Asanao Shimokawa[1] / **Etsuo Miyaoka**[2]

On Stratified Adjusted Tests by Binomial Trials

[1] Department of Mathematics, Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan, E-mail: shimokawa@rs.tus.ac.jp
[2] Department of Mathematics, Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku , Tokyo 162-8601, Japan

Abstract:
To estimate or test the treatment effect in randomized clinical trials, it is important to adjust for the potential influence of covariates that are likely to affect the association between the treatment or control group and the response. If these covariates are known at the start of the trial, random assignment of the treatment within each stratum would be considered. On the other hand, if these covariates are not clear at the start of the trial, or if it is difficult to allocate the treatment within each stratum, completely randomized assignment of the treatment would be performed. In both sampling structures, the use of a stratified adjusted test is a useful way to evaluate the significance of the overall treatment effect by reducing the variance and/or bias of the result. If the trial has a binary endpoint, the Cochran and Mantel-Haenszel tests are generally used. These tests are constructed based on the assumption that the number of patients within a stratum is fixed. However, in practice, the stratum sizes are not fixed at the start of the trial in many situations, and are instead allowed to vary. Therefore, there is a risk that using these tests under such situations would result in an error in the estimated variation of the test statistics. To handle the problem, we propose new test statistics under both sampling structures based on multinomial distributions. Our proposed approach is based on the Cochran test, and the difference between the two tests tends to have similar values in the case of a large number of patients. When the total number of patients is small, our approach yields a more conservative result. Through simulation studies, we show that the new approach could correctly maintain the type I error better than the traditional approach.

Keywords: binary data, random stratum sizes, risk difference, type I error
DOI: 10.1515/ijb-2016-0047

1 Introduction

In several medical studies conducted to test a treatment effect, such as a new drug or operative method, the response is frequently given as a binary endpoint such as "success" vs. "failure". The two groups compared are frequently referred to as the "treatment" and "control" groups. Then, the results of the study are summarized in a 2×2 contingency table that is dichotomous by treatment and result.

In practice, there is also a covariate that will likely affect the association between the treatment or control group and the response. When there is an imbalance of the covariate between the two groups, evaluating the overall treatment effect while ignoring this imbalance would lead to a biased result. This issue is equally relevant when considering either a multicenter clinical trial or a meta-analysis. That is, there is a possibility that the potential background factors of patient, treatment environment, or medical skills differ between each center or trial. To adjust for such imbalances, stratified analyses or analyses based on an appropriate regression model such as logistic regression are widely used [1, 2]. In this study, we focus on the use of a stratified analysis.

The most commonly used measures of a treatment effect are relative risk, odds ratio, and risk difference. The advantages and disadvantages of these measures are described in Lachin [1] and Fleiss et al. [3]. Although the following discussion may also be relevant when considering any criteria, we specifically consider the case based on risk difference for the sake of simplicity and to allow for natural interpretation. If the true success probabilities of the treatment and control groups in stratum i are represented as p_{Ti} and p_{Ci}, respectively, then the risk difference in the stratum is defined as $\delta_i = p_{Ti} - p_{Ci}$. Thus, the true overall treatment effect is expressed as $\delta = \sum_i \pi_i \delta_i$, where π_i is the true ratio of patients that would have entered stratum i in the population [4].

To examine the statistical significance of the true overall treatment effect under the case in which the strata are imbalanced, several approximation test statistics of the hypothesis $H_0 : \delta = 0$ are proposed. As a simple approach based on the weight of the reciprocal of the squared standard error of δ_i, the weighted mean divided by its standard error can be used. This statistic approximately follows the chi-square distribution with 1 degree of freedom under H_0 [3]. However, it is also known that the performance of this approach is not good in

Asanao Shimokawa is the corresponding author.

many situations. The most commonly used approaches are the Cochran test [5] and Mantel-Haenszel test [6]. There are several ways to derive these test statistics; for example, these can be viewed as summarized statistics based on risk differences weighted by the harmonic mean of the number of patients in each stratum [7]. From another perspective, the Cochran test can be viewed as a test based on two independent groups according to the binomial distribution in each stratum [1]. The Mantel-Haenszel test can be considered to be based on the hypergeometric distribution [1]. The results of these tests are often very similar, and quickly converge to the same value. Yusuf et al. [8] proposed a Mantel-Haenszel test with the continuity correction term removed. A comparison of these methods based on simulation studies is described in Sánchez-Meca and Mar\'in-Martńez [7]. Mehrotra and Railkar [4] proposed another test method that could minimize the squared error between the true overall treatment effect and the estimated treatment effect. A detailed discussion of this approach to a treatment by stratum interaction is provided by Mehrotra [9].

