Statistics and Its Interface Volume 11 (2018) 353–368

Testing the equality of risk difference among multiple incomplete two-way contingency tables

Huiqiong Li[‡], Niansheng Tang[‡], Guoliang Tian[‡], and Hongyuan Cao[*,†]

Contingency tables are used to summarize categorical data with multiple attributes, which frequently arise in natural and social sciences. In Tian and Li (2015), the equality of risk difference based on a 2×2 table is tested at the presence of non-response. In this paper, we derive the joint distribution of multiple contingency tables with non-response. Consequently, we propose a new homogeneity test statistic for the risk difference among multiple contingency tables. The limiting distribution of the proposed test statistic is established along with inferential procedures. Upon rejection of the global null hypothesis of homogeneity, to identify contingency tables with discordance, a multiple comparison procedure is proposed. Numerical studies corroborate theoretical results. We illustrate our method with dataset from a psychiatric study.

AMS 2000 subject classifications: Primary 62H17; secondary 62H15.
Keywords and phrases: Hypothesis testing, Incomplete contingency table, Risk difference.

1. INTRODUCTION

In clinical trials and epidemiological studies, it is common to summarize data in contingency tables. Incomplete contingency tables arise frequently due to various reasons, such as early withdrawn, recording errors, side effects, etc. As a motivating example, we consider dataset from a multi-center psychiatric study, which has been analyzed in Molenberghs and Lesaffre (1994), Kenward *et al.* (1994), Molenberghs *et al.* (1997), Michiels and Molenberghs (1997), and Kenward *et al.* (2001). In this study, 315 patients with psychiatric symptoms were treated by fluvoxamine, an antipsychotic drug. We focus on the presence or absence of side effects at the first and last visit after enrollment into the study. The outcome presence or absence of therapeutic effects is evaluated on an independent group of 315 patients at the first and last visit as well. Observed counts are summarized in Table 1. From Table 1, we can see that there are quite a number of non-responses for both side effects and therapeutic effects, resulting in incomplete contingency tables. The total number of subjects 315 is fixed in advance. We are interested in whether the risk differences between first visit and last visit are the same for side effects and therapeutic effects in the presence of missing data. This can be extended to K contingency tables. If risk differences between first visit and last visit are not homogeneous among K contingency tables, we would like to identify specific heterogeneous contingency tables.

Statistical inferences for the above two questions are usually conducted through hypothesis testing and confidence interval. Statistical methods for incomplete contingency tables have received a lot of attention in recent years. For instance, Choi and Stablein (1982) derived a method of analyzing incomplete paired data where the mechanisms are considered to be independent of treatment. Tang *et al.* (2009) proposed exact and approximate unconditional test-based methods for constructing confidence intervals for proportion and rate differences in the presence of incomplete paired binary data. Miller and Looney (2012) proposed a weighted average method for estimating the odds ratio when the sample consists of a combination of complete and incomplete matched pairs. However, all these papers work with a sin-

Table 1. Observed counts for patients at the first and last visit after the treatment of fluvoxamine in a multi-center psychiatric study (Kenward et al., 2001)

	Side effect		
	The last visit		
The first visit	Yes	No	Non-response
Yes	89	13	26
No	57	65	49
Non-response	2	0	14
	Therapeutic effect		
	The last visit		
The first visit	Yes	No	Non-response
Yes	11	1	7
No	124	88	68
Non-response	0	2	14

*Corresponding author.
†This research is partially supported by the National Natural Science Foundation of China (No. 11561075).
‡This research is partially supported by the Science Foundation of Yunnan Province (2016FB005).

gle incomplete 2×2 table and the literature on hypothesis testing with multiple incomplete 2×2 tables is scarce.

In this paper, we develop a new method for testing the homogeneity of risk differences under two conditions in multiple incomplete two-way contingency tables. Tian and Li (2015) derived the joint distribution of the observed counts in an $r \times c$ incomplete contingency table with fixed total counts under the missing at random assumption (MAR) (Rubin, 1976). Based on the derived joint sampling distribution, they provided a new framework for analyzing incomplete contingency tables. We extend this earlier result to the case of multiple contingency tables by treating the nonresponse as a new category in the contingency table. We will establish the limiting distribution of new test statistics, and conduct statistical inference through bootstrap. Upon rejection of the homogeneity hypothesis, to identify heterogeneous contingency tables, we propose multiple comparison methods based on Bonferroni procedure, Single-step adjusted MaxT procedure and Single-step adjusted MinP procedure.

The remainder of this paper is organized as follows. Section 2 presents the proposed methods, their asymptotic distributions, bootstrap resampling methods for testing homogeneity of risk differences under two conditions in multiple incomplete two-way contingency tables. Several multiple comparison procedures are developed in Section 3. Simulation studies are conducted to investigate finite sample performance of various methods in Section 4. In Section 5, an example is used to illustrate the proposed methodology. Concluding remarks are given in Section 6.

2. HOMOGENEITY TEST METHODS

2.1 The joint distribution of the observed counts in multiple incomplete 2×2 tables

Consider K independent incomplete 2×2 tables, each having $N_k (k = 1, 2, \ldots, K)$ subjects. Suppose that for each patient in stratum k, we take two treatments (X, Y) corresponding to two correlated dichotomous variables. In the kth stratum, $p_{00k} = \Pr(X = 0, Y = 0)$, $p_{01k} = \Pr(X = 0, Y = 1)$, $p_{10k} = \Pr(X = 1, Y = 0)$, $p_{11k} = \Pr(X = 1, Y = 1)$, where $\sum_{i=0}^{1} \sum_{j=0}^{1} p_{ijk} = 1$. For stratum k, consider paired observations from a total of N_k (N_k is predetermined and non-random) subjects which are classified into two classes containing n_k complete counts and $m_{xk}+m_{yk}+m_{xyk}$ incomplete counts. These incomplete counts consist of three categories, where m_{xk} is the number of incomplete observations on X (or missing on Y), m_{yk} is the number of incomplete observations on Y (or missing on X), and m_{xyk} is the number of missing data on both X and Y for the kth stratum. The observed counts and cell probabilities are displayed in Table 2, where $N_k = n_k + m_{xk} + m_{yk} + m_{xyk}$ is fixed, while $n_k = \sum_{i=0}^{1} \sum_{j=0}^{1} n_{ijk}$, $m_{xk} = m_{1xk} + m_{0xk}$, $m_{yk} = m_{y1k} +$ m_{y0k} and $m_{xyk} = N_k - n_k - m_{xk} - m_{yk}$ are random. Let $Y_{\text{obs}} = \{n_{00k}, \ldots, n_{11k}; m_{1xk}, m_{0xk}; m_{y1k}, m_{y0k}; m_{xyk}\}$ be the observed frequencies and $\mathbf{p} = (p_{00k}, p_{01k}, p_{10k}, p_{11k})^\top \in \mathbb{T}_4$ be the cell probability vector, where $\mathbb{T}_p$ is defined as $\mathbb{T}_p \hat{=} \Big\{(\theta_1, \ldots, \theta_p)^\top \colon\ \theta_j \geq 0,\ j = 1, \ldots, p,\ \sum_{j=1}^{p} \theta_j = 1\Big\}$.

Table 2. The observed counts and cell probabilities from a multi-center study with incomplete observations

	$Y=0$	$Y=1$	Missing on Y
$X=0$	n_{00k} $((p_{00k})$	n_{01k} (p_{01k})	m_{0xk} $(p_{00k}+p_{01k})$
$X=1$	n_{10k} (p_{10k})	n_{11k} (p_{11k})	m_{1xk} $(p_{10k}+p_{11k})$
Missing on X	m_{y0k} $(p_{00k}+p_{10k})$	m_{y1k} $(p_{01k}+p_{11k})$	m_{xyk}

NOTE: The total number $N_k = \sum_{i=0}^{1} \sum_{j=0}^{1} n_{ijk} + m_{1xk} + m_{0xk} + m_{y1k} + m_{y0k} + m_{xyk} = n_k + m_{xk} + m_{yk} + m_{xyk}$ is fixed in advance.

Following Choi and Stablein (1988) and Chang (2009), we assume that the missing mechanism is MAR; i.e., the probability of missing only depends on the observed counts (Little and Robin, 2002). According to Tian and Li (2015), to obtain sampling distribution of the observed counts Y_{obs} in incomplete 2×2 tables for the kth stratum, we first introduce a missing mechanism random variable R with four categories, where $R = 1$ (or 12) with probability ϕ_{1k} if both the status of X and the status of Y are reported; $R = 2$ (or $1\bar{2}$) with probability ϕ_{2k} if only the status of X is reported; $R = 3$ (or $\bar{1}2$) with probability ϕ_{3k} if only the status of Y is reported; $R = 4$ (or $\bar{1}\bar{2}$) with probability ϕ_{4k} if neither the status of X nor the status of Y is reported. Hence, $R \sim \text{Categorical}_4(\boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\phi_{1k}, \ldots, \phi_{4k})^\top \in \mathbb{T}_4$ is called the parameter vector of the missing-data mechanism. Next, based on the two binary variables X and Y, we can construct a new four-category random variable Z as follows:

$$Z = \begin{cases} 1, & \text{if } (X, Y) = (0, 0), \\ 2, & \text{if } (X, Y) = (0, 1), \\ 3, & \text{if } (X, Y) = (1, 0), \\ 4, & \text{if } (X, Y) = (1, 1). \end{cases}$$

Thus, $Z \sim \text{Categorical}_4(\mathbf{p})$, where $\mathbf{p} \in \mathbb{T}_\mathbf{4}$ is called the model parameter vector. The joint distribution of R and Z is defined by $\boldsymbol{\pi} = (\pi_{ijk})$, where $\pi_{ijk} = \Pr(R = i, Z = j)$ in the kth stratum for $i, j = 1, \ldots, 4; k = 1, \ldots, K$. Table 3 shows the observed counts, missing counts, marginal probabilities of R and Z, and joint probabilities of (R, Z). For kth stratum, the full observations are $\{n_{ijk}\}(i, j = 0, 1)$ with corresponding cell probabilities $\{\pi_{1jk}\}_{j=1}^{4}$, while the latent counts are $\{n'_{jk}\}_{j=1}^{4}$, $\{n''_{jk}\}_{j=1}^{4}$ and $\{n'''_{jk}\}_{j=1}^{4}$ with corresponding cell probabilities $\{\pi_{2jk}\}_{j=1}^{4}$, $\{\pi_{3jk}\}_{j=1}^{4}$ and

Table 3. The observed counts, missing counts, marginal probabilities of R and Z, and joint probabilities of (R, Z)

M.M.R.V. R	Four–category variable Z: 1	2	3	4	M.P. of R	Observed counts
1 (or 12)	π_{11k} (n_{00k})	π_{12k} (n_{01k})	π_{13k} (n_{10k})	π_{14k} (n_{11k})	ϕ_{1k}	$n_k = \sum_{i=0}^{1}\sum_{j=0}^{1} n_{ijk}$
2 (or $1\bar{2}$)	$\pi_{21k}(n'_{1k})$	$\pi_{22k}(n'_{2k})$	$\pi_{23k}(n'_{3k})$	$\pi_{24k}(n'_{4k})$	ϕ_{2k}	$m_{0xk} = n'_{1k} + n'_{2k}$ $m_{1xk} = n'_{3k} + n'_{4k}$
3 (or $\bar{1}2$)	$\pi_{31k}(n''_{1k})$	$\pi_{32k}(n''_{2k})$	$\pi_{33k}(n''_{3k})$	$\pi_{34k}(n''_{4k})$	ϕ_{3k}	$m_{y0k} = n''_{1k} + n''_{3k}$ $m_{y1k} = n''_{2k} + n''_{4k}$
4 (or $\bar{1}\bar{2}$)	$\pi_{41k}(n'''_{1k})$	$\pi_{42k}(n'''_{2k})$	$\pi_{43k}(n'''_{3k})$	$\pi_{44k}(n'''_{4k})$	ϕ_{4k}	$m_{xyk} = \sum_{j=1}^{4} n'''_{jk}$
M.P. of Z	p_{00k}	p_{01k}	p_{10k}	p_{11k}	1	$N_k = n_k + m_{xk}$ $+m_{yk} + m_{xyk}$

NOTE: M.M.R.V. = missing mechanism random variable, M.P.= marginal probability. Only n'_{1k}, n'_{3k}, n''_{1k}, n''_{2k}, n'''_{1k}, n'''_{2k}, n'''_{3k} are missing while $n'_{2k} = m_{0xk} - n'_{1k}$, $n'_{4k} = m_{1xk} - n'_{3k}$, $n''_{3k} = m_{y0k} - n''_{1k}$, $n''_{4k} = m_{y1k} - n''_{2k}$ and $n'''_{4k} = m_{xyk} - n'''_{1k} - n'''_{2k} - n'''_{3k}$. $R \sim \text{Categorical}_4(\boldsymbol{\phi})$ and $Z \sim \text{Categorical}_4(\mathbf{p})$. $m_{xk} = m_{0xk} + m_{1xk}$, $m_{yk} = m_{y0k} + m_{y1k}$.

