International Journal of Statistics and Probability; Vol. 7, No. 5; September 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Testing Simultaneous Marginal Homogeneity for Clustered Matched-Pair Multinomial Data

Bo Deng[1] & Keumhee Chough Carriere[1]

[1] University of Alberta, Canada.

Correspondence: Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada.

Received: June 5, 2018 Accepted: June 27, 2018 Online Published: August 9, 2018

doi:10.5539/ijsp.v7n5p86 URL: https://doi.org/10.5539/ijsp.v7n5p86

Abstract

For matched-pair data with a polychotomous outcome, the Stuart-Maxwell test (1955) and the Bhapkar test(1966) are commonly used to test marginal homogeneity. When the outcome is ordinal, the test proposed by Agresti (1983) can be used to test the marginal homogeneity against stochastic order. In practice, we often face the need to consider multiple categorical outcomes simultaneously to insure Type I error protection. In this paper, we propose three statistics to test simultaneous marginal homogeneity for multiple multinomial outcomes in two dependent samples. Furthermore, when the outcome is ordinal, we also propose a transformed version of the three statistics for testing simultaneous marginal homogeneity against stochastic order in two dependent samples. We then prove their asymptotic properties. Finally, Monte Carlo simulations are conducted to evaluate their performance in small samples with respect to empirical size and power.

Keywords: Simultaneous marginal homogeneity, Stochastic order, Chi-square distribution

1. Introduction

In phase II-III clinical trial of pharmaceutical products, the analysis of adverse event(AE) data is an important aspect of examining the safety of a new drug. The investigator are often interested in looking at if the severity distributions of the AEs patients experienced are different under two treatments or two dosages. The simplest way is to analyze each AE separately and combine individual $p-$values through various multiple adjustment techniques. However, many AEs are correlated to a huge extent. Without a proper adjustment, type I error will not be adequately controlled. Therefore, analyzing the incidence of related AEs simultaneously is ideal and preferable to treating them separately. Analyzing the severity distribution of multiple AEs simultaneously is challenging, as it leads to dealing with clustered and thus correlated matched-pair multinomial data.

Chuang-Stein and Mohberg (1993) developed the Wald and score-type tests for comparing the incidence of multiple AEs simultaneously in two independent groups. The score-type test used the pooled covariance estimator under two treatments. Agresti and Klingenberg (2005) recommended the score-type test over the Wald test because the empirical size of the score-type test tends to be closer to the nominal level than the Wald test does.

Klingenberg and Agresti (2006) developed the Wald and score-type test for comparing the incidence of multiple AEs simultaneously in two dependent groups. Both statistics asymptotically follow a Chi-square distribution. The score-type statistic is preferred because it maintains the nominal level much better than the Wald statistic. They also pointed out that for sample sizes less than 100, neither statistic is well approximated by a χ^2 distribution when the number of AE considered is 2 and 4. Therefore, the score-type statistic using bootstrap method is recommended in small samples. In situations when bootstrap method is too computationally intensive, a permutation test is recommended. But the permutation test is to test the identical joint distribution (IJD), which is a stronger hypothesis than SMH. Thus, the permutation test will lead to inflated type I error rates when testing SMH.

Klingenberg et al. (2008) developed the score-type test and a score-free statistic for testing SMH against stochastic order in two independent groups. The permutation test and bootstrap method based on these two tests were proposed. For small sample sizes, the bootstrap method for the two tests is either too conservative or liberal, while the permutation test has the empirical size closer to the nominal level. However, the permutation is only valid for testing SMH when a prior condition (stochastic order) is assumed or two samples are balanced.

In this paper, we focus on the methods for testing SMH in matched-pair multinomial data. In section 2, we define the hypothesis of SMH for matched-pair multinomial data and develop the statistics to test SMH. In section 3, we focus on a special case of multinomial data, which is ordered. For ordinal data, an alternative hypothesis, stochastic order, is defined. Then we present the statistics to test SMH against stochastic order. In section 4, Monte Carlo simulations are performed to

evaluate the performance of the statistics introduced in section 2 and 3. Section 5 discusses the drawbacks of the methods proposed and possible extensions. The paper is entirely framed in terms of a safety analysis comparing the marginal proportions of each AE severity categories under two treatments or dosages. However, our methods can be applied to any paired or repeated multinomial response.

