_{CMH} results are only given when $\rho = 0$. The results are summarized in Tables 6-8.

These tables show that when $\rho = 0$, the power of T_P and T_U is almost equal to the power of the T_{CMH}. When q is small, the power of T_P and T_U is much better than T_{EL}. As q increases, the difference in power between T_{EL} and T_P and T_U decreases. However, when $q = 32$, the power of T_P and T_U is still considerably better than the power of T_{EL}.

Table 6. Estimates of Power When $\alpha = 0.05$ and the Data from the Dirichlet-Multinomial (ρ) Distribution for $q = 8$

		$\rho = 0.0$		$\rho = 0.2$		$\rho = 0.8$	
Alternative	$n_{hik} =$	4	4-8	4	4-8	4	4-8
	T_{CMH}	.993	1.00				
	T_{EL}	.918	.993	.788	.838	.488	.493
Linear	T_P	.993	1.00	.941	.977	.644	.653
Trend	T_U	.993	1.00	.944	.977	.649	.660
	T_{CMH}	.986	1.00				
Means	T_{EL}	.594	.772	.426	.496	.246	.238
Differ	T_P	.983	1.00	.899	.952	.543	.562
	T_U	.986	1.00	.902	.953	.553	.568
	T_{CMH}	.979	1.00				
General	T_{EL}	.165	.201	.123	.135	.095	.069
Association	T_P	.978	1.00	.853	.933	.484	.502
	T_U	.979	1.00	.866	.943	.505	.530

T_{EL}: $(q - df)/(df(q-1))T_{EL}$ compare to a $F(df, q - df)$ distribution.

Table 7. Estimates of Power When $\alpha = 0.05$ and the Data from the Dirichlet-Multinomial (ρ) Distribution for $q = 16$

		$\rho = 0.0$		$\rho = 0.2$		$\rho = 0.8$	
Alternative	$n_{hik} =$	4	4-8	4	4-8	4	4-8
	T_{CMH}	.943	1.00				
Linear	T_{EL}	.900	.998	.730	.897	.437	.473
Trend	T_P	.932	1.00	.801	.931	.482	.550
	T_U	.936	1.00	.806	.933	.488	.558
	T_{CMH}	.901	1.00				
Means	T_{EL}	.807	.985	.611	.791	.423	.329
Differ	T_P	.896	1.00	.727	.883	.399	.444
	T_U	.899	1.00	.737	.886	.409	.458
	T_{CMH}	.852	.998				
General	T_{EL}	.629	.919	.395	.619	.204	.218
Association	T_P	.841	.998	.632	.833	.329	.355
	T_U	.846	.998	.663	.846	.337	.381

T_{EL}: $(q - df)/(df(q-1))T_{EL}$ compare to a $F(df, q - df)$ distribution.

Table 8. Estimates of Power When $\alpha = 0.05$ and the Data from the Dirichlet-Multinomial (ρ) Distribution for $q = 32$

		$\rho = 0.0$		$\rho = 0.2$		$\rho = 0.8$	
Alternative	$n_{hik} =$	4	4-8	4	4-8	4	4-8
	T_{CMH}	.996	1.00				
Linear	T_{EL}	.992	1.00	.921	.952	.622	.616
Trend	T_P	.995	1.00	.945	.969	.660	.659
	T_U	.996	1.00	.949	.970	.670	.664
	T_{CMH}	.986	1.00				
Means	T_{EL}	.979	1.00	.844	.904	.526	.509
Differ	T_P	.983	1.00	.890	.945	.568	.545
	T_U	.986	1.00	.893	.950	.582	.557
	T_{CMH}	.987	1.00				
General	T_{EL}	.963	1.00	.791	.867	.450	.446
Association	T_P	.983	1.00	.879	.933	.526	.507
	T_U	.982	1.00	.891	.943	.552	.528

T_{EL}: $(q - df)/(df(q-1))T_{EL}$ compare to a $F(df, q - df)$ distribution.

3 POWER CALCULATIONS

Since categorical response variables are usually based on ordinal scales, we will focus on power calculations for ordinal data. Furthermore, we will discuss the power calculation in the case of only one intervention group and one control group ($R = 2$).