$\{\pi_{4jk}\}_{j=1}^4$, respectively. Note that only n'_{1k}, n'_{3k}, n''_{1k}, n''_{2k}, n'''_{1k}, n'''_{2k}, and n'''_{3k} are missing while $n'_{2k} = m_{0xk} - n'_{1k}$, $n'_{4k} = m_{1xk} - n'_{3k}$, $n''_{3k} = m_{y0k} - n''_{1k}$, $n''_{4k} = m_{y1k} - n''_{2k}$, and $n'''_{4k} = m_{xyk} - n'''_{1k} - n'''_{2k} - n'''_{3k}$. Thus the complete data

$$Y_{\text{com}} = Y_{\text{obs}} \cup \{n'_{1k}, n'_{3k}, n''_{1k}, n''_{2k}, n'''_{1k}, n'''_{2k}, n'''_{3k}\} = \{n_{00k}, n_{01k}, n_{10k}, n_{11k}; n'_{1k}, \ldots, n'_{4k}; n''_{1k}, \ldots, n''_{4k}; n'''_{1k}, \ldots, n'''_{4k}\}$$

follow a multinomial distribution with the following joint *probability mass function* (pmf)

$$\begin{aligned} f(Y_{\text{com}}|\boldsymbol{\pi}) &= \binom{N_k}{n_{00k}, \ldots, n_{11k}, n'_{1k}, \ldots, n'_{4k}, n''_{1k}, \ldots, n''_{4k}, n'''_{1k}, \ldots, n'''_{4k}} \\ &\times (\pi_{11k}^{n_{00k}} \pi_{12k}^{n_{01k}} \pi_{13k}^{n_{10k}} \pi_{14k}^{n_{11k}}) \left(\prod_{j=1}^{4} \pi_{2jk}^{n'_{jk}}\right)\left(\prod_{j=1}^{4} \pi_{3jk}^{n''_{jk}}\right)\left(\prod_{j=1}^{4} \pi_{4jk}^{n'''_{jk}}\right), \end{aligned}$$

$\boldsymbol{\pi} \in \mathbb{T}_{16}$.

Then the joint pmf of the observed data Y_{obs} can be obtained by summing over all missing data in $f(Y_{\text{com}}|\boldsymbol{\pi})$, yielding

$$\begin{aligned} f(Y_{\text{obs}}|\boldsymbol{\pi}) = C_1^{-1} \cdot (\pi_{11k}^{n_{00k}} \pi_{12k}^{n_{01k}} \pi_{13k}^{n_{10k}} \pi_{14k}^{n_{11k}}) (\pi_{21k} + \pi_{22k})^{m_{0xk}} \\ (\pi_{23k} + \pi_{24k})^{m_{1xk}} (\pi_{31k} + \pi_{33k})^{m_{y0k}} \times \\ (\pi_{32k} + \pi_{34k})^{m_{y1k}} (\textstyle\sum_{j=1}^{4} \pi_{4jk})^{m_{xyk}}, \quad \boldsymbol{\pi} \in \mathbb{T}_{16}, \end{aligned}$$

where

(1)
$$C_1^{-1} = \binom{N_k}{n_{00k}, n_{01k}, n_{10k}, n_{11k}, m_{0xk}, m_{1xk}, m_{y0k}, m_{y1k}, m_{xyk}}.$$

When R and Z are independent, the missing-data generation mechanism is said to be *ignorable* or MAR. Under MAR, the joint pmf in the kth stratum reduces to

$$\begin{aligned} f_1(Y_{\text{obs}}|\boldsymbol{\phi}, \mathbf{p}) &= C_1^{-1} \cdot [(\phi_{1k})^{n_k} p_{00k}^{n_{00k}} p_{01k}^{n_{01k}} p_{10k}^{n_{10k}} p_{11k}^{n_{11k}}] \\ &\times [\phi_{2k}(p_{00k} + p_{01k})]^{m_{0xk}} [\phi_{2k}(p_{10k} + p_{11k})]^{m_{1xk}} \\ &\times [\phi_{3k}(p_{00k} + p_{10k})]^{m_{y0k}} [\phi_{3k}(p_{01k} + p_{11k})]^{m_{y1k}} \\ &\times [\phi_{4k}(p_{00k} + p_{01k} + p_{10k} + p_{11k})]^{m_{xyk}} \end{aligned}$$

(2)
$$= C_1^{-1} \cdot \left(\phi_{1k}^{n_k} \phi_{2k}^{m_{xk}} \phi_{3k}^{m_{yk}} \phi_{4k}^{m_{xyk}}\right) \cdot L_1(\mathbf{p}|\mathbf{Y}_{\text{obs}}),$$

(3)
$$\boldsymbol{\phi} \in \mathbb{T}_4,$$

where C_1 is defined in (1),

$$\begin{aligned} L_1(\mathbf{p}|Y_{\text{obs}}) = p_{00k}^{n_{00k}} p_{01k}^{n_{01k}} p_{10k}^{n_{10k}} p_{11k}^{n_{11k}} (p_{00k} + p_{01k})^{m_{0xk}} \\ \times (p_{10k} + p_{11k})^{m_{1xk}} (p_{00k} + p_{10k})^{m_{y0k}} (p_{01k} + p_{11k})^{m_{y1k}}. \end{aligned}$$

Then (2) indicates that

(4)
$$Y_{\text{obs}}|(\boldsymbol{\phi}, \mathbf{p}) \sim \text{Multinomial}_{\mathbf{9}}(\mathbf{N_k}, \boldsymbol{\phi})$$

where $\boldsymbol{\phi} = (\phi_{1k}\, p_{00k}, \phi_{1k}\, p_{01k}, \phi_{1k}\, p_{10k}, \phi_{1k}\, p_{11k}, \phi_{2k}(p_{00k} + p_{01k}), \phi_{2k}(p_{10k} + p_{11k}), \phi_{3k}(p_{00k} + p_{10k}), \phi_{3k}(p_{01k} + p_{11k}), \phi_{4k})^\top$ with only 6 free parameters. In other words, (4) is a special 9-dimensional multinomial distribution with different equality constraint on each component of $\boldsymbol{\phi}$.

2.2 Homogeneity test of risk differences

Denote $\delta_k = (p_{00k}+p_{10k})-(p_{00k}+p_{01k}) = p_{10k}-p_{01k}$ ($k = 1, \ldots, K$), which can be interpreted as the risk difference of two clinical entities (e.g., the first visit and the last visit as introduced in section 1) in the kth stratum. Here, our main interest is to test the following hypothesis:

$$(5)\qquad \begin{aligned} H_0: &\quad \delta_1 = \delta_2 = \cdots = \delta_K \triangleq \tau \quad \text{vs} \\ H_1: &\quad \delta_{j_1} \neq \delta_{j_2} \text{ for some } j_1 \neq j_2 \in \{1, 2, \ldots, K\}, \end{aligned}$$

where $\tau \in [-1, 1]$ is an unknown parameter. In the following, we will develop three most commonly used testing method: likelihood ratio test, score test and Wald test.

2.2.1 Likelihood ratio test

Let $\hat{p}_{ijk}$ be the maximum likelihood estimator (MLE) of p_{ijk} $(i, j = 0, 1)$ in the kth stratum. It follows from Campbell (1984) and Chang (2009) that MLEs of p_{00k}, p_{01k}, p_{10k} and p_{11k} satisfy the following equations:

$$\hat{p}_{00k} = \frac{\{n_{00k} + m_{0xk}\hat{p}_{00k}/(\hat{p}_{00k} + \hat{p}_{01k}) + m_{y0k}\hat{p}_{00k}/(\hat{p}_{00k} + \hat{p}_{10k})\}}{N_k},$$

$$\hat{p}_{01k} = \frac{\{n_{01k} + m_{0xk}\hat{p}_{01k}/(\hat{p}_{00k} + \hat{p}_{01k}) + m_{y1k}\hat{p}_{01k}/(\hat{p}_{01k} + \hat{p}_{11k})\}}{N_k},$$

$$\hat{p}_{10k} = \frac{\{n_{10k} + m_{y0k}\hat{p}_{10k}/(\hat{p}_{00k} + \hat{p}_{10k}) + m_{1xk}\hat{p}_{10k}/(\hat{p}_{10k} + \hat{p}_{11k})\}}{N_k},$$

$$\hat{p}_{11k} = \frac{\{n_{11k} + m_{y1k}\hat{p}_{11k}/(\hat{p}_{01k} + \hat{p}_{11k}) + m_{1xk}\hat{p}_{11k}/(\hat{p}_{10k} + \hat{p}_{11k})\}}{N_k}.$$

Thus, $\hat{p}_{00k}$, $\hat{p}_{01k}$, $\hat{p}_{10k}$ and $\hat{p}_{11k}$ can be obtained via the EM algorithm (Dempter *et al.*, 1977). It can be shown from (2) that the MLEs of $\phi_{1k}, \phi_{2k}, \phi_{3k}$ and ϕ_{4k} are given by $\hat{\phi}_{1k} = n_k/N_k, \hat{\phi}_{2k} = m_{xk}/N_k$, $\hat{\phi}_{3k} = m_{yk}/N_k$, and $\hat{\phi}_{4k} = m_{xyk}/N_k$, respectively.

Let $\tilde{p}_{ijk}$ be the constrained MLE of p_{ijk} under H_0: $p_{10k} - p_{01k} = \tau$ for $i, j = 0, 1$, $k = 1, \ldots, K$. Thus, it follows from (2) that $\tilde{p}_{00k}$, $\tilde{p}_{01k}$ and $\tilde{\tau}$ under H_0 are obtained by solving the following $2K + 1$ equations:

(6)
$$\begin{cases} \dfrac{n_{00k}}{\tilde{p}_{00k}} - \dfrac{n_{11k}}{1 - \tilde{p}_{00k} - 2\tilde{p}_{01k} - \tilde{\tau}} + \dfrac{m_{0xk}}{\tilde{p}_{00k} + \tilde{p}_{01k}} - \dfrac{m_{1xk}}{1 - \tilde{p}_{00k} - \tilde{p}_{01k}} \\ \quad + \dfrac{m_{y0k}}{\tilde{p}_{00k} + \tilde{p}_{01k} + \tilde{\tau}} - \dfrac{m_{y1k}}{1 - \tilde{p}_{00k} - \tilde{p}_{01k} - \tilde{\tau}} = 0, \\ \dfrac{n_{01k}}{\tilde{p}_{01k}} + \dfrac{n_{10k}}{\tilde{p}_{01k} + \tilde{\tau}} - \dfrac{2n_{11k}}{1 - \tilde{p}_{00k} - 2\tilde{p}_{01k} - \tilde{\tau}} + \dfrac{m_{0xk}}{\tilde{p}_{00k} + \tilde{p}_{01k}} - \\ \dfrac{m_{1xk}}{1 - \tilde{p}_{00k} - \tilde{p}_{01k}} + \dfrac{m_{y0k}}{\tilde{p}_{00k} + \tilde{p}_{01k} + \tilde{\tau}} - \dfrac{m_{y1k}}{1 - \tilde{p}_{00k} - \tilde{p}_{01k} - \tilde{\tau}} = 0, \\ \displaystyle\sum_{k=1}^{K} \{ \dfrac{n_{10k}}{\tilde{p}_{01k} + \tilde{\tau}} - \dfrac{n_{11k}}{1 - \tilde{p}_{00k} - 2\tilde{p}_{01k} - \tilde{\tau}} + \dfrac{m_{y0k}}{\tilde{p}_{00k} + \tilde{p}_{01k} + \tilde{\tau}} \\ \quad - \dfrac{m_{y1k}}{1 - \tilde{p}_{00k} - \tilde{p}_{01k} - \tilde{\tau}} \} = 0 \end{cases}$$