2. Tests of SMH for Clustered Matched-pair Multinomial Data

2.1 Simultaneous Marginal Homogeneity

Let $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \cdots, Y_{ijK})^T$ be a $K \times 1$ vector of multivariate responses for subject i at dose $j = 1, 2$, where K is the number of multinomial responses (AE severity), Y_{ijk} is the multinomial response with $C_k > 2$ categories, $k = 1, 2, \cdots, K$. In this paper, $C_k = C = 4$ is used for all k which denotes the 4 severity levels of AE, i.e. none, mild, moderate and severe. If subject i experienced AE k with severity c_k at dose j, then $Y_{ijk} = c_k$. For each subject, let $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \mathbf{Y}_{i2})^T = (Y_{i11}, Y_{i12}, \cdots, Y_{i1K}, Y_{i21}, Y_{i22}, \cdots, Y_{i2K})^T$ denote the subject i's AE severity profile. Assume that we have n subjects in the study, where $(\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n)$ are n independently and identically distributed random variables from a multinomial distribution with probability $\pi(c_1, c_2, \cdots, c_K, c_1^{'}, c_2^{'}, \cdots c_K^{'})$, where $\pi(c_1, c_2, \cdots, c_K, c_1^{'}, c_2^{'}, \cdots, c_K^{'})$ denotes the joint probability $Pr(Y_{i11} = c_1, \cdots, Y_{i1K} = c_K, \cdots, Y_{i21} = c_1^{'}, \cdots, Y_{i2K} = c_K^{'})$.

Let $\pi_j = (\pi_{j1}(1), \cdots, \pi_{j1}(C), \pi_{j2}(1), \cdots, \pi_{j2}(C), \cdots, \pi_{jK}(1), \cdots, \pi_{jK}(C))^{'}$, where $\pi_{jk}(c_k)$ denotes the probability $Pr(Y_{ijk} = c_k), c_k = 1, 2, \cdots, C$, for $C = 4$. The null hypothesis of SMH is defined as

$$H_0 : \pi_{1k}(c_k) = \pi_{2k}(c_k) \quad \text{for} \quad k = 1, 2, \cdots, K, \qquad c_k = 1, 2, \cdots, C.$$

2.2 Multivariate Tests of SMH

Motivated by the statistic proposed by Agresti and Klingenberg (2005) and Klingenberg and Agresti (2006), a statistic to test SMH is constructed by comparing the marginal proportions of each AE at two dosages. Let $\hat{\pi}_j = (\hat{\pi}_{j1}(1), \cdots, \hat{\pi}_{j1}(C-1), \hat{\pi}_{j2}(1), \cdots, \hat{\pi}_{j2}(C-1), \cdots, \hat{\pi}_{jK}(1), \cdots, \hat{\pi}_{jK}(C-1))^T$ denote the marginal proportions of each AE at dose j, where $j = 1, 2$ and $\hat{\pi}_{jk}(c_k)$ denotes the sample proportion of subjects with severity c_k of AE k at dose j, $c_k = 1, 2 \cdots, C-1$. Let $d = \hat{\pi}_2 - \hat{\pi}_1 = (\hat{d}_1(1), \cdots, \hat{d}_1(C-1), \hat{d}_2(1), \cdots, \hat{d}_2(C-1), \cdots, \hat{d}_K(1), \cdots, \hat{d}_K(C-1))^T$ denote the difference of the marginal sample proportions at two dosages.

Under the assumption of multinomial distribution, the covariance matrix V of d has elements:

$$\begin{aligned} Var(\hat{d}_k(c_k)) &= Var(\hat{\pi}_{2k}(c_k) - \hat{\pi}_{1k}(c_k)) \\ &= Var(\hat{\pi}_{2k}(c_k)) + Var(\hat{\pi}_{1k}(c_k)) - 2Cov(\hat{\pi}_{2k}(c_k), \hat{\pi}_{1k}(c_k)) \\ &= \frac{(\pi_{1k}(c_k) + \pi_{2k}(c_k) - 2\pi_k(c_k, c_k)) - (\pi_{1k}(c_k) - \pi_{2k}(c_k))^2}{n}, \end{aligned} \tag{1}$$

where $\pi_k(c_k, c_k)$ is the probability of experiencing AE k of severity c_k at both dosages,

$$\begin{aligned} Cov(\hat{d}_k(c_k), \hat{d}_k(c_k^{'})) &= Cov(\hat{\pi}_{2k}(c_k) - \hat{\pi}_{1k}(c_k), \hat{\pi}_{2k}(c_k^{'}) - \hat{\pi}_{1k}(c_k^{'})) \\ &= -\frac{\pi_{2k}(c_k)\pi_{2k}(c_k^{'})}{n} - \frac{\pi_{1k}(c_k)\pi_{1k}(c_k^{'})}{n} \\ &\quad -\frac{\pi_{12k}(c_k^{'}, c_k) - \pi_{1k}(c_k^{'})\pi_{2k}(c_k)}{n} \\ &\quad -\frac{\pi_{12k}(c_k, c_k^{'}) - \pi_{1k}(c_k)\pi_{2h}(c_k^{'})}{n}, \end{aligned} \tag{2}$$

where $\pi_{12k}(c_k^{'}, c_k)$ is the joint probability of experiencing AE k of severity $c_k^{'}$ at dose 1 and AE k of severity c_k at dose 2, and