Following the ideas of Wittes and Wallenstein (1987), direct calculation give us the following power approximation:

$$\Phi\left(\frac{2\sum_{h=1}^{q}(n_{h1.}n_{h2.}/N_h)\boldsymbol{C}_h\boldsymbol{\Delta}_h}{\sqrt{V}} - Z(1-\alpha/2)\right),$$

where Φ is the standard normal distribution function, $\boldsymbol{\Delta}_h = \pi_{h2} - \pi_{h1}$ is the difference of success probabilities between the treatment group and the control group for the hth stratum, $\boldsymbol{C}_h = (c_{h1}, c_{h2}, ..., c_{hC})'$ is the column score for the hth stratum, $Z(1-\alpha/2)$ is the $1-\alpha/2$ quantile of a standard normal distribution, and $V = \sum_{h=1}^{q}\{B_h\{\sum_{i=1}^{2} A_{hi}[\sum_{k=1}^{n_{hi}} Var(\boldsymbol{x}_{hi*k})]A_{hi}'\}B_h'\}$ is the variance of $\boldsymbol{G}$ with $B_h = (\boldsymbol{C}_h, -\boldsymbol{C}_h)$.

For simplicity, consider a study with equal numbers of repeated measurements for each subject ($n_{hik} = n_0$), the same alternative $\boldsymbol{\Delta}_h = \boldsymbol{\Delta}$, constant treatment and control probabilities $\pi_{h2} = \pi_2$ and $\pi_{h1} = \pi_1$ across the strata, the same covariance matrix $\boldsymbol{\Sigma}_2$ for $\boldsymbol{x}_{h2*k}$ in the treatment group and $\boldsymbol{\Sigma}_1$ for $\boldsymbol{x}_{h1*k}$ in the control group across the strata, constant response scores $\boldsymbol{C}_h = \boldsymbol{C}$ across the strata, and $\alpha = 0.05$. The power approximation simplifies to

$$\Phi\left(\frac{n_0\sqrt{n}\boldsymbol{C}\boldsymbol{\Delta}_h}{\sqrt{\boldsymbol{C}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\boldsymbol{C}'}} - 1.96\right). \tag{2}$$

When $n_0 = 1$, we can use the multinomial distribution to compute $\boldsymbol{\Sigma}_i$, where $\boldsymbol{\Sigma}_i = \boldsymbol{D}_{\pi_i} - \pi_i\pi_i'$, and $\boldsymbol{D}_{\pi_i}$ is a diagonal matrix with π_i on its main diagonal for $i = 1$ or 2. When $n_0 > 1$, we make the assumption that the $\boldsymbol{x}_{hi*k}$ has the Dirichlet-Multinomial distribution (ρ), so that the covariance matrix of $\boldsymbol{x}_{hi*k}$ is $\boldsymbol{\Sigma}_i = [1 + (n_0 - 1)\rho]n_0(\boldsymbol{D}_{\pi_i} - \pi_i\pi_i')$. Of course, different covariance structures for $\boldsymbol{x}_{hi*k}$ could be used to calculate $\boldsymbol{\Sigma}_i$.

Using the above formula (2), Figure 1 shows the power for $\pi_1 = (0.3, 0.44.\ 0.26)'$, $\boldsymbol{\Delta} = (-0.1, 0.06, 0.04)'$, $\boldsymbol{C} = (1, 2, 3)$, $\rho = 0.3, 0.5$ and 0.8, $n = 150$, and n_0 (the horizontal axis) varies from 1 to 50. Note that for n_0 is in the range 1 to 10, increasing n_0 will considerably increase the power of the test. However, when $n_0 > 10$, the gain in the power by increasing n_0 is minor. Therefore, when planning a study, one should jointly consider both the total sample size and the number of repeated measurements in order to achieve a balance between the optimal statistical power and realistic constraints.