As a more general framework, many estimators that adjust for covariates in randomized trials (to estimate the overall treatment effect) are proposed. The comparative research on the estimators that are focused on the binary outcome cases was published by Colantuoni and Rosenblum [10]. In their research, the seven estimators are studied in terms of the properties and performance based on simulations. Although these estimators can be applied to the case of the randomized stratification, as discussed in their paper, we consider simpler cases in this study. That is, we consider the case where the independence of the outcome and covariates (that are not used in the stratification) is satisfied.

There are two types of sampling structures for a stratified analysis in a randomized clinical trial: random assignment of a treatment within each stratum, and completely randomized assignment of the treatment. If the covariates that are likely to affect the association between the treatment or control group and the response are known at the start of the trial, random assignment of the treatment within each stratum would be considered. On the other hand, if these covariates are not clear at the start of the trial, or if it is difficult to allocate the treatment within each stratum, completely randomized assignment of the treatment would be performed. Although the debate as to which structure is preferable continues, in both sampling structures, the stratified adjusted tests described above are generally used to assess the significance of the overall treatment effect. Following the terminology employed by Ganju and Zhou [11], we refer to the random assignment of a treatment within each stratum as pre-stratification, and the completely randomized assignment is referred to as post-stratification.

Although the stratified adjusted tests such as the Cochran test show good performance for assessing the significance of an overall treatment effect in a stratified analysis, these tests are constructed based on the assumption that the number of patients within a stratum is fixed. However, in practice, the stratum sizes are not fixed at the start of the trial in many situations, and are instead allowed to vary. For example, in a multicenter trial using a pre-stratification design, the total sample size might be fixed but it is nevertheless difficult to control the sample size in each stratum in many cases [12]. In a randomized clinical trial using the post-stratification design, it would be impossible to fix the number of patients in each stratum (such as sex and age) before the start of the clinical trial. Therefore, there is a risk that using such tests under these situations will lead to potentially ignoring the variance in the number of patients in each stratum. Consequently, the variation in the test statistics is considered underestimated in many situations.

To handle this problem, we propose new test statistics applicable for the pre-stratification and post-stratification designs. In our approach, the number of patients in the strata are assumed to follow a multinomial distribution. This assumption is reasonable for many randomized clinical trials. In fact, [12] proposed a test statistic under this same assumption for evaluating the significance of the overall treatment effect using continuous data in multicenter trials with post-stratification. With this assumption, the number of success patients in each stratum is viewed as a result of a two-step process. That is, the number of patients in a stratum is given by following the multinomial distribution in the first stage, and then the number of success patients for the treatment in each stratum is given by following the binomial distribution. From another perspective, the number of success patients in each treatment can also be considered to be given by a finite mixture model.

Since we assume that the two independent groups have the binomial distributions in each stratum, our proposed approach is based on the Cochran test. The difference between the proposed test and ordinary Cochran test is given by the difference in the estimator of the variance of the overall treatment effect. When the total number of patients is small, this variance estimator of our proposed approach tends to become larger than that of the Cochran test because the variability of the estimator of the treatment effect in each stratum becomes large or the number of patients included in the two groups tends to become imbalanced in each stratum. This variability or imbalance is reduced by increasing the total number of patients, and as a result, our proposed approach and the Cochran test yield similar values.

In our proposed approach, the variation of the estimated treatment effect is expected to be higher than that obtained with the traditional approach in many situations. This fact seems to be similar to the approach considering a random-effects model for combining the evidence of several studies to compare two treatments in a meta-analysis [13, 14]. Although the results of these two approaches show trends in the same direction, their

basic underlying ideas and assumptions are distinct. In the random-effects approach, the model assumes random variation among the strata with respect to the expected treatment effect, and thus the estimated variation of the statistics is higher than that obtained with a fixed-effects approach. In our approach, on the other hand, the expected treatment effect is fixed in each stratum. The change in the variation of the estimated treatment effect is instead due to the variation in the number of patients in each stratum.