$$\begin{aligned} Cov(\hat{d}_k(c_k), \hat{d}_{k'}(c_{k'})) &= Cov(\hat{\pi}_{2k}(c_k) - \hat{\pi}_{1k}(c_k), \hat{\pi}_{2k'}(c_{k'}) - \hat{\pi}_{1k'}(c_{k'})) \\ &= \frac{\pi_{2kk'}(c_k, c_{k'}) - \pi_{2k}(c_k)\pi_{2k'}(c_{k'})}{n} \\ &\quad - \frac{\pi_{12k'k}(c_{k'}, c_k) - \pi_{1k'}(c_{k'})\pi_{2k}(c_k)}{n} \\ &\quad - \frac{\pi_{12k'k}(c_k, c_{k'}) - \pi_{1k}(c_k)\pi_{2k'}(c_{k'})}{n} \\ &\quad + \frac{\pi_{1kk'}(c_k, c_{k'}) + \pi_{1k}(c_k)\pi_{1k'}(c_{k'})}{n}, \end{aligned} \tag{3}$$

where $\pi_{jkk'}(c_k, c_{k'})$ is the joint probability of experiencing AE k of severity c_k and AE $k^{'}$ of severity $c_{k'}$ at dose j and $\pi_{jj'kk'}(c_k, c_{k'})$ is the joint probability of experiencing AE k of severity c_k at dose j and AE $k^{'}$ of severity $c_{k'}$ at dose $j^{'}$.

Let $\hat{V}$ denote the sample version of V. Then, a Wald statistic to test SMH is

$$W = d^T \hat{V}^{-1} d,$$

and based on the Central Limit Theorem, W has an asymptotic null χ^2 distribution with $df = K(C-1)$ when $n \longrightarrow \infty$. By replacing $\pi_{1k}(c_k)$ and $\pi_{2k}(c_k)$ by the pooled estimator, $\hat{\pi}_{0k}(c_k) = (\hat{\pi}_{1k}(c_k) + \hat{\pi}_{2k}(c_k))/2$, we have $\hat{V}_s$ as the pooled estimator of V, which has elements as:

$$\begin{aligned} \widehat{Var}(\hat{d}_k(c_k)) &= \frac{2(\hat{\pi}_{0k}(c_k) - \hat{\pi}_k(c_k, c_k))}{n} \\ \widehat{Cov}(\hat{d}_k(c_k), \hat{d}_k(c_k^{'})) &= \frac{-\hat{\pi}_{12k'k}(c_k^{'}, c_k) - \hat{\pi}_{12kk'}(c_k, c_k^{'})}{n} \\ \widehat{Cov}(\hat{d}_k(c_k), \hat{d}_{k'}(c_{k'})) &= \frac{\hat{\pi}_{1kk'}(c_k, c_{k'}) + \hat{\pi}_{2kk'}(c_k, c_{k'}) - \hat{\pi}_{12k'k}(c_{k'}, c_k) - \hat{\pi}_{12kk'}(c_k, c_{k'})}{n}. \end{aligned} \tag{4}$$

Then we have a score-type test statistic $W_s = d^T \hat{V}_s^{-1} d$, which also has an asymptotic null χ^2 distribution with $df = K(C-1)$ when $n \longrightarrow \infty$. In the binary case ($C = 2$), the W and W_s reduces to the multivariate McNemar's tests (Klingenberg and Agresti, 2006).

For the Wald and score-type statistics, it is assumed that $(\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n)$ are n independently and identically distributed random variables from a multinomial distribution. However, this assumption may not be feasible in practice. Therefore, a non-parametric covariance estimator of d is considered in this section. The non-parametric covariance estimator $\hat{V}_{np}$ of d is given as follows:

$$\begin{aligned} \widehat{Var}(\hat{d}_k(c_k)) &= \widehat{Var}(\hat{\pi}_{2k}(c_k) - \hat{\pi}_{1k}(c_k)) \\ &= \frac{\sum_{i=1}^n (Y_{i2k}(c_k) - \overline{Y}_{2k}(c_k))^2}{n(n-1)} + \frac{\sum_{i=1}^n (Y_{i1k}(c_k) - \overline{Y}_{1k}(c_k))^2}{n(n-1)} \\ &\quad - \frac{\sum_{i=1}^n 2((Y_{i1k}(c_k) - \overline{Y}_{1k}(c_k))(Y_{i2k}(c_k) - \overline{Y}_{2k}(c_k))}{n(n-1)}, \end{aligned} \tag{5}$$

where $Y_{i1k}(c_k) = 1$ if subject i experience AE k of severity c_k; $Y_{i1k}(c_k) = 0$ if subject i did not experience AE k of severity c_k.