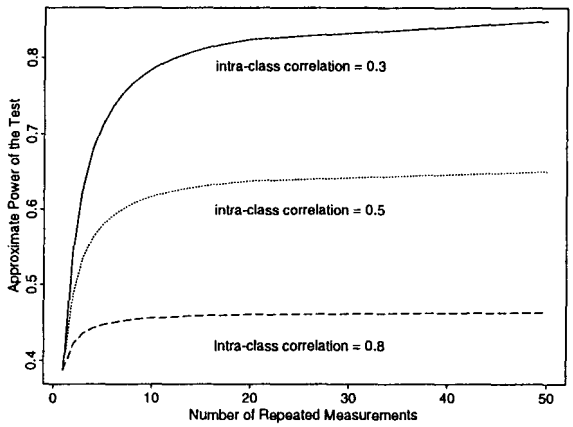


Figure 1: Power of the Test vs Number of Repeated Measurements n_0

4 EXAMPLE

The (slightly modified) data reported in Table A4 in the Appendix are from a multi-center clinical trial designed to compare a new drug for topical treatment of psoriasis with a placebo. Patients were randomly assigned to a drug (two different dose levels: low and high) or a placebo in each of the 16 clinical centers, and then were evaluated on a three-point ordinal scale from 1 to 3 (1 = "no improvement," 2 = "some improvement," and 3 = "marked improvement") at four follow-up visits. Notice that the number of patients in many centers is quite small (*sparse*), and the *sparse* situation will be more serious when adjustment for other prognostic factors, such as age, gender, and pretreatment severity, is required. Furthermore, the responses are correlated categorical data.

For illustrative purpose, we analyzed the data using all the three new statistics. Keep in mind that T_{EL} can be computed from Table A4, but in order to calculate T_P and T_U, data based on each individual subject are needed (see, for example, Table 2). The results are summarized in Table 9. All the p-values for T_P and T_U are smaller than those for T_{EL}, and this agrees with the simulations in Tables 6-8.

Table 9. Summary of the Analysis Results

Statistics	Alternative Hypothesis	df	Value	P-value
T_{EL}	Linear Trend	1	27.370	0.0001
	Mean Response Differ	2	27.939	0.0006
	General Association	4	32.397	0.0051
T_P	Linear Trend	1	25.522	4E-7
	Mean Response Differ	2	26.233	2E-6
	General Association	4	26.408	3E-5
T_U	Linear Trend	1	28.068	1E-7
	Mean Response Differ	2	28.820	6E-7
	General Association	4	29.159	7E-6

T_{EL}: $(q-df)/(df(q-1))T_{EL} \sim F(df, q-df)$, $q = 16$, $df = 1$, 2, and 4.
df: the rank of $\boldsymbol{B}_h$.

5 DISCUSSION

In this paper, we have extended the standard generalized Cochran-Mantel-Haenszel statistics to correlated categorical data and proposed three new test statistics. These new tests preserve the distinctive features of the standard generalized Cochran-Mantel-Haenszel statistics, that is, 1) they provide simultaneous adjustment for all covariates through stratification, 2) they are robust to the presence of sparse data, and 3) they have no analytical problems for missing data. Therefore, they have very broad application similar to the standard generalized Cochran-Mantel-Haenszel statistics.

Among the three statistics, we prefer T_P which uses pooled estimators in the variance estimator. The statistic T_U which uses unpooled estimators in the variance estimate performed very similar to T_P in the Monte Carlo studies. Both T_P and T_U have power advantages over T_{EL}, especially for a small number of strata. However, if we model the strata as random, only T_{EL} is valid. In addition, all three statistics have closed forms and are easy to compute. Finally, the approximate power calculations in Section 3 are easy to use when designing a study.

APPENDIX

Table A1. Parameters Used in the Simulation for $q = 8$

Stratum	n_{h1}	n_{h2}	n_{h3}	π_{hi1}	π_{hi2}	π_{hi3}
1	8	16	12	.3500	.6000	.0500
2	5	5	4	.2414	.4483	.3103
3	26	32	29	.4615	.5285	.0100
4	28	28	27	.6552	.1724	.1724
5	10	10	10	.1463	.5610	.2927
6	9	9	8	.2821	.6666	.0513
7	22	22	20	.1471	.6764	.1765
8	15	14	15	.1786	.8114	.0100