The remainder of this paper is organized as follows. In Section 2, we introduce the notation and traditional stratified adjusted test used to assess the significance of an overall treatment effect. In Section 3 and Section 4, the new approaches considering the variation in the number of patients in the strata are described for the pre-stratification and post-stratification contexts. In Section 5, the results of simulation studies to assess the type I error for each approach are described. The results of applying the proposed method to data from previous randomized clinical trials are described in Section 6. Finally, concluding remarks are given in Section 7.

2 Notation and traditional approaches

2.1 Notation

Let π_i be the true ratio of patients that would have entered stratum i in the target population, where $i = 1, 2, \cdots, K$, with the number of strata K fixed. If the true success probabilities of the treatment and control groups in stratum i are represented as p_{Ti} and p_{Ci}, respectively, the risk difference in i is defined as $\delta_i = p_{Ti} - p_{Ci}$. As used in the general medical research (e.g. Ganju and Mehrotra [12]), the true overall treatment effect is defined as follows:

$$\delta = \sum_{i=1}^{K} \pi_i \delta_i.$$

Now, our goal is to assess the significance of δ by testing the null hypothesis $H_0 : \delta = 0$.

Let n_{Ti} and n_{Ci} be the sample sizes for the treatment and control groups, respectively. In the traditional approaches, n_{Ti} and n_{Ci} are fixed. In our approach, on the other hand, n_{Ti} and n_{Ci} are treated as random variables. The total number of patients in stratum i is represented as $n_i = n_{Ti} + n_{Ci}$. The number of success patients in the treatment and control groups for stratum i is represented as a_i and b_i, respectively. If n_{Ti} is fixed, a_i follows a binomial distribution with n_{Ti} and p_{Ti}. Similarly, if n_{Ci} is fixed, b_i follows a binomial distribution with n_{Ci} and p_{Ci}. The total number of patients in the treatment and control groups is denoted by $N_T = \sum_i n_{Ti}$ and $N_C = \sum_i n_{Ci}$, respectively, and $N = N_T + N_C$ represents the total number of patients included in the trial.

Depending on the context, π_i can be either known or unknown. For example, if the target population comprises patients with a disease that has been previously well-studied on a large scale, the composition ratio of such patients such as age and sex can be used for π_i. On the other hand, if there is no available information about the target disease, then π_i needs to be estimated by

$$\hat{\pi}_i = \frac{n_{Ti} + n_{Ci}}{N}. \tag{1}$$

2.2 The test for assessing the overall treatment effect

As described in the Introduction, several asymptotic tests for assessing the overall treatment effect can be considered. The typical and simple methods are Cochran are Mantel-Haenszel tests. Because the difference between these tests is given by the difference between the sample size of 1 in denominators of the test statistics [1], these statistics take similar values. The difference between these tests can be viewed as the difference in the model assumptions. That is, the Cochran and Mantel-Haenszel tests are based on the assumption that the patients included in a stratum follow two independent binomial distributions or a hypergeometric distribution, respectively. Because in this study, we assume that the two independent groups have the binomial distributions in each stratum, we consider the following argument based on the Cochran test. The argument based on the other complicated methods, e.g., including continuous correction or covariate adjustments described by Colantuoni and Rosenblum [10] would be possible. These extensions are further works.

The Cochran test can be viewed as summarized statistics based on risk differences weighted by the harmonic mean of the number of patients in each stratum. The individual risk differences of the strata are estimated by

$\hat{\delta}_i = \hat{p}_{Ti} - \hat{p}_{Ci}$, where $\hat{p}_{Ti} = a_i/n_{Ti}$ and $\hat{p}_{Ci} = b_i/n_{Ci}$ are the observed success probabilities in the treatment and control groups, respectively. Then, the test statistic can be calculated as follows:

$$\chi^2_C = \frac{\hat{\delta}^2}{\widehat{Var}_C(\hat{\delta})}, \tag{2}$$

where $\hat{\delta} = \sum_i w_i \hat{\delta}_i$ represents the estimator of the overall treatment effect. w_i represents the harmonic mean of the number of patients in i:

$$w_i = \frac{n_{Ti} n_{Ci}}{n_{Ti} + n_{Ci}}. \tag{3}$$

The variance of $\hat{\delta}$ is estimated as

$$\widehat{Var}_C(\hat{\delta}) = \sum_i w_i \hat{p}_i (1 - \hat{p}_i), \tag{4}$$

where $\hat{p}_i$ is the success probability of two treatments under the null hypothesis in i:

$$\hat{p}_i = \frac{a_i + b_i}{n_i}. \tag{5}$$

From the independence of strata and Slutsky's theorem, χ^2_C is asymptotically distributed following the chi square distribution with 1 degree of freedom under H_0.