$$\begin{aligned}\widehat{Cov}(\hat{d}_k(c_k), \hat{d}_{k'}(c_{k'})) &= \widehat{Cov}(\hat{\pi}_{2k}(c_k) - \hat{\pi}_{1k}(c_k), \hat{\pi}_{2k'}(c_{k'}) - \hat{\pi}_{1k'}(c_{k'})) \\ &= \frac{\sum_{i=1}^{n}((Y_{i2k}(c_k) - \overline{Y}_{2k}(c_k))(Y_{i2k'}(c_{k'}) - \overline{Y}_{2k'}(c_{k'})))}{n(n-1)} \\ &\quad - \frac{\sum_{i=1}^{n}((Y_{i2k}(c_k) - \overline{Y}_{2k}(c_k))(Y_{i1k'}(c_{k'}) - \overline{Y}_{1k'}(c_{k'})))}{n(n-1)} \\ &\quad - \frac{\sum_{i=1}^{n}((Y_{i1k}(c_k) - \overline{Y}_{1k}(c_k))(Y_{i2k'}(c_{k'}) - \overline{Y}_{2k'}(c_{k'})))}{n(n-1)} \\ &\quad + \frac{\sum_{i=1}^{n}((Y_{i1k}(c_k) - \overline{Y}_{1k}(c_k))(Y_{1ik'}(c_{k'}) - \overline{Y}_{1k'}(c_{k'})))}{n(n-1)}.\end{aligned} \tag{6}$$

Next, we will show (5) and (6) are consistent estimators of $Var(\hat{d}_k(c_k))$ and $Cov(\hat{d}_k(c_k), \hat{d}_{k'}(c_{k'}))$, respectively. First, we will show $\sum_{i=1}^{n}(Y_{i2k}(c_k) - \overline{Y}_{2k}(c_k))^2/n(n-1)$ in (5) is a consistent estimator of $Var(\hat{\pi}_{2k}(c_k))$.

Assume that $Y_{12k}, Y_{22k}, \cdots, Y_{n2k}$ are independently and identically distributed with mean μ and variance σ^2. It is well-known that sample variance $\sum_{i=1}^{n}(Y_{i2k}(c_k) - \overline{Y}_{2k}(c_k))^2/(n-1)$ is a consistent estimator of σ^2. Furthermore, we have

$$Var(\hat{\pi}_{2k}(c_k)) = Var(\frac{\sum_{i=1}^{n}(Y_{i2k}(c_k))}{n}) = \frac{Var(Y_{i2k})}{n} = \frac{\sigma^2}{n}.$$

Hence, $\sum_{i=1}^{n}(Y_{i2k}(c_k) - \overline{Y}_{2k}(c_k))^2/n(n-1)$ is a consistent estimator of $Var(\hat{\pi}_{2k}(c_k))$. Similarly, it can be shown that $\sum_{i=1}^{n}(Y_{i1k}(c_k) - \overline{Y}_{1k}(c_k))^2/n(n-1)$ and $\sum_{i=1}^{n}((Y_{i1k}(c_k) - \overline{Y}_{1k}(c_k))(Y_{i2k}(c_k) - \overline{Y}_{2k}(c_k))/n(n-1)$ are consistent estimators of $Var(\hat{\pi}_{1k}(c_k))$ and $Cov(\hat{\pi}_{2k}(c_k), \hat{\pi}_{1k}(c_k))$, respectively. Therefore, (5) is a consistent estimator of $Var(\hat{d}_k(c_k))$.

Similarly, it can be shown that (6) is a consistent estimator of $Cov(\hat{d}_k(c_k), \hat{d}_{k'}(c_{k'}))$. Since (5) and (6) are both consistent, $\hat{V}_{np}$ is a consistent estimator of variance of d.

From the Central Limit Theorem, we have a non-parametric Wald test statistic $W_{np} = d^T \hat{V}_{np} d$, which asymptotically follows a χ^2 distribution with $df = K(C-1)$ when $n \longrightarrow \infty$.

3. Tests of SMH against Stochastic Order in Clustered Matched-pair Ordinal Data

3.1 Stochastic Order

All the statistics introduced in section 2.2 treat the outcome as nominal. They can be used to test marginal homogeneity against any alternatives. When the outcome is ordinal, however, they ignore the ordinal nature of the outcome. Quite naturally, the tests that utilize the ordinal information will be more powerful. For instance, for ordinal outcome, one is usually interested if the classifications based on one variable are higher than those based on the other variable. For a $I \times I$ square table, let Y_1 denote the observation from the row marginal distribution $\{\pi_{i+}\}$ and Y_2 denote the observation from the column marginal distribution $\{\pi_{+j}\}$. Y_1 is stochastically higher than Y_2 (Agresti, 2010) if the cumulative density function of Y_1 is uniformly below the cumulative density function of Y_2, i.e.

$$\pi_{1+} + \cdots + \pi_{j+} \leq \pi_{+1} + \cdots + \pi_{+j}, \qquad \text{for } j = 1, \cdots, I-1.$$

This means that Y_1 is more likely to have larger values than Y_2.