Table A2. Parameters Used in the Simulation for $q = 16$

Stratum	n_{h1}	n_{h2}	n_{h3}	π_{hi1}	π_{hi2}	π_{hi3}
1	3	10	7	.4138	.5762	.0100
2	5	6	5	.3500	.6000	.0500
3	5	5	4	.2414	.4483	.3103
4	9	10	10	.4615	.5285	.0100
5	9	10	8	.4444	.5456	.0100
6	4	4	3	.5789	.4111	.0100
7	4	8	8	.3158	.6742	.0100
8	8	7	6	.1250	.7187	.1563
9	7	6	7	.6552	.1724	.1724
10	13	15	14	.2619	.5952	.1429
11	10	10	10	.1463	.5610	.2927
12	9	9	8	.2821	.6666	.0513
13	7	7	6	.1471	.6764	.1765
14	15	15	14	.2500	.6364	.1136
15	7	6	7	.1786	.8114	.0100
16	8	8	8	.4688	.5212	.0100

Table A3. Parameters Used in the Simulation for $q = 32$

Stratum	n_{h1}	n_{h2}	n_{h3}	π_{hi1}	π_{hi2}	π_{hi3}
1	3	3	4	.4138	.5762	.0100
2	4	3	3	.3638	.4762	.1600
3	5	6	5	.3500	.6000	.0500
4	5	5	4	.2414	.4483	.3103
5	3	3	3	.4615	.5285	.0100
6	3	3	3	.3615	.3785	.2600
7	3	3	5	.3615	.4285	.2100
8	4	4	5	.4444	.5456	.0100
9	4	4	6	.2944	.3956	.3100
10	4	4	3	.5789	.4111	.0100
11	4	4	3	.3158	.6742	.0100
12	3	3	4	.3158	.6742	.0100
13	3	3	4	.2250	.5187	.2563
14	3	3	5	.1250	.7187	.1563
15	3	3	4	.6552	.1724	.1724
16	3	3	4	.6552	.1724	.1724
17	5	5	4	.2619	.5952	.1429
18	5	5	4	.2619	.5952	.1429
19	5	5	4	.2619	.5952	.1429
20	5	5	5	.1463	.4610	.3927
21	5	5	5	.1463	.5610	.2927
22	4	4	4	.2821	.5666	.1513
23	4	4	6	.2821	.6666	.0513
24	3	3	3	.1471	.6764	.1765
25	3	3	5	.1471	.6764	.1765
26	3	3	4	.3500	.5364	.1136
27	4	4	4	.2500	.5364	.2136
28	4	5	13	.2500	.6364	.1136
29	3	3	4	.1786	.8114	.0100
30	3	3	4	.1786	.8114	.0100
31	3	3	4	.4688	.3212	.2100
32	4	4	6	.4688	.5212	.0100

Table A4. Frequency Distribution of Patient Responses from a Multicenter Clinical Trial to Compare a New Drug (Low and High Dose Levels) with a Placebo