Note that there are no closed-form solutions for $\tilde{\tau}$, $\tilde{p}_{00k}$ and $\tilde{p}_{01k}$ $(k = 1, 2, \ldots, K)$, which can be obtained by iteratively solving the above equations via the Fisher scoring algorithm. Then, the likelihood ratio statistic for testing $H_0: \delta_1 = \cdots = \delta_K = \tau$ is given by

$$T_l = 2\{l(\hat{\mathbf{p}}) - l(\tilde{\mathbf{p}}_0, \tilde{\tau})\},$$

which is asymptotically distributed as the χ^2 distribution with $K - 1$ degree of freedom, where $l(\hat{\mathbf{p}}) = \sum_{k=1}^{K} l_k(\hat{\mathbf{p}}_k)$ with $l_k(\hat{\mathbf{p}}_k) = n_{00k}\log(\hat{p}_{00k}) + n_{01k}\log(\hat{p}_{01k}) + n_{10k}\log(\hat{p}_{10k}) + n_{11k}\log(\hat{p}_{11k}) + m_{0xk}log(\hat{p}_{00k} + \hat{p}_{01k}) + m_{1xk}log(1 - \hat{p}_{00k} - \hat{p}_{01k}) + m_{y0k}log(\hat{p}_{00k} + \hat{p}_{10k}) + m_{y1k}log(1 - \hat{p}_{00k} - \hat{p}_{10k})$+constant, $l(\tilde{\mathbf{p}}_0, \tilde{\tau}) = \sum_{k=1}^{K} l_k(\tilde{\mathbf{p}}_{0k}, \tilde{\tau})$ with $l_k(\tilde{\mathbf{p}}_{0k}, \tilde{\tau}) = n_{00k}\log(\tilde{p}_{00k}) + n_{01k}\log(\tilde{p}_{01k}) + n_{10k}\log(\tilde{p}_{01k} + \tilde{\tau}) + n_{11k}\log(1 - \tilde{p}_{00k} - 2\tilde{p}_{01k} - \tilde{\tau}) + m_{0xk}\log(\tilde{p}_{00k} + \tilde{p}_{01k}) + m_{1xk}\log(1 - \tilde{p}_{00k} - \tilde{p}_{01k}) + m_{y0k}\log(\tilde{p}_{00k} + \tilde{p}_{01k} + \tilde{\tau}) + m_{y1k}\log(1 - \tilde{p}_{00k} - \tilde{p}_{01k} - \tilde{\tau})$+constant, for $k = 1, \ldots, K$. We reject H_0 at significance level α if $T_l \geq \chi^2_{K-1,\alpha}$, where $\chi^2_{K-1,\alpha}$ is the upper α percentile of the χ^2 distribution with $K-1$ degrees of freedom. Rejecting H_0 implies that ignoring stratification is unreasonable.

2.2.2 Score test

Following the arguments of Tang et al. (2016), it follows from (2) that the score function of log-likelihood with respect to τ under H_0 is

$$S(\tilde{\mathbf{p}}_0) = \frac{\partial l}{\partial \tau} = \sum_{k=1}^{K} \{ \frac{n_{10k}}{\tilde{p}_{01k} + \tilde{\tau}} - \frac{n_{11k}}{1 - \tilde{p}_{00k} - 2\tilde{p}_{01k} - \tilde{\tau}}$$

$$+ \frac{m_{y0k}}{\tilde{p}_{00k} + \tilde{p}_{01k} + \tilde{\tau}} - \frac{m_{y1k}}{1 - \tilde{p}_{00k} - \tilde{p}_{01k} - \tilde{\tau}} \},$$

where $\tilde{\tau}$, $\tilde{p}_{00k}$ and $\tilde{p}_{01k}$ $(k = 1, 2, \ldots, K)$ are given in (6). The Fisher information matrix with respect to τ, p_{00k} and p_{01k} under H_0 is given by

$$\mathbf{I} = \begin{pmatrix} I_{11} & I_{12} & I_{13} & I_{14} & I_{15} \\ I_{12} & I_{22} & I_{23} & 0 & 0 \\ I_{13} & I_{23} & I_{33} & 0 & 0 \\ I_{14} & 0 & 0 & I_{44} & I_{45} \\ I_{15} & 0 & 0 & I_{45} & I_{55} \end{pmatrix},$$

where $I_{11} = \sum_{k=1}^{2} (N_{10k} + N_{11k} + b_k)$, $I_{12} = N_{111} + b_1$, $I_{13} = N_{10} + 2N_{11} + b_1$, $I_{14} = N_{112} + b_2$, $I_{15} = N_{102} + 2N_{112} + b_2$, $I_{22} = N_{001} + N_{111} + a_1 + b_1$, $I_{23} = 2N_{111} + a_1 + b_1$, $I_{33} = N_{011} + N_{101} + 4N_{111} + a_1 + b_1$, $I_{44} = N_{002} + N_{112} + a_2 + b_2$, $I_{45} = 2N_{112} + a_2 + b_2$, $I_{55} = N_{012} + N_{102} + 4N_{112} + a_2 + b_2$, where $N_{ijk} = n_k/p_{ijk}$ for $i, j = 0, 1$, $a_k = m_{xk}/\{(p_{00k} + p_{01k})(1 - p_{00k} - p_{01k})\}$, $b_k = m_{yk}/\{(p_{00k} + p_{01k} + \delta_k)(1 - p_{00k} - p_{01k} - \delta_k)\}$, $p_{10k} = p_{01k} + \delta_k$, and $p_{11k} = 1 - p_{00k} - 2p_{01k} - \delta_k$. It follows that the upper left element I^{11} of $\mathbf{I}^{-1}$ can be calculated. The score statistic for testing $H_0 : \delta_1 = \delta_2 =$

$\cdots = \delta_K = \tau$ is given by

$$T_s = \sum_{k=1}^{K} \{ \frac{n_{10k}}{\tilde{p}_{01k} + \tilde{\tau}} - \frac{n_{11k}}{1 - \tilde{p}_{00k} - 2\tilde{p}_{01k} - \tilde{\tau}} + \frac{m_{y0k}}{\tilde{p}_{00k} + \tilde{p}_{01k} + \tilde{\tau}}$$

$$- \frac{m_{y1k}}{1 - \tilde{p}_{00k} - \tilde{p}_{01k} - \tilde{\tau}} \}^2 I^{11}$$

which is asymptotically distributed as the χ^2 distribution with $K-1$ degrees of freedom. We reject H_0 at significance level α if $T_s \geq \chi^2_{K-1,\alpha}$.

2.2.3 Wald test

Testing hypothesis $H_0 : \delta_1 = \delta_2 = \cdots = \delta_K = \tau$ is equivalent to testing the following hypothesis $H_0' : \boldsymbol{A\delta} = \mathbf{0}$, where $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_K)^T$, $\mathbf{0} = (0, 0, \cdots, 0)^T$ and

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}_{(K-1)\times K} .$$

The naïve MLE of $\boldsymbol{\delta}$ is given by $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_K)^T$. Since variance of $\hat{\boldsymbol{\delta}}$ is given by var($\hat{\boldsymbol{\delta}}$) =diag(var($\hat{\delta}_1$), ..., var($\hat{\delta}_K$)), an estimate of var($\hat{\boldsymbol{\delta}}$) under H_0 is given by $\widehat{\text{var}}(\hat{\boldsymbol{\delta}})$ =diag($\widehat{\text{var}}(\hat{\delta}_1|H_0), \ldots, \widehat{\text{var}}(\hat{\delta}_K|H_0)$), where

$$\widehat{\text{Var}}(\hat{\delta}_k|H_0) = \widehat{\text{Var}}(\hat{p}_{00k} + \hat{p}_{01k}) + \widehat{\text{Var}}(\hat{p}_{00k} + \hat{p}_{10k})$$

$$-2\frac{\hat{\phi}_{1k}}{N_k \hat{A}_{0k}} (\hat{p}_{00k}\hat{p}_{11k} - \hat{p}_{01k}\hat{p}_{10k}),$$

$$\widehat{\text{Var}}(\hat{p}_{00k} + \hat{p}_{01k}) = \frac{1}{N_k \hat{A}_{0k}} \Big[\hat{\phi}_{1k}\hat{\mathbb{D}}_{1k} + \hat{\phi}_{2k}\frac{\hat{\mathbb{D}}_{3k}}{\hat{\mathbb{D}}_{2k}} \Big],$$

and

$$\widehat{\text{Var}}(\hat{p}_{00k} + \hat{p}_{10k}) = \frac{1}{N_k \hat{A}_{0k}} \Big[\hat{\phi}_{1k}\hat{\mathbb{D}}_{2k} + \hat{\phi}_{3k}\frac{\hat{\mathbb{D}}_{3k}}{\hat{\mathbb{D}}_{1k}} \Big],$$

where $\hat{\mathbb{D}}_{1k} = (\hat{p}_{00k} + \hat{p}_{01k})(1 - \hat{p}_{00k} - \hat{p}_{01k})$, $\hat{\mathbb{D}}_{2k} = (\hat{p}_{00k} + \hat{p}_{10k})(1 - \hat{p}_{00k} - \hat{p}_{10k})$, $\hat{\mathbb{D}}_{3k} = \hat{p}_{00k}\hat{p}_{01k}(\hat{p}_{11k} + \hat{p}_{01k}) + (\hat{p}_{00k} + \hat{p}_{01k})\hat{p}_{10k}\hat{p}_{11k}$, and $\hat{A}_{0k} = \hat{\phi}_{1k} + \hat{\phi}_{2k}\hat{\phi}_{3k}\hat{\mathbb{D}}_{3k}/(\hat{\mathbb{D}}_{1k}\hat{\mathbb{D}}_{2k})$ for $k = 1, \ldots, K$. Thus, the Wald-type statistic for testing $H_0 : \delta_1 = \cdots = \delta_K = \tau$ can be expressed as

$$T_w = \hat{\boldsymbol{\delta}}^T \boldsymbol{A}^T (\mathbf{A}\widehat{\text{var}}(\hat{\boldsymbol{\delta}})\boldsymbol{A}^{\mathbf{T}})^{-1} \boldsymbol{A}\hat{\boldsymbol{\delta}},$$

which is asymptotically distributed as the χ^2 distribution with $K-1$ degrees of freedom. We reject H_0 at significance level α if $T_w \geq \chi^2_{K-1,\alpha}$.

2.3 Statistical inference through bootstrap

The condition for using the above developed asymptotic procedures to test hypotheses $H_0 : \delta_1 = \delta_2 = \cdots = \delta_K = \tau$ is that $N_1, \ldots, N_K$ are sufficiently large. In practical applications, it is rather difficult to satisfy the above condition. In this case, using the above developed asymptotic test procedures to test hypotheses H_0 may lead to inaccurate results. Therefore, a nonparametric bootstrap resampling method is developed to solve the above mentioned difficulties in this subsection.