3.2 Multivariate Tests of SMH against Stochastic Order

For clustered matched-pair AE severity data, it may be of interest to test whether one margin is stochastically higher than the other for each AE. Motivated by the statistic proposed by Agresti (1983), a statistic for testing SMH against stochastic ordering is formed by comparing the marginal mean scores under two treatments.

Let $\hat{\pi}_j = (\hat{\pi}_{j1}(1), \cdots, \hat{\pi}_{j1}(C), \hat{\pi}_{j2}(1), \cdots, \hat{\pi}_{j2}(C), \cdots, \hat{\pi}_{jk}(1), \cdots, \hat{\pi}_{jK}(C))^T$, where $\hat{\pi}_{jk}(c_k)$ denotes the sample proportion of subjects with severity c_k of AE k at dose j, $c_k = 1, 2 \cdots, C$. Let

$$d = \hat{\pi}_2 - \hat{\pi}_1 = (\hat{d}_1(1), \cdots, \hat{d}_1(C), \hat{d}_2(1), \cdots, \hat{d}_2(C), \cdots, \hat{d}_K(1), \cdots, \hat{d}_K(C))^T$$

denote the difference of the marginal sample proportions at two dosages. The difference of the marginal mean scores at two dosages is formed by $S = Ad$, where $A = diag({u_k}^T, k = 1, 2, \cdots, K)$ is a matrix with score ${u_k}^T = (u_k(1), u_k(2), \cdots, u_k(C))^T$ for severity levels of AE k.

Under the assumption of multinomial distribution, the covariance matrix V of d is given in (1), (2) and (3). From $S = Ad$, we have the covariance matrix of S as $\Sigma = AVA^T$. Let $\hat{\Sigma}$ denote the sample version of Σ. A multivariate Wald test is constructed by

$$W_o = S^T \hat{\Sigma}^{-1} S$$

Based on the Central Limit Theorem, W_o asymptotically follows a Chi-square distribution with $df = K$ when $n \longrightarrow \infty$, where K is the number of AEs considered simultaneously. When $K = 1$, it reduces to the statistic proposed by Agresti (1983).

By replacing $\pi_{1k}(c_k)$ and $\pi_{2k}(c_k)$ with the pooled estimator $\hat{\pi}_{0k}(c_k) = (\hat{\pi}_{1k}(c_k) + \hat{\pi}_{2k}(c_k))/2$, we have $\hat{V}_s$ as the pooled estimator of V, which is given in (4). Then we have a score-type statistic

$$W_{os} = S^T \hat{\Sigma}_s^{-1} S,$$

which also asmptotically follows the χ^2 distribution with $df = K$ when $n \longrightarrow \infty$, where $\hat{\Sigma}_s = A\hat{V}_s A^T$.

The above two statistics assume that $(Y_1, Y_2, \cdots, Y_n)$ are n independently and identically distributed random variables from a multinomial distribution. Similar to the non-parametric statistic in section 2.2, a non-parametric covariance estimator of d can be considered, which is as given in (5) and (6). Then we have a non-parametric Wald test statistic $W_{onp} = S^T \hat{\Sigma}_{np}^{-1} S$ which also asymptotically follows a χ^2 distribution with $df = K$ when $n \longrightarrow \infty$, where $\hat{\Sigma}_{np} = A\hat{V}_{np}A^T$.

4. Asymptotics and Power

4.1 Asymptotics of W, W_s and W_{np}

In this section, the empirical size of W, W_s and W_{np} are compared for sample sizes n = 25; 50; 100; 200; 300; 500 and for the number of AEs K = 2; 3; 4, under the nominal level = 0.05. And we use $C = 4$ categories for each AE, To simulate a dataset under the null hypothesis (SMH), a $4^{2K} \times 1$ random vector of (a length of 4^{2K} is used because two treatments are considered) multinomial probabilities is generated. Then the iterative proportional fitting procedure (Deming and Stephan, 1940) is performed to adjust the vector to make each one of K pairs of marginal probabilities under two treatments equivalent to a specified 4×1 vector of probabilites.

In practice, higher levels of severity of an AE may be observed less frequenly than lower levels and our interest is in the performance of the tests on low incidence events. Therefore, high probabilities are assigned to the lower levels of severity, while low probabilities are assigned to the higher levels in our simulation design. More specifically, when simulating the data sets under SMH for $K = 2$, $(0.5, 0.25, 0.24, 0.01)$ is used as the marginal probabilities for AE = 1 and $(0.7, 0.2, 0.05, 0.05)$ is used as the marginal probabilities for AE = 2 under both treatments. For $K = 3$, $(1/3, 1/3, 0.25, 1/12)$ is used as the marginal probabilities for AE = 1, $(0.45, 1/3, 1/6, 0.05)$ is used as the marginal probabilities for the AE = 2, and $(0.45, 0.25, 0.25, 0.05)$ is used as the marginal probabilities for the AE = 3 under both treatments. For $K = 4$, they are $(0.5, 0.25, 0.24, 0.01)$, $(0.7, 0.2, 0.05, 0.05)$, $(0.4, 0.25, 0.25, 0.1)$ and $(0.7, 0.15, 0.11, 0.04)$ as the marginal probabilities for the AE = 1, 2, 3, 4, respectively, under both treatments.