Center	Treatment	Score 1	Score 2	Score 3	Total number of patients
		$x_{hi1.}$ (Prop)	$x_{hi2.}$ (Prop)	$x_{hi3.}$ (Prop)	n_{hi}
1	placebo	12 (0.41)	17 (0.59)	0 (0.00)	10
	low dose	4 (0.09)	31 (0.70)	9 (0.21)	13
	high dose	13 (0.35)	18 (0.49)	6 (0.16)	12
2	placebo	7 (0.35)	12 (0.60)	1 (0.05)	5
	low dose	4 (0.17)	18 (0.75)	2 (0.08)	6
	high dose	3 (0.15)	13 (0.65)	4 (0.20)	6
3	placebo	7 (0.24)	13 (0.45)	9 (0.31)	8
	low dose	0 (0.00)	22 (0.85)	4 (0.15)	8
	high dose	0 (0.00)	12 (0.60)	8 (0.40)	8
4	placebo	18 (0.46)	21 (0.54)	0 (0.00)	10
	low dose	9 (0.23)	29 (0.72)	2 (0.05)	10
	high dose	2 (0.05)	33 (0.83)	5 (0.12)	10
5	placebo	20 (0.44)	25 (0.56)	0 (0.00)	12
	low dose	7 (0.15)	36 (0.78)	3 (0.07)	12
	high dose	4 (0.10)	34 (0.81)	4 (0.09)	12
6	placebo	11 (0.58)	8 (0.42)	0 (0.00)	5
	low dose	7 (0.37)	10 (0.53)	2 (0.10)	5
	high dose	1 (0.08)	9 (0.69)	3 (0.23)	4
7	placebo	6 (0.32)	13 (0.68)	0 (0.00)	8
	low dose	16 (0.50)	12 (0.38)	4 (0.12)	8
	high dose	16 (0.50)	15 (0.47)	1 (0.03)	8
8	placebo	4 (0.13)	23 (0.72)	5 (0.15)	8
	low dose	3 (0.10)	16 (0.51)	12 (0.39)	8
	high dose	0 (0.00)	14 (0.48)	15 (0.52)	8
9	placebo	19 (0.66)	5 (0.17)	5 (0.17)	8
	low dose	5 (0.21)	12 (0.50)	7 (0.29)	6
	high dose	4 (0.13)	18 (0.58)	9 (0.29)	8
10	placebo	11 (0.26)	25 (0.60)	6 (0.14)	11
	low dose	12 (0.25)	28 (0.58)	8 (0.17)	12
	high dose	5 (0.11)	28 (0.62)	12 (0.27)	12
11	placebo	6 (0.15)	23 (0.56)	12 (0.29)	11
	low dose	9 (0.22)	25 (0.61)	7 (0.17)	11
	high dose	4 (0.10)	26 (0.62)	12 (0.28)	11
12	placebo	11 (0.28)	26 (0.67)	2 (0.05)	10
	low dose	5 (0.13)	31 (0.79)	3 (0.08)	10
	high dose	7 (0.22)	16 (0.50)	9 (0.28)	8

The number in parentheses is the proportion which belongs to that category.

Continue from Table A4

Center	Treatment	Score 1	Score 2	Score 3	Total number of patients
13	placebo	5 (0.15)	23 (0.68)	6 (0.17)	9
	low dose	0 (0.00)	28 (0.85)	5 (0.15)	9
	high dose	7 (0.23)	19 (0.63)	4 (0.14)	8
14	placebo	11 (0.25)	28 (0.64)	5 (0.11)	11
	low dose	14 (0.29)	26 (0.54)	8 (0.17)	12
	high dose	5 (0.11)	33 (0.75)	6 (0.14)	12
15	placebo	5 (0.18)	23 (0.82)	0 (0.00)	7
	low dose	3 (0.13)	14 (0.61)	6 (0.26)	6
	high dose	3 (0.10)	16 (0.52)	12 (0.39)	8
16	placebo	15 (0.47)	17 (0.53)	0 (0.00)	8
	low dose	8 (0.25)	23 (0.72)	1 (0.03)	8
	high dose	0 (0.00)	28 (0.87)	4 (0.13)	8

The number in parentheses is the proportion which belongs to that category.

Proof of Theorem 1: Direct calculations give

$$E(\boldsymbol{x}_{hi*k} - n_{hik}\hat{\boldsymbol{\pi}}_h)(\boldsymbol{x}_{hi*k} - n_{hik}\hat{\boldsymbol{\pi}}_h)' =$$
$$Var(\boldsymbol{x}_{hi*k})(1 - n_{hik}/N_h) - \frac{n_{hik}}{N_h}Var(\boldsymbol{x}_{hi*k}) + \frac{n_{hik}^2}{N_h^2}\{\sum_{i=1}^{R}\sum_{k=1}^{n_{hi}} Var(\boldsymbol{x}_{hi*k})\} .$$

When we substitute $Var(\boldsymbol{x}_{hi*k}) = n_{hik}\boldsymbol{\Sigma}_h$, the last two terms above cancel. Thus

$$E\{\sum_{k=1}^{n_{hi}} \frac{(\boldsymbol{x}_{hi*k} - n_{hik}\hat{\boldsymbol{\pi}}_h)(\boldsymbol{x}_{hi*k} - n_{hik}\hat{\boldsymbol{\pi}}_h)'}{(1 - n_{hik}/N_h)}\} = \sum_{k=1}^{n_{hi}} Var(\boldsymbol{x}_{hi*k}) .$$

Then, Theorem 1 follows.