3 Proposed test under pre-stratification

In this section and Section 4, we introduce a new statistic for testing $H_0 : \delta = 0$ based on eq. (2) under the assumption that the number of patients in each stratum follows multinomial distributions:

$$(n_{T1}, n_{T2}, \cdots, n_{TK}) \sim Multinomial\left(N_T, (\pi_1, \pi_2, \cdots, \pi_K)\right),$$
$$(n_{C1}, n_{C2}, \cdots, n_{CK}) \sim Multinomial\left(N_C, (\pi_1, \pi_2, \cdots, \pi_K)\right).$$

In a pre-stratification analysis, the patients are randomized within each stratum. Therefore, we can assume that the ratio of the number of patients in the two groups is exactly $n_{Ti}/n_{Ci} = k_T/k_C$ in arbitrary stratum i, where k_T and k_C represent positive integers. That is, n_{Ci} can be represented as

$$n_{Ci} = \frac{k_C}{k_T} n_{Ti} \tag{6}$$

for $i = 1, 2, \cdots, K$.

The variance of the estimator of the overall treatment effect $\hat{\delta}$ is represented as

$$Var(\hat{\delta}) = E\left[Var\left(\sum_i w_i \hat{\delta}_i \middle| n_{Ti}, n_{Ci}\right)\right] + Var\left[E\left(\sum_i w_i \hat{\delta}_i \middle| n_{Ti}, n_{Ci}\right)\right]. \tag{7}$$

By tedious calculation, the exact formulation of this variance is derived as

$$Var_{pre}(\hat{\delta}) = N_T \left(\frac{k_C}{k_T + k_C}\right)^2 \left[\frac{1}{k_C} \sum_i \pi_i \{k_C p_{Ti}(1 - p_{Ti}) + k_T p_{Ci}(1 - p_{Ci})\} + \sum_i (\delta_i - \delta)^2\right]. \tag{8}$$

The detailed derivation of eq. (8) is given in Appendix A. With replacement of the unknown parameters p_{Ti} and p_{Ci} by $\hat{p}_i$ in eq. (5), replacement of δ_i and δ by the estimator, and under the null hypothesis, a new test statistic is proposed as follows:

$$\chi^2_{pre} = \frac{k_C \left(\sum_i n_{Ti}\hat{\delta}_i\right)^2}{N_T(k_T + k_C)\left[\sum_i \pi_i \hat{p}(1-\hat{p}) + \sum_i (\hat{\delta}_i - \hat{\delta})^2\right]}. \tag{9}$$

When π_i is unknown, its value is replaced by $\hat{\pi}_i$ in eq. (1). This test statistic tests the significance of the overall treatment effect as χ^2_C by using the chi square distribution with 1 degree of freedom.

In the case of pre-stratification, by using eq. (6), the estimator of the variance in eq. (4) can be reduced to

$$\widehat{Var}_{C\cdot pre}(\hat{\delta}) = \frac{k_C}{k_T + k_C} \sum_i n_{Ti}\hat{p}_i(1-\hat{p}_i).$$

On the other hand, if π_i is replaced by $\hat{\pi}_i$, the estimator of the variance in eq. (8) can be represented as

$$\widehat{Var}_{pre}(\hat{\delta}) = \widehat{Var}_{C\cdot pre}(\hat{\delta}) + N_T \left(\frac{k_C}{k_T + k_C}\right)^2 \sum_i (\hat{\delta}_i - \hat{\delta})^2. \tag{10}$$

Because the second term in eq. (10) is always a positive value, if the number of patients is not fixed at the beginning of a trial, and randomization is conducted by pre-stratification, the value of the statistic in eq. (2) will be larger than that in eq. (9). As a result, using the Cochran test in such situations carries the risk of over-detecting the treatment effect. In addition, the value of this term approaches 0 when the true value of δ_i equals δ in the arbitrary stratum because $\hat{\delta}_i$ and $\hat{\delta}$ are consistent estimators of δ_i and δ, respectively. In this case, the value of the statistic in eq. (9) will be close to that in eq. (2) because of the increase in the sample size. When the sample size is small, because the variability of the difference between $\hat{\delta}_i$ and $\hat{\delta}$ in each stratum is large, the second term in eq. (10) tends to have a large value, and as a result, our proposed statistic will yield a more conservative result.