Given the observed data $Y_{\text{obs}} = \{n_{00k}, \ldots, n_{11k}; m_{1xk}, m_{0xk}; m_{y1k}, m_{y0k}; m_{xyk}; k = 1, 2, \ldots, K\}$, the naïve MLEs $\hat{p}_{ijk}$ of p_{ijk} are obtained through the EM algorithm for $i, j = 0, 1$ and $k = 1, 2, \ldots, K$. Also, the observed value, denoted by t_L, of statistic T_L $(L = l, s, w)$ can be obtained from available data. For $k = 1, 2, \ldots, K$, based on $\hat{p}_{00k}, \hat{p}_{01k}, \hat{p}_{10k}, \hat{p}_{11k}$, $\phi_{1k} = n_k/N_k, \phi_{2k} = m_{xk}/N_k, \phi_{3k} = m_{yk}/N_k$, and $\phi_{4k} = m_{xyk}/N_k$, we can independently generate

$$Y^{(b)}_{\text{obs,k}} = \Big\{ \mathbf{N}^{(b)} = (n^{(b)}_{ijk}),\ \mathbf{m}^{(b)}_x = (m^{(b)}_{1xk}, m^{(b)}_{0xk})^\top, \\ \mathbf{m}^{(b)}_y = (m^{(b)}_{y1k}, m^{(b)}_{y0k})^\top,\ m^{(b)}_{xyk} \Big\} \sim \text{Multinomial}_9(N_k, \hat{\boldsymbol{\phi}}),$$

where $\hat{\boldsymbol{\phi}} = (\hat{\phi}_{1k}\,\hat{p}_{00k}, \hat{\phi}_{1k}\,\hat{p}_{01k}, \hat{\phi}_{1k}\,\hat{p}_{10k}, \hat{\phi}_{1k}\,\hat{p}_{11k}, \hat{\phi}_{2k}(\hat{p}_{00k} + \hat{p}_{01k}), \hat{\phi}_{2k}(\hat{p}_{10k} + \hat{p}_{11k}), \hat{\phi}_{3k}(\hat{p}_{00k} + \hat{p}_{10k}), \hat{\phi}_{3k}(\hat{p}_{01k} + \hat{p}_{11k}), \hat{\phi}_{4k})^\top$. For each generated $\boldsymbol{Y}^{(b)}_{obs} = \{\boldsymbol{Y}^{(b)}_{obs,1}, \boldsymbol{Y}^{(b)}_{obs,2}, \ldots, \boldsymbol{Y}^{(b)}_{obs,K}\}$, we can calculate a bootstrap replicate $t^{(b)}_L$ of statistic T_L $(L = l, s, w)$. Repeating this process B times, we obtain B bootstrap replicates $\{t^{*(b)}_L\}^B_{b=1}$. Thus, the p-value for testing hypotheses $H_0 : \delta_1 = \delta_2 = \cdots = \delta_T = \tau$ via statistic T_L $(L = l, s, w)$ can be computed by

$$\hat{p}_B = \frac{1}{B}\sum_{b=1}^{B} I(t^{(b)}_L \geq t_L),$$

where $I(\cdot)$ is an indictor function which is 1 when $t^{(b)}_L \geq t_L$, and 0 otherwise.

3. MULTIPLE COMPARISON PROCEDURE

We test $H_0 : \delta_1 = \delta_2 = \cdots = \delta_K = \tau_0$ vs $H_1 : \delta_k \neq \tau_0$ for some $k \in \{1, \cdots, K\}$, where τ_0 is a fixed constant. If H_0 is rejected, there is at least one $k \in \{1, \cdots, K\}$ such that $\delta_k \neq \tau_0$. We would like to identify such heterogeneous strata. Towards this goal, we consider the following multiple testing problem:

$$H_{k0} : \delta_k = \tau_0 \quad \text{vs} \quad H_{k1} : \delta_k \neq \tau_0 \quad \text{for } k = 1, 2, \cdots, K.$$

At significance level α, $H_{k0}, k = 1, \ldots, K$ is rejected if the corresponding p-value is smaller than α/K by Bonferroni method. If none of the p-values corresponding to $H_{k0}, k = 1, \ldots, K$ is smaller than α/K, we fail to reject the global null H_0.

In this section, we shall propose three statistics for the multiple testing of H_{k0} based on the likelihood ratio test, score test, and Wald-type test.

3.1 Test statistics

3.1.1 Likelihood ratio statistic

The likelihood ratio statistic for testing $H_{k0}: \delta_k = \tau_0$ is given by

$$T_{lk} = 2[l_k(\hat{p}_{00k}, \hat{p}_{01k}, \hat{\delta}_k) - l_k(\tilde{p}^*_{00k}, \tilde{p}^*_{01k}, \tau_0)],$$

which is asymptotically distributed as the χ^2 distribution with one degree of freedom under $H_{0k}: \delta_k = \tau_0$, where

$$\begin{aligned}
&l_k(\hat{p}_{00k}, \hat{p}_{01k}, \hat{\delta}_k) = n_{00k}\log(\hat{p}_{00k}) + n_{01k}\log(\hat{p}_{01k}) \\
&+n_{10k}\log(\hat{p}_{01k} + \hat{\delta}_k) + n_{11k}\log(1 - \hat{p}_{00k} - 2\hat{p}_{01k} - \hat{\delta}_k) \\
&+m_{0xk}\log(\hat{p}_{00k} + \hat{p}_{01k}) + m_{1xk}\log(1 - \hat{p}_{00k} - \hat{p}_{01k}) + m_{y0k} \\
&\log(\hat{p}_{00k} + \hat{p}_{01k} + \hat{\delta}_k) + m_{y1k}\log(1 - \hat{p}_{00k} - \hat{p}_{01k} - \hat{\delta}_k), \\
&l_k(\tilde{p}^*_{00k}, \tilde{p}^*_{01k}, \tau_0) = n_{00k}\log(\tilde{p}^*_{00k}) + n_{01k}\log(\tilde{p}^*_{01k}) + \\
&n_{10k}\log(\tilde{p}^*_{01k} + \tau_0) + n_{11k}\log(1 - \tilde{p}^*_{00k} - 2\tilde{p}^*_{01k} - \tau_0) + \\
&m_{0xk}\log(\tilde{p}^*_{00k} + \tilde{p}^*_{01k}) + m_{1xk}\log(1 - \tilde{p}^*_{00k} - \tilde{p}^*_{01k}) + m_{y0k} \\
&\log(\tilde{p}^*_{00k} + \tilde{p}^*_{01k} + \tau_0) + m_{y1k}\log(1 - \tilde{p}^*_{00k} - \tilde{p}^*_{01k} - \tau_0),
\end{aligned}$$

where $\hat{p}_{00k}, \hat{p}_{01k}, \hat{\delta}_k = \hat{p}_{10k} - \hat{p}_{01k}$ can be obtained as in section 2.2.1, $\tilde{p}^*_{00k}, \tilde{p}^*_{01k}$ are the solutions of the following equations:

$$\begin{cases}
\dfrac{n_{00k}}{\tilde{p}^*_{00k}} - \dfrac{n_{11k}}{1 - \tilde{p}^*_{00k} - 2\tilde{p}^*_{01k} - \tau_0} + \dfrac{m_{0xk}}{\tilde{p}^*_{00k} + \tilde{p}^*_{01k}} \\
- \dfrac{m_{1xk}}{1 - \tilde{p}^*_{00k} - \tilde{p}^*_{01k}} + \dfrac{m_{y0k}}{\tilde{p}^*_{00k} + \tilde{p}^*_{01k} + \tau_0} \\
- \dfrac{m_{y1k}}{1 - \tilde{p}^*_{00k} - \tilde{p}^*_{01k} - \tau_0} = 0, \\
\dfrac{n_{01k}}{\tilde{p}^*_{01k}} + \dfrac{n_{10k}}{\tilde{p}^*_{01k} + \tau_0} - \dfrac{2n_{11k}}{1 - \tilde{p}^*_{00k} - 2\tilde{p}^*_{01k} - \tau_0} + \dfrac{m_{0xk}}{\tilde{p}^*_{00k} + \tilde{p}^*_{01k}} \\
- \dfrac{m_{1xk}}{1 - \tilde{p}^*_{00k} - \tilde{p}^*_{01k}} + \dfrac{m_{y0k}}{\tilde{p}^*_{00k} + \tilde{p}^*_{01k} + \tau_0} \\
- \dfrac{m_{y1k}}{1 - \tilde{p}^*_{00k} - \tilde{p}^*_{01k} - \tau_0} = 0.
\end{cases}$$

Fisher scoring method can be used to iteratively solve the above equations to obtain $\tilde{p}^*_{00k}$ and $\tilde{p}^*_{01k}$.

3.1.2 Score statistic

The score statistic for testing H_{k0} is given by

$$T_{sk} = \{\frac{n_{10k}}{\tilde{p}^*_{01k} + \tau_0} - \frac{n_{11k}}{1 - \tilde{p}^*_{00k} - 2\tilde{p}^*_{01k} - \tau_0} + \frac{m_{y0k}}{\tilde{p}^*_{00k} + \tilde{p}^*_{01k} + \tau_0}$$

$$- \frac{m_{y1k}}{1 - \tilde{p}^*_{00k} - \tilde{p}^*_{01k} - \tau_0}\}\sqrt{\frac{\tilde{\mathbb{A}}^*_2 + (\tilde{a}^* + \tilde{b}^*)\tilde{\mathbb{A}}^*_1}{\tilde{\mathbb{B}}^*_1 + \tilde{\mathbb{A}}^*_1\tilde{a}^*\tilde{b}^* + \tilde{\mathbb{B}}^*_2\tilde{a}^* + \tilde{\mathbb{B}}^*_3\tilde{b}^*}},$$

where $\tilde{N}^*_{ij} = n_k/\tilde{p}^*_{ijk}$ for $i,j = 0,1$, $\tilde{a}^* = m_{xk}/\{(\tilde{p}^*_{00k} + \tilde{p}^*_{01k})(1 - \tilde{p}^*_{00k} - \tilde{p}^*_{01k})\}$, $\tilde{b}^* = m_{yk}/\{(\tilde{p}^*_{00k} + \tilde{p}^*_{01k} + \tau_0)(1 - \tilde{p}^*_{00k} - \tilde{p}^*_{01k} - \tau_0)\}$, $\tilde{p}^*_{10k} = \tilde{p}^*_{01k} + \tau_0$, $\tilde{p}^*_{11k} = 1 - \tilde{p}^*_{00k} - 2\tilde{p}^*_{01k} - \tau_0$, $\tilde{\mathbb{A}}^*_1 = \sum_{i=0}^1\sum_{j=0}^1 \tilde{N}_{ij}$, $\tilde{\mathbb{A}}^*_2 = (\tilde{N}^*_{00} + \tilde{N}^*_{11})(\tilde{N}^*_{01} + \tilde{N}^*_{10}) + 4\tilde{N}^*_{00}\tilde{N}^*_{11}$, $\tilde{\mathbb{B}}^*_1 = \tilde{N}^*_{00}\tilde{N}^*_{01}\tilde{N}^*_{1+} + \tilde{N}^*_{10}\tilde{N}^*_{11}\tilde{N}^*_{0+}$ where $\tilde{N}^*_{1+} = \tilde{N}^*_{10} + \tilde{N}^*_{11}$, $\tilde{N}^*_{0+} = \tilde{N}^*_{00} + \tilde{N}^*_{01}$, $\tilde{\mathbb{B}}^*_2 = \tilde{N}^*_{0+}\tilde{N}^*_{1+}$, and $\tilde{\mathbb{B}}^*_3 = (\tilde{N}^*_{00} + \tilde{N}^*_{10})(\tilde{N}^*_{01} + \tilde{N}^*_{11})$. Under $H_{k0}: \delta_k = \tau_0$, T_{sk} is asymptotically distributed as standard normal distribution.

3.1.3 Wald-type statistic

It follows from Chang (2009) that under H_{0k}, the asymptotic mean of $\hat{\delta}_k$ is given by $E(\hat{\delta}_k) \approx \delta_k$, and the asymptotic variance of $\hat{\delta}_k$ can be estimated by

$$\widehat{\mathrm{Var}}(\hat{\delta}_k) = \widehat{\mathrm{Var}}(\hat{p}_{00k} + \hat{p}_{10k}) + \widehat{\mathrm{Var}}(\hat{p}_{00k} + \hat{p}_{10k})$$

$$-2\frac{\hat{\phi}_{1k}}{N_k\hat{A}_{0k}}(\hat{p}_{00k}\hat{p}_{11k} - \hat{p}_{01k}\hat{p}_{10k}),$$

Hence, the Wald-type statistic for testing H_{k0} is given by $T_{wk} = (\hat{\delta}_k - \tau_0)/\sqrt{\widehat{\mathrm{Var}}(\hat{\delta}_k)}$, which is asymptotically distributed as the standard normal distribution under $H_{k0}: \delta_k = \tau_0$.

3.2 Testing procedures

In this subsection, several multiple comparison procedures are proposed to test the hypothesis $H_0: \delta_1 = \delta_2 = \cdots = \delta_K = \tau_0$ vs $H_1: \delta_k \neq \tau_0$ for some $k \in \{1, \cdots, K\}$.

3.2.1 Bonferroni procedure

Following Westfall and Young (1993) and Hochberg and Tamhane (1997), we reject the null hypothesis $H_{k0}: \delta_k = \tau_0$ if $T_{rk}(r = l, s, w)$ is greater than the critical value c ($k = 1, 2, \ldots, K$). In this case, we can define the p-value for controlling the family-wise error rate as follows:

$$p = \mathrm{P}(\max_{k=1,\ldots,K}|T_{rk}| > \mathrm{c} \mid H_0),$$

where the critical value c can be taken to be $z_{\alpha/2K}$, which is the upper $\alpha/2K$-percentile of the standard normal distribution. According to the Bonferroni procedure, one rejects $H_0: \delta_1 = \delta_2 = \cdots = \delta_K = \tau_0$ if $\max_{k=1,\ldots,K}|t_{rk}| > z_{\alpha/2K}$, where t_{rk} is the observed value of statistic T_{rk}.