Table 1. Empirical size of the W, W_s and W_{np} in 5000 simulated data sets under the nominal level 0.05

		Empirical Size					
K	Method	n=25	n=50	n=100	n=200	n=300	n=500
2	W	**0.168**	**0.112**	**0.077**	**0.058**	0.055	**0.057**
	W_s	**0.015**	**0.035**	0.047	**0.042**	0.045	0.056
	W_{np}	**0.147**	**0.106**	**0.074**	0.056	0.054	0.051
3	W	**0.348**	**0.168**	**0.088**	**0.067**	**0.061**	**0.058**
	W_s	**0.009**	**0.03**	**0.038**	**0.042**	**0.042**	0.049
	W_{np}	**0.325**	**0.157**	**0.084**	**0.066**	**0.060**	**0.058**
4	W	**0.553**	**0.24**	**0.12**	**0.077**	**0.079**	**0.06**
	W_s	**0.003**	**0.02**	**0.034**	**0.037**	0.05	0.046
	W_{np}	**0.514**	**0.22**	**0.114**	**0.074**	**0.077**	**0.059**

Note: The bold text indicates that the empirical size falls outside the 95% confidence interval $(0.044, 0.056)$ of the nominal level 0.05 .

When the simulated data set was very sparse and the test failed to work, the data set was eliminated from the summary analysis.

Table 1 shows the empirical size of W, W_s and W_{np} at a nominal type I error rate of 0.05. The W and W_{np} are too liberal and the empirical size improves as the sample size increases. On the contrary, the W_s is too conservative and the empirical size improves as the sample size increases. When $K = 2$, the W_s maintain the nominal level well when $n \geq 100$ and a larger sample size is required for W and W_{np} to control the empirical size at the nominal level. When $K = 3, 4$, none of the W, W_s and W_{np} seem to perform reasonably if $n \leq 200$. A minimum sample size of 300 is required for W_s when number of outcomes is greater than 2.

4.2 Asymptotics of W_o, W_{os} and W_{onp}

Table 2 shows the empirical size of W_o, W_{os} and W_{onp} with scores $(1, 2, 3, 4)$. It shows that W_{os} maintains the nominal level much better than W_o and W_{onp} and is therefore recommended to use. When $K = 2$, W_{os} can be used even at $n = 25$. When $K = 3, 4$, W_{os} can be used when sample size is at least 50.

Table 2. Empirical size of the W_o, W_{os} and W_{onp} with scores $(1, 2, 3, 4)$ in 5000 simulated data sets under the nominal level 0.05

		Empirical Size					
K	Method	n=25	n=50	n=100	n=200	n=300	n=500
2	W_o	**0.093**	**0.066**	**0.066**	0.055	0.055	0.054
	W_{os}	0.046	0.046	0.056	0.05	0.051	0.053
	W_{onp}	**0.084**	**0.062**	**0.065**	0.055	0.054	0.054
3	W_o	**0.1**	**0.073**	**0.063**	0.049	0.052	0.051
	W_{os}	**0.039**	0.044	0.051	**0.042**	0.047	0.049
	W_{onp}	**0.092**	**0.07**	**0.062**	0.048	0.051	0.051
4	W_o	**0.127**	**0.09**	**0.068**	**0.063**	**0.058**	0.055
	W_{os}	**0.031**	0.044	0.049	0.054	0.051	0.052
	W_{onp}	**0.113**	**0.085**	**0.064**	**0.063**	**0.057**	0.054

Note: The bold text indicates that the empirical size falls outside the 95% confidence interval $(0.044, 0.056)$ of the nominal level 0.05.

4.3 Power Comparison of W_s and W_{os} in Testing SMH against Stochastic Order

For the case of one adverse event, Agresti (1983) showed that the test using ordinal scales outperforms the test ignoring its ordinal nature when they are used to test SMH against stochastic ordering. In this section, a simulation is performed to verify if the test using ordinal scales is also more powerful to test SMH against stochastic ordering for the multiple AEs case.

Based on the simulation results in section 4.1 and 4.2, it appears that the W_s and W_{os} performs the best. Thus, the W_s and W_{os} are contrasted in this section. To ensure the nominal level can be well controlled for W_s and W_{os}, the sample size of 100 and 200 are utilized in the simulation.