4 Proposed test under post-stratification

In a post-stratification analysis, the randomization is conducted for all patients. Although the total number of patients that will be assigned to either of the two groups is fixed, the ratios of the number of patients in the strata could differ according to each stratum. Therefore, in this case, we can only assume that N_T and N_C are fixed. Of course, in this case, the variation in the number of patients is expected to be larger than that in the case of pre-stratification, because n_{Ti} and n_{Ci} follow independent distributions.

The variance of $\hat{\delta}$ can be calculated from eq. (7) as in the pre-stratification case. However, as an additional difficulty that is distinct from the previous case, it is necessary to calculate the expected value and variance of a complex function that is constructed from $n_{T1}, n_{T2}, \cdots, n_{TK}$ and $n_{C1}, n_{C2}, \cdots, n_{CK}$. Although the details of the calculation are given in Appendix B, to address this problem, we used the multivariate first-order Taylor expansion around the expected values. As a result, the approximation of the variance of $\hat{\delta}$ is given by

$$\begin{aligned} Var(\hat{\delta}) \approx\ & \frac{N_T N_C}{N^2} \sum_i \{N_C \pi_i p_{Ti}(1-p_{Ti}) + N_T \pi_i p_{Ci}(1-p_{Ci})\} \\ +\ & \frac{N_T N_C}{N^4}(N_T^3 + N_C^3) \sum_i \pi_i (\delta_i - \sum_j \pi_j \delta_j)^2. \end{aligned} \tag{11}$$

With replacement of unknown parameters p_{Ti} and p_{Ci} by $\hat{p}_i$ in eq. (5), and replacement of δ_i by the estimator, the estimation of eq. (11) is given by

$$\begin{aligned} \widehat{Var}_{post}(\hat{\delta}) =\ & \frac{N_T N_C}{N} \sum_i \pi_i \hat{p}_i(1-\hat{p}_i) \\ +\ & \frac{N_T N_C}{N^4}(N_T^3 + N_C^3) \sum_i \pi_i (\hat{\delta}_i - \sum_j \pi_j \hat{\delta}_j)^2. \end{aligned} \tag{12}$$

Then, we propose a new test statistic based on eq. (2) as follows:

$$\chi^2_{post} = \frac{\hat{\delta}^2}{\widehat{Var}_{post}(\hat{\delta})}. \tag{13}$$

As in the case of χ^2_{pre}, π_i is replaced by $\hat{\pi}_i$ in eq. (1) when π_i is unknown. The test statistic is compared to the chi square distribution with 1 degree of freedom to test the significance.

In post-stratification, the difference between $\widehat{Var}_C(\hat{\delta})$ and $\widehat{Var}_{post}(\hat{\delta})$ cannot be as easily compared as in the pre-stratification case with eq. (10). If π_i is unknown, the difference between the first term of eq. (12) and $\widehat{Var}_C(\hat{\delta})$ is given as

$$\sum_i \left\{ \frac{N_T N_C}{N^2} - \frac{n_{Ti} n_{Ci}}{n_i^2} \right\} n_i \hat{p}_i (1 - \hat{p}_i). \tag{14}$$

Therefore, the difference between these two values can be considered by the magnitude of the difference between the heterogeneities of the two groups, which are calculated from the total number of patients ($N_T N_C / N^2$) and the number of patients in each stratum ($n_{Ti} n_{Ci} / n_i^2$). When the total number of patients in the two groups is completely balanced ($N_T = N_C$), the value of eq. (14) always becomes positive, because the value of $n_i \hat{p}_i (1 - \hat{p}_i)$ is at least 0. In this case, if the numbers of patients in each stratum are completely balanced ($n_{Ti} = n_{Ci}, \forall i$), then eq. (14) becomes 0. When the total number of patients becomes large, the number of patients in each stratum becomes balanced, and as a result, the value of eq. (13) approximates eq. (2). On the other hand, if the total number of patients is small, the imbalance of patients in each stratum becomes large within the poststratification framework, and as a result, the value of eq. (13) becomes more conservative than eq. (2).