3.2.2 Single-step adjusted MaxT procedure

It is well known that the Bonferroni procedure is rather conservative. The p-value adjustment procedure is one of the most commonly used alternatives for multiple hypothesis testing. we consider the following single-step adjusted MaxT procedure.

Step 1. Compute the observed values $t_{r1}, \ldots, t_{rK}$ of test statistics $T_{r1}, \ldots, T_{rK}(r = l, s, w)$ based on the original data.

Step 2. Given the MLEs $\hat{p}_{00k}$, $\hat{p}_{01k}$, $\hat{\delta}_k$, $\hat{\phi}_{1k}$, $\hat{\phi}_{2k}$, $\hat{\phi}_{3k}$, $\hat{\phi}_{4k}$ $(k=1,\ldots,K)$ for the original data, we generate B bootstrap samples

$$\{n_{00k}^{(b)},\ldots,n_{11k}^{(b)};m_{1xk}^{(b)},m_{0xk}^{(b)};m_{y1k}^{(b)},m_{y0k}^{(b)};m_{xyk}^{(b)};b=1,\ldots,B\}$$

$$\sim \text{Multinomial}_9(N_k,\hat{\boldsymbol{\phi}}),$$

where $\hat{\boldsymbol{\phi}}=(\hat{\phi}_{1k}\,\hat{p}_{00k},\hat{\phi}_{1k}\,\hat{p}_{01k},\hat{\phi}_{1k}\,(\hat{p}_{01k}+\hat{\delta}_k),\hat{\phi}_{1k}\,(1-\hat{p}_{00k}-2\hat{p}_{01k}-\hat{\delta}_k),\hat{\phi}_{2k}(\hat{p}_{00k}+\hat{p}_{01k}),\hat{\phi}_{2k}(1-\hat{p}_{00k}-\hat{p}_{11k}),\hat{\phi}_{3k}(\hat{p}_{00k}+\hat{p}_{01k}+\hat{\delta}_k),\hat{\phi}_{3k}(1-\hat{p}_{00k}-\hat{p}_{01k}-\hat{\delta}_k),\hat{\phi}_{4k})^{\top}$.

Step 3. Based on the bth bootstrap sample $(b=1,\ldots,\text{B})$, we calculate the observed values $t_{r1}^{(b)},\ldots,t_{rK}^{(b)}$ of statistics $T_{r1},\ldots,T_{rK}(r=l,s,w)$. Let $\omega_b=\max_{k=1,\ldots,K}|t_{rk}^{(b)}|$.

Step 4. Sort $\omega_1,\ldots,\omega_B$ to obtain the ordered values $\omega_{(1)}\le\omega_{(2)}\le\cdots\le\omega_{(B)}$ and compute the critical value $c_\alpha=\omega_{[B(1-\alpha)+1]}$, where $[\alpha]$ is the largest integer not greater than α.

Step 5. Reject global null hypothesis $H_0:\delta_1=\delta_2=\cdots=\delta_K=\tau_0$ if $\max_{k=1,\ldots,K}|t_{rk}|\ge c_\alpha$. In particular, one can reject the hypothesis $H_{k0}:\delta_k=\tau_0$ if $|t_{rk}|\ge c_\alpha$ for $k=1,\ldots,K$, where $r=l,s,w$.

3.2.3 Single-step adjusted MinP procedure

Similar to the singe-step adjusted MaxT procedure, in this section, we propose an algorithm based on the single-step adjusted MinP procedure as follows.

Step 1. Compute the observed values $t_{r1},\ldots,t_{rK}$ of test statistics $T_{r1},\ldots,T_{rK}(r=l,s,w)$ based on the original data.

Step 2. Given the MLEs $\hat{p}_{00k}$, $\hat{p}_{01k}$, $\hat{\delta}_k$, $\hat{\phi}_{1k}$, $\hat{\phi}_{2k}$, $\hat{\phi}_{3k}$, $\hat{\phi}_{4k}$ $(k=1,\ldots,K)$ for the original data, we generate B bootstrap samples

$$\{n_{00k}^{(b)},\ldots,n_{11k}^{(b)};m_{1xk}^{(b)},m_{0xk}^{(b)};m_{y1k}^{(b)},m_{y0k}^{(b)};m_{xyk}^{(b)};b=1,\ldots,B\}$$

$$\sim \text{Multinomial}_9(N_k,\hat{\boldsymbol{\phi}}),$$

where $\hat{\boldsymbol{\phi}}=(\hat{\phi}_{1k}\,\hat{p}_{00k},\hat{\phi}_{1k}\,\hat{p}_{01k},\hat{\phi}_{1k}\,(\hat{p}_{01k}+\hat{\delta}_k),\hat{\phi}_{1k}\,(1-\hat{p}_{00k}-2\hat{p}_{01k}-\hat{\delta}_k),\hat{\phi}_{2k}(\hat{p}_{00k}+\hat{p}_{01k}),\hat{\phi}_{2k}(1-\hat{p}_{00k}-\hat{p}_{11k}),\hat{\phi}_{3k}(\hat{p}_{00k}+\hat{p}_{01k}+\hat{\delta}_k),\hat{\phi}_{3k}(1-\hat{p}_{00k}-\hat{p}_{01k}-\hat{\delta}_k),\hat{\phi}_{4k})^{\top}$.

Step 3. Based on the bth bootstrap sample $(b=1,\ldots,\text{B})$, we calculate the observed values $t_{r1}^{(b)},\ldots,t_{rK}^{(b)}$ of statistics $T_{r1},\ldots,T_{rK}(r=l,s,w)$. Let $\omega_b=\max_{k=1,\ldots,K}|t_{rk}^{(b)}|$.

Step 4. The adjusted p-value is calculated by $\tilde{p}_{rk}=\frac{1}{B}\sum_{b=1}^{B}I(\omega_b\ge|t_{rk}|)(k=1,\ldots,K)$.

Step 5. Reject global null hypothesis $H_0:\delta_1=\delta_2=\cdots=\delta_K=\tau_0$ if $\tilde{p}=\min_{k=1,\ldots,K}\tilde{p}_{rk}\le\alpha$. In particular, one rejects the hypothesis $H_{k0}:\delta_k=\tau_0$ if $\tilde{p}_{rk}\le\alpha$, where $r=l,s,w$.

4. SIMULATION STUDIES

In this section, we investigate the finite sample performance of the proposed methods. Specifically, we examine the type I error control and power in a variety of settings via Monte Carlo simulations. First, we generate 10,000 samples $\{\mathbf{N}_k,\mathbf{m}_{xk},\mathbf{m}_{yk},m_{xyk}\}$ from the multinomial distribution $\text{Multinomial}_9(N_k;\phi_{1k}\,p_{00k},\phi_{1k}\,p_{01k},\phi_{1k}\,p_{10k},\phi_{1k}\,p_{11k},\phi_{2k}(p_{00k}+p_{01k}),\phi_{2k}(p_{10k}+p_{11k}),\phi_{3k}(p_{00k}+p_{10k}),\phi_{3k}(p_{01k}+p_{11k}),\phi_{4k})$ in stratum k $(k=1,\ldots,K)$ to calculate the empirical type I error rate and the empirical power. We generate 5,000 bootstrap samples. In this simulation, we consider $K=2$. Our interest is to test $\delta_1=\delta_2$, where $\delta_k=p_{10k}-p_{01k},k=1,2$ as in Table 2.

Denote marginal probabilities $p_{0+k}=p_{00k}+p_{01k},p_{1+k}=p_{10k}+p_{11k},p_{+0k}=p_{00k}+p_{10k}$, and $p_{+1k}=p_{01k}+p_{11k},k=1,2$. To induce dependence among entries in the contingency table, we define the correlation coefficient $\rho_k=(p_{00k}-p_{0+k}p_{+0k})/(p_{0+k}p_{1+k}p_{+0k}p_{+1k})^{1/2},k=1,2$ (Choi and Stablein, 1982). Parameter configurations are as follows. $(\phi_{11},\phi_{21},\phi_{31},\phi_{41})^{\top}=(0.7,0.1,0.1,0.1)^{\top}$, $(\phi_{12},\phi_{22},\phi_{32},\phi_{42})^{\top}=(0.5,0.1,0.1,0.3)^{\top}$, $p_{0+1}=0.4,0.6$, $p_{0+2}=0.3,0.5,0.7$, $\rho_1=\rho_2=0.1,0.3,0.5$, $N_1=N_2=30,50$ and 100 for the balanced design and $(N_1,N_2)=(30,50)$ and $(50,100)$ for the unbalanced design.

Empirical type I error rates are summarized in Table 4-5, 8-9, and empirical powers are summarized in Table 6-7, 10-11, where $\delta_1=0.1$ and $\delta_2=0.3$. Statistical significance level is set to be 0.05.

We summarize the main findings from the simulations as follows.

(i) The bootstrap re-sampling test procedure demonstrates robust behavior and outperforms the asymptotic test procedure (see, e.g., Table 4-5) in the sense that all the estimated type I error rates for the bootstrap re-sampling test procedure are close to the nominal level $\alpha=0.05$ under various settings.
(ii) The MaxT and MinP procedures usually outperform the Bonferroni procedure regardless of the test statistics.
(iii) The empirical power increases as ρ increases.
(iv) From Table 6-7 and Table 10-11, we can see that as the proportion of the missing data increases, the empirical power decreases.
(v) The Wald-type statistics on the basis of the Bonferroni procedure are liberal regardless of sample size.
(vi) The empirical powers for the Wald-type statistics with the Bonferroni procedure are greater than those with the single-step MaxT procedure and the single-step MinP procedure. Yet the Wald-type statistics have inflated type I error rates.
(vii) The score statistic is pretty robust in all settings. The type I error rates are close to the nominal level and the powers are reasonably high.

Table 4. *Empirical type I error rates for testing hypothesis $H_0: \delta_1 = \delta_2$ based on 10000 trials, $N_1 = N_2 = 30$, $K = 2$ at $\alpha = 5\%$*

$(\phi_1, \phi_2, \phi_3, \phi_4)$	p_{0+1}	p_{0+2}	ρ	Asymptotic test			Bootstrap resampling test		
				T_l	T_s	T_w	T_{bl}	T_{bs}	T_{bw}
(0.7,0.1,0.1,0.1)	0.4	0.3	0.1	4.40	3.36	4.56	5.30	4.80	5.30
			0.3	4.44	3.42	4.20	5.20	5.50	5.30
			0.5	2.78	2.30	3.08	5.70	5.20	5.40
		0.5	0.1	4.40	3.50	4.92	5.30	4.80	5.30
			0.3	4.42	2.84	3.92	5.20	5.50	5.30
			0.5	3.42	2.28	2.92	5.70	5.20	5.40
		0.7	0.1	4.68	3.60	4.72	5.30	4.80	5.30
			0.3	4.08	2.96	4.26	5.20	5.50	5.30
			0.5	3.64	2.12	2.88	5.70	5.20	5.40
	0.6	0.3	0.1	4.58	3.84	5.10	5.60	5.60	5.80
			0.3	4.66	3.74	5.18	5.90	5.10	5.40
			0.5	4.46	2.80	3.88	4.90	5.10	5.10
		0.5	0.1	5.30	3.62	4.84	5.60	5.60	5.80
			0.3	4.54	3.22	4.22	5.90	5.10	5.40
			0.5	3.72	2.54	3.70	4.90	5.10	5.10
		0.7	0.1	5.14	3.80	5.22	5.60	5.60	5.80
			0.3	4.30	3.16	4.62	5.90	5.10	5.40
			0.5	3.52	2.40	3.42	4.90	5.10	5.10
(0.5,0.1,0.1,0.3)	0.4	0.3	0.1	4.26	3.22	6.00	5.10	5.20	5.20
			0.3	3.34	2.36	4.86	5.00	5.40	4.90
			0.5	2.82	1.70	3.90	5.03	5.50	5.10
		0.5	0.1	3.88	3.38	6.00	5.10	5.20	5.20
			0.3	3.48	2.68	5.00	5.00	5.40	4.90
			0.5	2.66	1.34	3.52	5.03	5.50	5.10
		0.7	0.1	3.68	3.30	5.90	5.10	5.20	5.20
			0.3	3.54	2.86	5.50	5.00	5.40	4.90
			0.5	3.12	1.42	3.60	5.03	5.50	5.10
	0.6	0.3	0.1	4.52	3.62	6.88	4.60	5.33	5.43
			0.3	3.42	2.78	5.68	4.47	5.13	5.53
			0.5	3.04	1.82	3.80	4.47	5.27	5.30
		0.5	0.1	4.38	3.38	6.82	4.60	5.33	5.43
			0.3	4.08	2.92	5.48	4.47	5.13	5.53
			0.5	2.80	2.18	4.10	4.47	5.27	5.30
		0.7	0.1	4.44	3.34	6.60	4.60	5.33	5.43
			0.3	3.78	2.24	5.54	4.47	5.13	5.53
			0.5	2.96	2.00	4.26	4.47	5.27	5.30