Their power performance is investigated by extending the simulation design introduced by Agresti (1983). We randomly sample from an underlying multivariate normal distribution having mean 0 and within-AE correlation $\rho_1 = 0.6$ and between-AE correlation $\rho_2 = 0.2$. A half of the dimensions of the multivarite random vector are divided as the severity levels $c = 1, 2, 3, 4$ of all AEs under treatment 1 and the other half are divided as the severity levels of all AEs under treatment 2. The boundries for AE categories under treatment 1 are set as -0.6, 0 and 0.6. The boundries for AE categories under treatment 2 are obtained by placing a shift $\Delta = 0.2$ relative to the boundaries for treatment 1. Hence, the boundaries of the AE categories under treatment 2 are -0.4, 0.2 and 0.8. The division produces the marginal probabilities of AE categories under treatment 1 to be $(0.2743, 0.2257, 0.2257, 0.2743)$, and the marginal probabilities of AE categories under treatment 2 to be $(0.3446, 0.2347, 0.2089, 0.2119)$. Our simulation includes the settings representing the combinations of sample size $n = 100, 200$ and number of AEs $K = 2, 3, 4$.

Table 3 shows the power of the W_s and W_{os} under the nominal levels of 0.05. From table 3, we have the following observations:

1. The power of the W_{os} is consistently greater than that of W_s for all the combinations of K and n we considered.
2. The ratio (1−power of the W_s)/(1−power of the W_{os}) is consistently greater at $n = 200$ than $n = 100$ for all K.
3. As K increases, the ratio (1−power of the W_s)/(1−power of the W_{os}) consistently increases for all n.

Therefore, the W_{os} outperforms the W_s with respect to the power when they are used to test SMH against stochastic order. Furthermore, as the sample size or number of AEs increases, the advantage of the W_{os} compared to the W_s increases.

Table 3. Empirical power of the W_{os} and W_s in 5000 simulated data sets under the nominal level of 0.05

Empirical Power						
Method	n=100	n=200	n=100	n=200	n=100	n=200
W_s	0.532	0.828	0.675	0.934	0.578	0.955
W_{os}	0.676	0.933	0.799	0.984	0.847	0.993

Note: The marginal probabilities of AE categories are $(0.2743, 0.2257, 0.2257, 0.2743)$ and $(0.3446, 0.2347, 0.2089, 0.2119)$ for treatment 1 and treatment 2, respectively.

5. Example

The data used in this article was part of a longitudinal study of criminal career patterns of former California youth authority wards between 1965 and 1984 (Haapanen, 1990). This study investigated the patterns of criminal behavior that occurred over 10 to 15 years for 1308 subjects whose early criminal involvement was serious enough to result in commitment to California Youth Authority institutions. The outcomes selected from this study are yearly arrest rates for 12 types of offenses, including murder, rape, felony assault, misdemeanor assault, armed robbery, strong-arm robbery, other personal offenses(extortion, kidnapping), burglary, receiving stolen property, grand theft, forgery and grand theft auto, over four four-year age blocks (18-21, 22-25, 26-29, 30-33). The 12 types of offenses are categorized according to the definition used in the earlier study (Haapanen, 1990), which are as follows:

Violent-aggressive: murder, rape, felony assault and misdemeanor assault.

Vilent-economic: armed robbery, strong-arm robbery and other personal offenses (extortion, kidnapping).

Property: burglary, receiving stolen property, grand theft, forgery and grand theft auto. Our main interest is to test if the yearly arrest rates for the three categories decline over time as indicated in Haapanen (1900). To apply the method we proposed, the yearly arrest rates for the 12 types of offenses are all converted to a scale of 1 to 4 (represents the offense rate as none, mild, moderate and severe), using the range: 0, (0, 1], (1, 2] and $(2, \infty)$. Scores $(0.1, 0.5, 1.5, 2.5)$ are assigned to the outcome categories which are approximately midpoints of the above ranges. Table 4 shows P-values from our method and Hotelling's Paired T^2 test. Note that the Hotelling's Paired T^2 test is applied to continuous variables before categorization.

Table 4. P-values for comparing the yearly arrest rates in four four-year age blocks, from W_{os} with scores $(0.1, 0.5, 1.5, 2.5)$ and Hotelling's Paired T^2 test

	P-values		
	$18-21$ vs $22-25$	$22-25$ vs $26-29$	$26-29$ vs $30-33$
Offense	W_{os} / $Hotelling's T^2$	W_{os} / $Hotelling's T^2$	W_{os} / $Hotelling's T^2$
Violent-aggressive	0.202 / 0.183	0.055 /0.317	0.097 /0.056
Violent-economic	**0.025 /0.322**	0.005 / 0.013	0.99/0.922
Property	0.000/ 0.046	0.000/ 0.007	0.000 / 0.004

Note: The bold text indicates that W_{os} and Hotelling's Paired T^2 test give different conclusion for the comparison.