Even if the total number of patients in the two groups is unbalanced, the value is expected to become positive since the number of patients in each stratum will be unbalanced in many cases. As a special case, when the total number of patients is unbalanced and the true values of p_{Ti}, p_{Ci}, and π_i are extremely skewed, eq. (14) could become a negative value. In almost all cases, however, the first term of eq. (12) becomes greater than $\widehat{Var}_C(\hat{\delta})$. In addition, the second term of eq. (12) always becomes greater than 0. As a result, the value of the test statistic based on eq. (13) will be less than χ^2_C in many situations. In addition, for the case of this unbalanced total number of patients, when the total number of patients becomes large, the difference between eqs (2) and (13) is expected to become small because the value in the first pair of parentheses of eq. (14) approaches 0.

5 Simulations

We here present the results of simulation studies to demonstrate the excessively high type I error obtained with the Cochran test and the reasonably controlled type I error obtained using the proposed test under the situation of a stratified randomized clinical trial. As mentioned above, there are several tests in addition to the Cochran test for assessing the overall treatment effects. However, since our proposed approaches are constructed based on eqs (2) and (3), we focus only on the comparison to the Cochran test in this study. To assess the type I error, we generated the simulated data from hypothetical stratified trials under several situations of $H_0 : \delta = 0$ and compared the percentage of times the null hypothesis was rejected among all tests according to the significance level $\alpha = 0.05$.

For each setting, the simulations were repeated 50,000 times. The number of treatment patients in strata $(n_{T1}, n_{T2}, \cdots, n_{TK})$ is given by a multinomial random number with N_T and $(\pi_1, \pi_2 \cdots, \pi_K)$. The number of patients in the control group is given by $(n_{C1}, n_{C2}, \cdots, n_{CK})$=$(n_{T1}, n_{T2}, \cdots, n_{TK}) \times k_C / k_T$ for the pre-stratification case. In the post-stratification case, $(n_{C1}, n_{C2}, \cdots, n_{CK})$ is given by a multinomial random number with N_C and $(\pi_1, \pi_2 \cdots, \pi_K)$.

For both pre-stratification and post-stratification, the settings of the simulations were equivalent. The only difference was whether or not there is the exact setting of the ratio of the number of patients in the two groups k_T / k_C. We set $k_T / k_C = 1$ for all simulations in the pre-stratification case. In the post-stratification case, we set $N_T = N_C$ for all settings. Therefore, the total number of patients in the two groups was set to be equivalent in all simulations for both cases. The number of strata K was set to 2, 5, or 10. The setting of the low number of strata is assumed for a situation in which there is a general covariate such as sex or stratified age that should be considered to affect the associations between the two groups and the response in a randomized clinical trial. On the other hand, the setting with a high number of strata is assumed to represent a large-scale study such as a multicenter trial or meta-analysis.

In the case of $K = 2$, the success probabilities of the two groups were set to $(p_{T1}, p_{T2}) = (p_{C1}, p_{C2}) = (0.7, 0.4)$. In the cases of $K = 5$ and $K = 10$, the probabilities were set to $(p_{T1}, \cdots, p_{T5}) = (p_{C1}, \cdots, p_{C5}) = (0.7, 0.6, 0.5, 0.4, 0.3)$ and $(p_{T1}, \cdots, p_{T10}) = (p_{C1}, \cdots, p_{C10}) = (0.75, 0.7, 0.65, 0.6, 0.55, 0.5, 0.45, 0.4, 0.35, 0.3)$, respectively. The other settings and results are shown in Table 1-Table 4. In all Tables, the percentages of times the null hypothesis was rejected by the Cochran test (χ^2_C) and the proposed tests (χ^2_{pre} or χ^2_{post}) are shown for the cases when $(\pi_1, \pi_2, \cdots, \pi_K)$ is known and unknown.

Table 1: Results of the 50,000 simulated randomized trials in the case of prestratification, and the stratum size $K = 2$. The type I errors for the Cochran test and proposed tests when π is known and unknown are shown. The significance level is set to $\alpha = 0.05$.