5. A REAL DATA EXAMPLE

In this section, we shall revisit the multi-center study introduced in Section 1. In this dataset, patients are grouped by their effect, i.e., side effect and therapeutic effect group. From Table 1, we observe $n_{001} = 89$, $n_{011} = 13$, $n_{101} = 57$, $n_{111} = 65$, $m_{0x1} = 26$, $m_{1x1} = 49$, $m_{y01} = 2$, $m_{y11} = 0$, $m_{xy1} = 14$, $N_1 = 315$, $n_{002} = 11$, $n_{012} = 1$, $n_{102} = 124$,

Table 5. *Empirical type I error rates for testing hypothesis $H_0 : \delta_1 = \delta_2$ based on 10000 trials, $N_1 = N_2 = 50$, $K = 2$ at $\alpha = 5\%$*

$(\phi_1, \phi_2, \phi_3, \phi_4)$	p_{0+1}	p_{0+2}	ρ	Asymptotic test			Bootstrap resampling test		
				T_l	T_s	T_w	T_{bl}	T_{bs}	T_{bw}
(0.7,0.1,0.1,0.1)	0.4	0.3	0.1	4.46	3.70	4.66	5.30	5.70	5.00
			0.3	4.76	3.82	4.90	5.40	5.40	5.00
			0.5	4.26	3.66	4.16	5.60	4.30	4.60
		0.5	0.1	4.72	4.48	5.56	5.30	5.70	5.00
			0.3	3.88	4.32	5.22	5.40	5.40	5.00
			0.5	4.36	3.36	3.74	5.60	4.30	4.60
		0.7	0.1	4.44	3.64	4.52	5.30	5.20	5.00
			0.3	4.22	3.64	4.36	5.40	5.40	5.00
			0.5	4.10	3.26	3.78	5.60	4.30	4.60
	0.6	0.3	0.1	4.96	4.12	5.76	5.40	5.10	5.50
			0.3	4.84	3.96	5.06	4.10	5.10	4.90
			0.5	4.32	3.32	4.44	4.30	4.50	4.70
		0.5	0.1	4.46	4.18	5.14	5.40	5.10	5.50
			0.3	4.60	4.08	5.06	4.10	5.10	4.90
			0.5	4.28	3.26	3.76	4.30	4.50	4.70
		0.7	0.1	4.72	4.52	5.96	5.40	5.10	5.50
			0.3	5.06	4.14	5.08	4.10	5.10	4.90
			0.5	4.04	3.42	3.98	4.30	4.50	4.70
(0.5,0.1,0.1,0.3)	0.4	0.3	0.1	4.30	4.32	7.24	4.47	4.40	4.27
			0.3	4.74	4.12	6.78	4.63	4.50	4.27
			0.5	3.62	2.78	4.78	5.10	4.57	4.30
		0.5	0.1	4.76	4.04	6.54	4.47	4.40	4.27
			0.3	4.14	3.60	6.22	4.63	4.50	4.27
			0.5	3.78	2.82	5.04	5.10	4.57	4.30
		0.7	0.1	4.06	4.20	6.88	4.47	4.40	4.27
			0.3	3.58	3.68	6.64	4.63	4.50	4.27
			0.5	3.98	2.98	5.34	5.10	4.57	4.30
	0.6	0.3	0.1	4.30	3.68	7.22	5.10	5.40	5.60
			0.3	5.04	4.32	7.02	4.90	4.90	5.30
			0.5	3.40	2.70	5.08	4.23	4.50	4.80
		0.5	0.1	4.36	4.02	6.86	5.10	5.40	5.60
			0.3	4.30	3.72	6.86	4.90	4.90	5.30
			0.5	4.32	3.22	5.98	4.23	4.50	4.80
		0.7	0.1	5.22	4.08	7.38	5.10	5.40	5.60
			0.3	4.24	3.32	6.26	4.90	4.90	5.30
			0.5	3.70	3.00	6.00	4.23	4.50	4.80

$n_{112} = 88$, $m_{0x2} = 7$, $m_{1x2} = 68$, $m_{y02} = 0$, $m_{y12} = 2$, $m_{xy2} = 14$, $N_2 = 315$. We calculate $\hat{p}_{00k}, \hat{p}_{01k}, \hat{p}_{10k}$ and $\hat{p}_{11k}, k = 1, 2$ through the EM algorithm as illustrated in Section 2.2.1. We compute $\hat{\delta}_1 = \hat{p}_{101} - \hat{p}_{011} = 0.2147$ and $\hat{\delta}_2 = \hat{p}_{102} - \hat{p}_{012} = 0.5375$. To investigate if there is a statistical significant difference between the side effect group

Table 6. Empirical power for testing $H_0 : \delta_1 = \delta_2$, where $\delta_1 = 0.1, \delta_2 = 0.3, N_1 = N_2 = 30$ and $K = 2$

$(\phi_1, \phi_2, \phi_3, \phi_4)$	p_{0+1}	p_{0+2}	ρ	Asymptotic test			Bootstrap resampling test		
				T_l	T_s	T_w	T_{bl}	T_{bs}	T_{bw}
(0.7,0.1,0.1,0.1)	0.4	0.3	0.1	35.20	54.36	58.42	45.70	71.50	69.80
			0.3	34.24	54.04	57.66	47.40	75.00	74.10
			0.5	34.94	53.62	58.04	50.00	79.80	79.50
		0.5	0.1	25.52	37.56	44.76	27.70	43.30	44.20
			0.3	29.66	43.86	50.00	32.90	53.40	53.60
			0.5	35.68	52.88	58.04	45.20	68.00	67.70
		0.7	0.1	23.98	35.52	42.24	27.60	44.20	45.50
			0.3	27.82	40.64	47.70	34.90	51.10	52.20
			0.5	34.08	49.20	55.68	41.20	62.50	62.40
	0.6	0.3	0.1	33.52	53.32	57.40	45.70	71.50	69.80
			0.3	34.94	53.54	57.70	47.40	75.00	74.10
			0.5	34.02	52.70	56.82	50.00	79.80	79.50
		0.5	0.1	26.94	38.72	45.64	27.70	43.30	44.20
			0.3	31.38	44.88	50.98	32.90	53.40	53.60
			0.5	36.26	52.64	57.80	45.20	68.00	67.70
		0.7	0.1	23.88	34.58	41.76	27.60	44.20	45.50
			0.3	28.08	39.56	46.06	34.90	51.10	52.20
			0.5	33.94	48.80	54.72	41.20	62.50	62.40
(0.5,0.1,0.1,0.3)	0.4	0.3	0.1	17.70	33.48	48.02	29.00	53.50	54.80
			0.3	16.26	31.96	47.66	32.60	59.30	59.10
			0.5	16.98	33.48	48.00	37.20	64.40	64.80
		0.5	0.1	15.90	27.62	40.18	20.90	34.80	34.90
			0.3	17.56	29.62	42.94	23.20	42.60	41.90
			0.5	18.92	32.16	47.98	28.10	52.20	52.30
		0.7	0.1	16.76	26.44	39.38	19.40	32.60	33.70
			0.3	17.66	28.92	43.46	23.10	38.90	39.80
			0.5	18.82	31.02	46.56	28.90	49.50	49.30
	0.6	0.3	0.1	17.50	33.70	48.42	29.00	53.50	54.80
			0.3	16.50	32.24	46.64	32.60	59.30	59.10
			0.5	17.64	33.86	47.64	37.20	64.40	64.80
		0.5	0.1	16.72	26.94	39.96	20.90	34.80	34.90
			0.3	18.48	29.66	42.60	23.20	42.60	41.90
			0.5	17.24	30.68	45.56	28.10	52.20	52.30
		0.7	0.1	15.74	25.12	37.24	19.40	32.60	33.70
			0.3	18.24	29.22	42.52	23.10	38.90	39.80
			0.5	18.14	29.78	44.88	28.90	49.50	49.30

and therapeutic effect group, we test $H_0 : \delta_1 = \delta_2$ vs $H_1 : \delta_1 \neq \delta_2$. The proposed homogeneity testing procedures, i.e., asymptotic method and bootstrap re-sampling method are used. The results are summarized in Table 12. We observe that though the observed test statistics span a wide range, the resulting p-values are similar, providing overwhelming evidence to reject the null hypothesis that risk differences between the side effect group and therapeutic effect group are the same.

Although there are some differences between $\hat{\delta}_1$ and $\hat{\delta}_2$, they are very close to their mean 0.3761. To examine whether there is a substantial difference in proportion be-

Table 7. Empirical power for testing $H_0 : \delta_1 = \delta_2$, where $\delta_1 = 0.1, \delta_2 = 0.3, N_1 = N_2 = 50$ and $K = 2$

$(\phi_1, \phi_2, \phi_3, \phi_4)$	p_{0+1}	p_{0+2}	ρ	Asymptotic test			Bootstrap resampling test		
				T_l	T_s	T_w	T_{bl}	T_{bs}	T_{bw}
(0.7,0.1,0.1,0.1)	0.4	0.3	0.1	74.26	86.98	86.52	80.90	93.30	92.60
			0.3	72.50	86.44	86.06	82.50	94.00	93.60
			0.5	72.74	86.18	86.14	81.70	94.50	94.40
		0.5	0.1	44.96	57.68	63.32	44.00	60.60	60.90
			0.3	55.02	67.86	71.96	56.00	73.10	73.50
			0.5	72.84	84.54	84.78	76.10	88.70	86.70
		0.7	0.1	41.74	55.30	61.50	43.50	57.10	58.70
			0.3	51.62	64.00	69.20	54.80	67.20	68.30
			0.5	67.22	79.74	81.80	70.90	83.60	83.00
	0.6	0.3	0.1	73.40	86.32	86.16	80.90	93.30	92.60
			0.3	73.30	85.74	85.54	82.50	94.00	93.60
			0.5	72.98	85.88	85.64	81.70	94.50	94.40
		0.5	0.1	46.48	59.12	64.66	44.00	60.60	60.90
			0.3	55.36	68.26	72.30	56.00	73.10	73.50
			0.5	74.14	85.34	85.66	76.10	88.70	86.70
		0.7	0.1	42.92	55.24	61.46	43.50	57.10	58.70
			0.3	52.28	65.26	70.46	54.80	67.20	68.30
			0.5	67.68	79.68	81.48	70.90	83.60	83.00
(0.5,0.1,0.1,0.3)	0.4	0.3	0.1	50.02	68.10	77.38	61.50	81.30	81.80
			0.3	50.24	68.26	76.62	63.50	83.80	84.90
			0.5	50.72	69.82	78.54	66.50	87.50	88.60
		0.5	0.1	34.24	46.68	59.78	35.70	51.10	51.90
			0.3	40.18	54.34	66.66	43.60	59.80	60.80
			0.5	49.84	65.16	74.92	56.00	74.10	74.30
		0.7	0.1	32.60	44.18	57.14	32.20	49.30	49.60
			0.3	38.56	50.08	63.20	41.30	58.80	60.30
			0.5	47.80	60.90	72.60	51.80	70.60	71.70
	0.6	0.3	0.1	51.80	68.90	77.38	61.50	81.30	81.80
			0.3	51.52	69.56	77.58	63.50	83.80	84.90
			0.5	51.50	69.70	77.86	66.50	87.50	88.60
		0.5	0.1	32.52	46.26	58.38	35.70	51.10	51.90
			0.3	39.46	52.50	64.78	43.60	59.80	60.80
			0.5	49.56	65.58	75.84	56.00	74.10	74.30
		0.7	0.1	32.12	43.82	57.62	32.20	49.30	49.60
			0.3	38.40	51.18	63.54	41.30	58.80	60.30
			0.5	47.02	60.24	72.18	51.80	70.60	71.70

tween the first visit and the last visit, we consider testing

$$H_{k0} : \delta_k = 0.3761 \quad \text{vs} \quad H_{k1} : \delta_k \neq 0.3761 \quad \text{for } k = 1, 2.$$

Testing results are summarized in Table 13, where test statistics based on the likelihood ratio statistic, score statistic and Wald statistic are presented and p-values are recorded in parentheses. Again, at significance level 0.05, we have overwhelming evidence to reject $H_{k0} : \delta_k = 0.3761$ for $k = 1, 2$ and conclude that risk differences between the first visit and the last visit are not the same for the side effect group and therapeutic effect group.