Proof of Theorem 2: From (1) we get

$$E\{\sum_{k=1}^{n_{hi}} \frac{(\boldsymbol{x}_{hi*k} - n_{hik}\hat{\boldsymbol{\pi}}_{hi})(\boldsymbol{x}_{hi*k} - n_{hik}\hat{\boldsymbol{\pi}}_{hi})'}{(1 - 2n_{hik}/n_{hi\cdot})}\}$$
$$= \{1 + \sum_{k=1}^{n_{hi}}(\frac{n_{hik}^2/n_{hi\cdot}^2}{1 - 2n_{hik}/n_{hi\cdot}})\}\sum_{k=1}^{n_{hi}} Var(\boldsymbol{x}_{hi*k})$$
$$= \delta_{hi}Var(\boldsymbol{x}_{hi*k}) .$$

Then, Theorem 2 follows.

Proof of Theorem 3: Note that

$$\boldsymbol{G} = \sum_{h=1}^{q} \boldsymbol{B}_h(\boldsymbol{x}_h - \boldsymbol{m}_h)$$

$$= \sum_{h=1}^{q}\sum_{i=1}^{R}\sum_{k=1}^{n_{hi}} \boldsymbol{B}_h \boldsymbol{A}_{hi} \boldsymbol{x}_{hi*k}$$
$$= \sum_{h=1}^{q}\sum_{i=1}^{R}\sum_{k=1}^{n_{hi}} \boldsymbol{y}_{hik} \ .$$

Since each element of vector $\boldsymbol{y}_{hik}$ is bounded, the Lindeberg condition is satisfied. By the assumption of $Var(\boldsymbol{G})/n \rightarrow \boldsymbol{\Sigma}$ and the central limit theorem, we have

$$\boldsymbol{G}/\sqrt{n} \xrightarrow{d} \boldsymbol{Z}, \boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}) \ .$$

Furthermore, $E(\boldsymbol{V}_P)/n = Var(\boldsymbol{G})/n \rightarrow \boldsymbol{\Sigma}$. If v denotes an element of $\boldsymbol{V}_P$, it is easy to verify that $Var(v/n) \rightarrow 0$. Then we have $\boldsymbol{V}_P/n \xrightarrow{p} \boldsymbol{\Sigma}$. Therefore,

$$\begin{aligned} \boldsymbol{G}'\boldsymbol{V}_P^{-1}\boldsymbol{G} &= \frac{\boldsymbol{G}'}{\sqrt{n}}\left(\frac{\boldsymbol{V}_P}{n}\right)^{-1}\frac{\boldsymbol{G}}{\sqrt{n}} \\ &\xrightarrow{d} \boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} \\ &= \chi^2_{df} \ . \end{aligned}$$

where df = rank($\boldsymbol{\Sigma}$) = rank($\boldsymbol{B}_h$). A similar proof works for T_U.

REFERENCES

Beitler, P.J. and Landis, J.R. (1985). A mixed-effects model for categorical data. *Biometrics* 41, 991-1000.

Boos, D.D. and Brownie, C. (1992). A rank-based mixed model approach to multisite clinical trials. *Biometrics* 48, 61-72.

Brier, S.S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika* 67, 591-596.

Koch, G.G., Landis, J.R., Freeman, D.H. and Lehnen, R.G. (1977) A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* 33, 133-158.

Kuritz, S.J., Landis, J.R. and Koch, G.G. (1988). A general overview of Mantel-Haenszel methods. *Annual Review of Public Health* 9, 123-160.

Landis, J.R., Heyman, E.R. and Koch, G.G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review* 46, 237-254.

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

Liang, K.Y. and Zeger, S.L. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science* 10, 158-199.

Robins, J., Breslow, N., and Greenland, S. (1986) Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 42, 311-323.

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

Zhang, J. and Boos, D.D. (1995). Mantel-Haenszel test statistics for correlated binary data. *Institute of Statistics Mimeo* Series No. 2274.