From Table 4, W_{os} and Hotelling's Paired T^2 test give the same conclusion for all the comparisons except the property offense in age blocks $18-21$ and $22-25$. The difference might be due to the categorization and the scores assigned to the outcome categories. with scores $(0.1, 0.5, 1.5, 10)$, W_{os} gives P-value= 0.314 instead of 0.025 in Table 4. Therefore, careful consideration should be given to scores assigned to outcome categories before applying W_{os}.

More specifically, in table 4, W_{os} indicates the yearly arrest rate in age block $22-25$ is significantly lower than that of $18-21$, the yearly arrest rate in age block $26-29$ is significantly lower than that of $22-25$, for violent economic offense. For property offense, the yearly arrest rate in latter age block is always significantly lower than that of earlier age block. And significant difference is not found for other comparisons. Haapanen (1990) calculated the yearly arrest rates for three types of offenses by using the total number of arrests occurring to all subjects divided by the number of street years accumulated by all subjects together during the follow-up period. Table 5 (Haapanen, 1990) shows the yearly arrest rates for the three types of offenses over four four-year age blocks.

Table 5. Mean yearly arrest rates for active adult period by age block

Offense Type	$18-21$	$22-25$	$26-29$	$30-33$
Violent-aggressive	0.202	0.198	0.153	0.143
Violent-economic	0.169	0.125	0.094	0.120
Property	0.597	0.435	0.334	0.425

No statistical test was applied by Haapanen (1990) to investigate if the yearly arrest rate declines with age. Instead, by looking at the aggregated yearly arrest rates, it was declared that the arrest rates for the three categories of offenses over the first three age blocks showed the same general decline, and property and violent-economic arrest rates showed an increase for the last age block. The aggregation ignores the change over age blocks within subject. It gives valid conclusion only when the change in yearly arrest rate over age blocks is in the same direction for at least most of the subjects. Furthermore, the change is very slight for some comparisons, for instance, for violent aggressive offense in age blocks $18-21$ and $22-25$, which are 0.202 and 0.198, differ slightly. It might be caused by mere random chance rather than true difference in yearly arrest rates. By cross-checking the result from W_{os} in table 4 and table 5, it is found W_{os} gives P-value < 0.05 when the mean yearly arrest rates in two adjacent age blocks differ relatively large (> 0.03), which confirms, to a certain extent, that our method works properly.

6. Conclusion and Discussion

In this paper, we considered simultaneous testing of marginal homogeneity in matched-pair data with multivariate multinomial outcome. We developed three general test statistics and proved their asymptotic χ^2 distributions and then demonstrated their performance in small samples. In general, the sample size requirement for these tests seems rather large, as even at $n = 200$, the tests were too liberal, with the exception of a score-type test, which was rather conservative.

When the ordered nature of the adverse effects are taken into consideration, the power of the tests appear to improve, judging by the empirical sizes. Again none other than the score-type test controlled the nominal level, but the improvement by the additional information of ordering effects in the small sample problem is evident. The power to detect a small departure from the SMH assumption is also demonstrated, especially when the ordered information is taken into consideration.

The similarity between the two tests W and W_{np} (also in W_o and W_{onp}) in the simulation results are expected, because the data in our simulation was generated under the parametric distribution assumption for W. While W_{np} may appear to correct any departure from the assumption in the simulated data, it is essentially the same tests as W, because the only difference is in the variance of d, which is estimated by V and V_{np} under a multinomial assumption and distribution-free, respectively. However, in actual applications where data do not conform to the specified parametric distribution, it would make a difference, as W_{np} and W_{onp} do not force any parametric assumptions. For example, when the data exibits heterogeneity, severe sparsity, or overdispersion problems, we can anticipate the two non-parametrically based tests, W_{np} and W_{onp}, to outperform the others that are based on often unverifiable parametric assumptions. We did not pursue to demonstrate this latter aspect of the tests in simulations in this paper. Generating overdispersed multivariate multinomial outcome data in a matched pair is quite challenging, but we plan to investigate it further in our future research. As well, one possible way to deal with the sparsity of the data is to add a slightly restrictive condition, for instance, an equal correlation among the outcomes, which might be able to make the covariance matrix invertible.

In conclusion, we draw the following conclusions and recommendations:

1. For clustered matched-pair multinomial data, when $K = 2$, the W_s is recommended when sample size is at least 100. Moreover, When the the number of outcomes $K > 2$, it requires sample size larger than 200 to be used.
2. For clustered matched-pair ordinal data, when $K = 2$, the W_{os} can be used when sample size is 25. If $K > 2$, then W_{os} is suggested to use when sample size is at least 50.
3. If the outcomes considered are ordinal, the W_{os} is more powerful than W_s when the stochastic order holds in the two treatments.
4. If an overdispersion is suspected or when the data seem sparse, we recommend to use W_{np} or W_{onp}, especially when the sample size is large.