Table 8. Empirical type I error for testing $H_0 : \delta_1 = \delta_2 = 0.1$, where $N_1 = 30, N_2 = 50$ and $K = 2$

$(\phi_1, \phi_2, \phi_3, \phi_4)$	p_{0+1}	p_{0+2}	ρ	Bonferroni			MaxT			MinP		
				T_{lk}	T_{sk}	T_{wk}	T_{lk}	T_{sk}	T_{wk}	T_{lk}	T_{sk}	T_{wk}
(0.7,0.1,0.1,0.1)	0.4	0.3	0.1	3.48	2.92	6.20	5.95	5.80	5.15	4.20	4.60	4.70
			0.3	3.26	2.76	6.22	5.85	5.80	5.95	4.60	4.80	4.80
			0.5	2.94	1.88	5.76	5.40	5.90	4.90	4.30	4.60	4.80
		0.5	0.1	4.24	3.96	7.28	5.25	5.40	5.60	5.60	5.60	4.40
			0.3	3.90	3.12	6.08	5.05	5.50	5.25	5.00	5.20	5.10
			0.5	3.18	2.16	5.92	5.35	5.80	5.35	5.00	5.60	5.20
		0.7	0.1	4.00	3.64	6.72	4.90	5.00	5.15	5.40	5.70	4.10
			0.3	4.20	3.48	6.98	5.75	5.85	5.20	5.10	5.30	5.20
			0.5	2.98	2.02	5.62	5.80	5.35	5.10	5.20	5.50	5.60
	0.6	0.3	0.1	4.32	3.56	6.60	5.80	5.40	5.40	5.10	5.90	5.60
			0.3	4.48	3.36	7.14	5.75	5.40	5.90	4.90	4.60	5.00
			0.5	3.00	1.90	6.12	5.85	5.45	6.00	5.90	5.50	4.40
		0.5	0.1	4.20	4.26	6.90	5.30	5.15	5.20	5.60	5.30	5.50
			0.3	4.14	3.42	7.18	5.40	4.90	5.10	5.20	4.50	4.90
			0.5	3.50	2.44	6.76	5.70	5.30	5.95	4.60	4.10	4.90
		0.7	0.1	3.86	3.64	6.64	4.90	4.90	5.10	5.20	5.40	5.40
			0.3	4.44	3.76	7.06	5.70	5.45	5.70	5.70	5.70	5.80
			0.5	3.72	2.72	6.34	5.10	5.00	4.95	5.40	5.30	5.70
(0.5,0.1,0.1,0.3)	0.4	0.3	0.1	3.56	3.04	9.92	5.15	4.60	5.25	5.00	4.40	5.30
			0.3	3.16	2.50	10.66	4.95	5.00	4.70	4.70	4.50	4.80
			0.5	2.22	1.44	9.82	4.95	4.95	4.95	4.80	5.00	5.00
		0.5	0.1	3.62	3.32	9.68	4.95	4.50	5.00	4.70	4.00	4.80
			0.3	2.98	2.70	10.10	4.25	4.00	4.15	4.40	4.50	4.40
			0.5	2.76	1.88	9.60	4.15	4.40	4.05	4.00	4.00	4.00
		0.7	0.1	3.20	2.92	10.06	5.05	4.45	4.70	5.40	4.70	5.10
			0.3	3.24	2.86	9.96	4.65	4.20	4.55	4.40	4.70	4.50
			0.5	2.34	1.68	9.48	4.10	4.40	4.20	4.00	4.00	4.70
	0.6	0.3	0.1	3.56	3.20	10.60	4.90	5.80	5.05	4.10	4.80	4.40
			0.3	3.28	2.56	10.04	4.85	5.90	4.80	4.80	4.90	4.30
			0.5	2.56	1.40	10.60	5.25	5.45	5.45	4.70	5.30	5.20
		0.5	0.1	3.56	3.34	10.96	5.00	5.05	5.15	4.10	4.00	4.40
			0.3	3.12	2.80	10.82	4.10	4.70	4.35	4.80	4.70	4.00
			0.5	2.68	2.02	10.10	4.85	4.95	4.75	4.10	4.30	4.90
		0.7	0.1	3.88	3.70	10.22	4.85	5.45	5.00	4.30	5.00	5.00
			0.3	3.38	3.06	10.40	4.70	5.25	4.70	4.80	4.70	4.90
			0.5	2.58	1.98	9.72	4.50	4.85	4.60	4.80	4.10	4.70

6. DISCUSSION

In this paper, we derive the joint distribution of the observed counts in an 2×2 contingency table with fixed total number of observations under missing at random assumption and propose new methods to test the equality of risk differences among multiple contingency tables. A post-hoc analysis is proposed to identify heterogeneous contingency tables. Numerical results support the proposed theory and the method is shown to be able to address practical problems of interest.

An important assumption is that data are missing at random. For non-ignorable missing data, we are not able to explicitly write down the joint distribution of the observed

Table 9. Empirical type I error for testing $H_0 : \delta_1 = \delta_2 = 0.1$, where $N_1 = 50, N_2 = 100$ and $K = 2$

$(\phi_1, \phi_2, \phi_3, \phi_4)$	p_{0+1}	p_{0+2}	ρ	Bonferroni			MaxT			MinP		
				T_{lk}	T_{sk}	T_{wk}	T_{lk}	T_{sk}	T_{wk}	T_{lk}	T_{sk}	T_{wk}
(0.7,0.1,0.1,0.1)	0.4	0.3	0.1	4.12	4.08	6.70	4.35	4.40	4.45	4.40	4.20	4.40
			0.3	4.56	3.96	6.82	4.30	4.90	4.00	4.60	5.10	4.10
			0.5	4.26	3.62	6.68	4.10	4.25	3.90	4.50	4.50	4.20
		0.5	0.1	4.30	4.08	6.48	4.70	4.90	4.90	4.70	5.40	5.10
			0.3	4.28	3.96	6.74	4.45	4.60	4.40	4.90	5.10	4.70
			0.5	4.56	3.60	6.72	4.65	4.55	4.70	4.00	4.00	4.10
		0.7	0.1	4.50	4.26	6.70	4.35	4.10	4.30	4.40	4.00	4.40
			0.3	4.12	3.80	6.54	4.35	4.50	4.25	5.20	5.10	4.90
			0.5	4.00	2.98	6.40	4.50	4.45	4.30	5.30	5.10	4.90
	0.6	0.3	0.1	4.32	3.98	6.66	4.60	4.50	4.55	4.50	4.70	4.60
			0.3	4.72	4.22	6.80	4.35	4.60	4.30	4.50	4.90	4.30
			0.5	3.74	3.24	5.96	4.40	4.95	4.20	4.40	5.10	4.00
		0.5	0.1	4.50	4.46	6.68	5.05	4.80	5.00	5.10	5.10	4.90
			0.3	4.16	3.82	6.34	4.45	4.35	4.45	4.60	4.80	4.50
			0.5	4.10	3.72	6.18	3.80	4.05	3.90	4.50	4.30	3.90
		0.7	0.1	4.26	4.22	6.60	4.50	4.30	4.35	4.90	4.40	4.40
			0.3	4.74	4.40	6.78	4.60	4.45	4.55	5.20	5.10	5.10
			0.5	4.16	3.48	6.44	5.15	4.95	4.90	5.70	5.50	5.30
(0.5,0.1,0.1,0.3)	0.4	0.3	0.1	4.50	4.00	10.88	5.60	5.65	5.55	4.90	4.50	4.80
			0.3	3.46	3.32	10.54	5.45	5.35	5.45	4.00	4.10	4.70
			0.5	3.98	3.14	10.90	5.25	5.35	5.25	5.20	5.60	5.40
		0.5	0.1	4.38	4.28	11.60	5.45	5.15	5.10	5.70	5.30	5.70
			0.3	3.78	3.40	10.92	5.45	5.15	5.10	5.00	4.90	4.30
			0.5	3.76	3.10	10.54	5.50	5.30	5.95	5.50	5.70	5.20
		0.7	0.1	4.40	4.32	10.86	5.30	5.15	5.15	5.00	5.70	5.50
			0.3	4.66	4.10	11.74	4.70	4.65	4.75	4.90	4.50	4.10
			0.5	3.76	2.68	10.18	5.00	5.00	5.20	5.70	5.30	5.10
	0.6	0.3	0.1	4.52	4.12	11.38	5.10	5.25	5.30	4.80	4.20	4.10
			0.3	4.74	4.32	11.68	5.30	5.10	5.45	4.90	4.20	4.90
			0.5	3.58	2.88	10.28	4.80	4.80	4.90	4.20	4.40	4.10
		0.5	0.1	4.28	4.68	11.46	4.80	4.45	4.60	4.80	4.50	4.40
			0.3	4.16	3.74	11.02	5.40	4.85	5.05	5.00	4.70	4.20
			0.5	4.04	3.10	10.76	5.25	4.85	5.80	5.00	4.80	5.20
		0.7	0.1	4.26	4.38	11.20	4.60	4.50	4.50	4.30	4.80	4.30
			0.3	4.52	4.54	11.68	4.70	4.40	4.55	4.00	4.00	4.90
			0.5	3.44	2.84	9.96	4.90	4.70	5.15	5.10	5.40	5.30

counts and consequently the asymptotic results would not be valid. However, we conjecture that bootstrap resampling method to calculate p-values remain valid for the global test $H_0 : \delta_1 = \delta_2 = \ldots = \delta_K$ and multiple comparison $H_{0k} : \delta_k = \tau_0, k = 1, \ldots, K$. In practice, it is recommended to use the bootstrap resampling method when sample size is small. In addition, bootstrap resampling method also provides robust inference against various modeling assumptions, including MAR.

Molenberghs *et al.* (1999) provided examples, in the contingency table setting, where different non-ignorable missing models that produce the same fit to the observed data,

Table 10. *Empirical power for testing* $H_0 : \delta_1 = \delta_2 = 0.1$ *vs* $H_1 : \delta_1 = 0.1, \delta_2 = 0.3,$ *where* $N_1 = 30, N_2 = 50$ *and* $K = 2$

$(\phi_1, \phi_2, \phi_3, \phi_4)$	p_{0+1}	p_{0+2}	ρ	Bonferroni			MaxT			MinP		
				T_{lk}	T_{sk}	T_{wk}	T_{lk}	T_{sk}	T_{wk}	T_{lk}	T_{sk}	T_{wk}
(0.7,0.1,0.1,0.1)	0.4	0.3	0.1	60.08	58.92	65.14	67.15	71.05	60.85	66.30	70.80	61.00
			0.3	60.94	59.92	66.00	68.75	74.05	60.40	67.50	73.20	60.10
			0.5	58.12	57.12	64.08	72.50	79.35	60.50	72.80	78.90	62.00
		0.5	0.1	35.52	33.68	44.42	38.35	38.35	38.70	39.00	39.30	39.20
			0.3	43.08	41.06	51.08	50.00	49.50	47.80	50.80	50.30	48.80
			0.5	59.52	56.58	64.08	66.65	69.40	58.35	65.20	67.20	57.60
		0.7	0.1	32.50	30.66	41.44	36.80	36.05	37.45	36.90	37.10	37.90
			0.3	40.18	37.52	48.22	44.55	44.45	43.65	44.20	44.90	43.60
			0.5	54.60	50.24	59.18	61.95	63.15	55.00	61.60	63.50	54.40
	0.6	0.3	0.1	58.86	58.04	63.88	67.10	70.65	60.20	65.70	69.40	59.40
			0.3	60.40	59.30	65.50	68.80	72.65	60.45	67.50	71.30	59.80
			0.5	60.38	58.86	65.44	72.60	78.50	60.40	73.30	78.70	62.10
		0.5	0.1	35.38	33.72	43.44	38.25	38.00	38.40	39.60	38.80	39.30
			0.3	44.54	42.52	51.52	49.80	48.85	48.00	50.80	49.40	48.70
			0.5	59.34	56.08	63.34	66.40	68.55	58.90	64.70	66.50	57.60
		0.7	0.1	32.28	30.44	40.82	36.95	35.75	37.50	37.50	36.50	38.00
			0.3	39.86	37.30	48.82	44.15	43.25	43.75	44.40	43.90	44.20
			0.5	55.22	51.06	60.66	61.60	62.35	55.85	61.10	62.30	55.30
(0.5,0.1,0.1,0.3)	0.4	0.3	0.1	35.52	34.08	56.48	45.83	49.20	40.97	46.70	50.10	42.00
			0.3	35.38	33.78	56.86	48.60	54.40	41.47	49.00	53.80	42.20
			0.5	34.42	32.54	55.86	51.87	60.37	39.10	53.00	61.00	40.10
		0.5	0.1	22.96	21.16	40.96	29.27	29.37	29.10	30.50	29.50	30.20
			0.3	28.86	25.88	46.98	35.87	36.63	34.80	34.70	36.20	34.40
			0.5	34.42	29.88	55.06	46.97	49.40	39.50	45.80	47.90	39.10
		0.7	0.1	22.68	20.96	40.60	26.20	25.30	26.57	26.70	25.00	27.50
			0.3	27.08	23.94	45.40	32.07	32.03	30.90	32.50	32.40	31.00
			0.5	33.16	28.32	52.20	43.90	44.47	38.57	45.20	45.60	39.90
	0.6	0.3	0.1	35.00	33.06	55.16	45.73	48.67	40.67	46.90	49.00	41.60
			0.3	34.90	33.04	55.52	48.37	53.00	41.27	48.80	53.50	41.80
			0.5	34.40	32.26	55.50	51.17	59.17	38.80	51.90	60.00	39.40
		0.5	0.1	24.88	23.82	42.90	29.00	28.57	29.33	29.30	28.80	29.20
			0.3	29.32	26.64	48.92	35.37	35.97	34.27	34.00	35.60	32.50
			0.5	35.08	30.14	54.80	46.07	48.63	39.23	45.10	47.90	38.80
		0.7	0.1	22.94	21.04	40.24	26.27	25.43	26.37	26.10	24.90	27.30
			0.3	27.52	24.56	45.36	31.73	31.63	30.93	31.50	32.50	31.70
			0.5	32.92	27.80	51.28	43.50	43.13	37.93	45.00	44.80	39.20

are different in their prediction of the unobserved counts. This implies that such models cannot be examined using data alone. Indeed, even if two models fit the observed data equally well, one still needs to reflect on the plausibility of the assumptions made as discussed in Molenberghs *et al.* (1999). In this scenario, *prior* knowledge about some of the parameters should be incorporated into data analysis. It is desirable to consider a range of plausible models in the sensitivity analysis. Such an analysis might show that some parameter estimates are very variable and no precise conclusions can be reached from the range of models considered, whereas other parameter estimates may be shown to be fairly stable. This certainly warrants additional development but is beyond the scope of this paper.

Table 11. Empirical power for testing $H_0 : \delta_1 = \delta_2 = 0.1$ vs $H_1 : \delta_1 = 0.1, \delta_2 = 0.3$, where$N_1 = 50, N_2 = 100$ and $K = 2$

$(\phi_1, \phi_2, \phi_3, \phi_4)$	p_{0+1}	p_{0+2}	ρ	Bonferroni			MaxT			MinP		
				T_{lk}	T_{sk}	T_{wk}	T_{lk}	T_{sk}	T_{wk}	T_{lk}	T_{sk}	T_{wk}
(0.7,0.1,0.1,0.1)	0.4	0.3	0.1	96.98	97.48	96.72	97.15	97.95	95.10	96.80	97.60	95.20
			0.3	97.62	97.96	97.64	97.40	98.25	94.85	97.00	98.00	94.90
			0.5	97.76	98.18	97.62	98.85	99.15	95.40	98.90	99.10	95.00
		0.5	0.1	70.12	69.36	76.18	72.45	72.40	72.55	71.50	72.10	71.30
			0.3	81.94	81.50	85.32	84.20	85.05	83.45	84.20	85.40	83.00
			0.5	96.52	96.72	96.06	96.90	97.50	95.10	96.20	97.20	94.10
		0.7	0.1	65.14	64.12	71.40	66.85	66.05	67.15	68.00	67.10	68.70
			0.3	76.26	75.08	81.28	79.20	79.20	78.85	79.90	79.20	79.30
			0.5	93.38	92.56	94.24	93.55	94.10	91.25	93.20	94.00	91.00
	0.6	0.3	0.1	97.16	97.80	97.02	97.30	97.75	95.00	97.00	97.50	95.20
			0.3	97.28	97.78	97.26	97.40	98.15	95.00	97.20	97.80	95.10
			0.5	96.96	97.34	96.88	98.70	99.05	95.35	99.00	99.00	95.10
		0.5	0.1	69.92	69.00	76.04	72.70	71.80	72.50	71.60	71.10	71.40
			0.3	82.26	81.78	86.26	84.35	85.10	83.80	83.80	84.80	83.50
			0.5	97.06	97.24	96.64	96.95	97.40	95.15	96.10	97.10	94.00
		0.7	0.1	66.50	65.74	72.84	67.25	65.75	67.15	68.80	66.60	68.70
			0.3	76.84	75.36	81.62	79.30	79.10	78.85	80.10	79.30	79.50
			0.5	93.14	92.66	93.68	93.60	94.55	91.65	93.60	94.90	91.70
(0.5,0.1,0.1,0.3)	0.4	0.3	0.1	87.50	88.44	93.60	90.05	91.95	85.85	90.70	92.70	87.30
			0.3	88.28	88.64	93.56	90.45	92.75	85.15	91.60	93.40	86.80
			0.5	88.26	88.68	93.92	92.25	95.00	84.10	92.00	94.80	83.80
		0.5	0.1	54.10	52.64	70.58	57.80	57.35	57.30	57.10	56.00	56.70
			0.3	66.42	64.94	80.42	68.85	68.85	67.55	68.10	67.80	66.40
			0.5	85.50	85.04	91.46	88.05	89.15	83.40	87.90	88.90	83.20
		0.7	0.1	49.38	48.14	66.46	52.50	50.35	52.45	52.20	50.50	53.00
			0.3	60.62	57.80	76.44	63.55	63.00	62.55	63.60	62.20	62.50
			0.5	79.96	77.52	88.34	82.45	83.10	78.55	82.30	83.20	77.90
	0.6	0.3	0.1	88.30	88.52	93.86	89.90	91.05	85.00	90.80	92.00	86.60
			0.3	87.94	88.46	93.86	90.15	92.25	84.60	91.30	93.00	86.50
			0.5	87.58	87.92	92.82	92.20	94.70	84.30	92.00	94.10	83.70
		0.5	0.1	54.60	53.02	70.90	57.65	56.80	56.85	56.50	54.90	55.70
			0.3	64.70	63.92	79.00	69.20	68.80	67.65	68.40	68.00	66.70
			0.5	85.62	85.14	91.44	87.85	89.15	83.45	87.90	88.90	83.60
		0.7	0.1	50.38	48.92	67.32	51.45	49.65	51.55	51.20	50.20	51.50
			0.3	61.12	59.40	75.92	63.00	62.45	62.15	62.70	61.90	61.90
			0.5	80.26	77.56	88.34	82.55	83.00	79.15	82.40	83.20	78.20

Table 12. Three test statistics and corresponding bootstrap p-values for testing H_0: $\delta_1 = \delta_2$ vs H_1: $\delta_1 \neq \delta_2$

Test statistic	Test statistic value	Bootstrap p-value
Likelihood ratio test	90.0464	< 0.0000001
Score test	115.6219	< 0.0000001
Wald test	51.1145	< 0.0000001

Table 13. Three test statistics and corresponding bootstrap p-values

	T_{lk}	T_{sk}	T_{wk}
$k = 1$	22.7484	-4.1134	-4.9511
	(< 0.000001)	(< 0.000001)	(< 0.000001)
$k = 2$	20.6479	4.3000	5.1826
	(< 0.000001)	(< 0.000001)	(< 0.000001)

ACKNOWLEDGEMENT

The research of Li, H. is partially supported by a grant from the Natural Science Foundation of China (11561075) and a grant from the Science Foundation of Yunnan Province (2016FB005).

SUPPLEMENTARY MATERIAL

Supplementary materials (http://intlpress.com/site/pub/pages/journals/items/sii/content/vols/0011/0002/s002) available include additional simulation results.

Received 1 February 2017

REFERENCES

Chang, M. (2009). Estimation of multiple response rates in phase II clinical trials with missing observations. *Journal of Biopharmaceutical Statistics* **19**, 791–802. MR2751088

Choi, S.C. and Stablein, D.M. (1982). Practical tests for comparing two proportions with incomplete data. *Applied Statistics* **31**, 256–262.

Choi, S. C., Stablein, D. M. (1988). Comparing incomplete paired binomial data under non-random mechanisms. *Statistics in Medicine* **7**, 929–939.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38. MR0501537

Kadane, J.B. (1985). Is victimization chronic? A Bayesian analysis of multinomial missing data. *Journal of Econometrics* **29**, 47–67. MR0801653

Kenward, M.G., Goethgebeur, E.J.T. and Molenberghs, G. (2001). Sensitivity analysis for incomplete categorical data. *Statistical Modelling* **1**, 31–48.

Kenward, M.G., Lesaffre, E. and Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50**, 945–953.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data, 2nd Edition. *New York: John Wiley & Sons.* MR1925014

Michiels, B. and Molenberghs, G. (1997). Protective estimation of longitudinal categorical data with nonrandom dropout. *Communications in Statistics — Theory and Methods* **26**, 65–94.

Miller, K.M. and Looney, S.W. (2012). A simple method for estimating the odds ratio in matched case-control studies with incomplete paired data. *Statistics in Medicine* **31**, 3299–3312. MR3041811

Molenberghs, G., Kenward, M.G. and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika* **84**, 33–44.

Molenberghs, G., Goetghebeur, E., Lipsitz, S. R. and Kenward, M. G. (1999). Non-random missinngness in categorical data: strengths and limitations. *The American Statistician* **53**, 110–118.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using an n-way Plackett distribution. *Journal of the American Statistical Association* **89**, 633–644.

Rubin, D.R. (1976). Inference and missing data. *Biometrika* **63**, 581–592. MR0455196

Tang, M.L., Ling, M.H. and Tian, G.L. (2009). Exact and approximate unconditional confidence intervals for proportion difference in the presence of incomplete data. *Statistics in Medicine* **28**, 625–641. MR2655734

Tang, M.L., Ng, K.W., Tian, G.L. and Tan, M. (2007). On improved EM algorithm and confidence interval construction for incomplete $r \times c$ tables. *Computational Statistics & Data Analysis* **51**, 2919–2933. MR2345614

Tang, N.S., Li, H.Q., Tang, M.L. and Li, J. (2016). Confidence interval construction for the difference between two correlated proportions with missing observations. *Journal of Biopharmaceutical Statistics* **26**, 323–338.

Tian, G.L. and Li, H.Q. (2015). A new framework of statistical inferences based on the valid joint sampling distribution of observed counts in an incomplete contingency table. *Statistical Methods in Medical Research*, in press. MR3687174

Tian, G.L., Ng, K.W. and Geng Z. (2003). Bayesian computation for contingency tables with incomplete cell-counts. *Statistica Sinica* **13**, 189–206. MR1963928

Huiqiong Li
Department of Statistics
Yunnan University
Kunming 650091
P. R. China
E-mail address: lihuiqiong@ynu.edu.cn

Niansheng Tang
Department of Statistics
Yunnan University
Kunming 650091
P. R. China
E-mail address: nstang@ynu.edu.cn

Guoliang Tian
Department of Mathematics
Southern University of Science and Technology
Shenzhen 518055
P. R. China
E-mail address: tiangl@sustc.edu.cn

Hongyuan Cao
Department of Statistics
University of Missouri-Columbia
Columbia, MO 65201
U.S.A.
E-mail address: caohong@missouri.edu