$[\pi_1, \pi_2]$	N_T	$\alpha_{\chi^2_C}$	$\alpha_{\chi^2_{pre}}$ (π **known**)	$\alpha_{\chi^2_{pre}}$ (π **unknown**)
[0.5, 0.5]	10	5.84	5.13	4.86
	20	5.46	4.89	4.86
	50	4.88	4.72	4.75
	100	4.83	4.73	4.73
	200	5.06	5.03	5.02
[0.7, 0.3]	10	5.74	4.67	4.36
	20	5.50	4.92	4.93
	50	4.97	4.80	4.81
	100	5.07	4.90	4.90
	200	5.00	4.96	4.96
[0.1, 0.9]	50	5.04	4.91	4.94
[0.2, 0.8]	50	4.97	4.89	4.90
[0.3, 0.7]	50	4.96	4.89	4.89
[0.4, 0.6]	50	4.94	4.82	4.80
[0.6, 0.4]	50	4.96	4.82	4.83
[0.7, 0.3]	50	4.97	4.80	4.81
[0.8, 0.2]	50	4.99	4.82	4.83
[0.9, 0.1]	50	5.16	4.93	4.91

Note: $\alpha_{\chi^2_C}$: type I errors for the Cochran test ×100, $\alpha_{\chi^2_{pre}}$ (π known): type I errors for the proposed tests when π is known ×100, $\alpha_{\chi^2_{pre}}$ (π unknown): type I errors for the proposed tests when π is unknown ×100.

Table 1 shows the results for the pre-stratification and $K = 2$ case. As expected, the percentage of times the null hypothesis was rejected was lower with the proposed approach than with the Cochran approach in all cases. In the two-strata case, the Cochran test showed a nearly nominal significance level in almost all cases. When the sample size was small, however, the percentage of times H_0 was rejected became too high, regardless of the strata proportion in the population. In addition, the proportion of times H_0 was rejected increased when the heterogeneity of the strata became very high. However, in such situations, the proposed test could also control the nominal significance level well. During a comparison of the results of the proposed tests when (π_1, π_2) was known and unknown, the test that estimates (π_1, π_2) tended to become more conservative when the sample size was small. After the number of samples was increased, these two test statistics quickly showed convergent values.

Table 2: Results of the 50,000 simulated randomized trials in the case of prestratification and the stratum size $K = 5$ and 10. The type I errors for the Cochran test and proposed tests when π is known and unknown are shown. The significance level is set to $\alpha = 0.05$.

K	$[\pi_1, \cdots, \pi_K]$	N_T	$\alpha_{\chi^2_C}$	$\alpha_{\chi^2_{pre}}$ (π **known**)	$\alpha_{\chi^2_{pre}}$ (π **unknown**)
5	[0.2, 0.2, 0.2, 0.2, 0.2]	25	6.00	4.34	4.32
		50	5.31	4.61	4.64
		125	5.38	4.94	4.92
		250	5.08	4.82	4.82
		500	5.04	4.93	4.94
5	[0.1, 0.1, 0.2, 0.3, 0.3]	25	6.05	3.88	3.95
		50	5.36	4.57	4.60
		125	5.35	4.86	4.87
		250	5.09	4.85	4.85
		500	5.13	5.05	5.05
10	[0.1, 0.1, $\cdots$, 0.1]	50	6.20	3.96	4.10

	100	5.52	4.47	4.49
	250	5.25	4.76	4.74
	500	5.03	4.83	4.84
	1000	4.95	4.81	4.81

Note: See note in Table 1.

Table 2 shows the results for the pre-stratification and $K = 5$ and 10 cases. If the number of strata was large, the Cochran test tended to over reject H_0 in almost all situations. On the other hand, the proposed test tended to be more conservative, especially when the sample size was small. Both tests approximated to the nominal level of the test as the number of samples increased. As the reason for this finding, the effect of the second term in eq. (10) can be considered. That is, when the total number of patients is small, the variation between estimators of the treatment effect in each stratum becomes large, and the value of the second term in eq. (10) tends to become large. Consequently, our proposed test tends to be conservative in this case. This second term in eq. (10) approximates 0 if the number of patients is increased in this simulation setting, and the result of our proposed test approximates the result of the Cochran test. Judging by the results in Table 1 and Table 2, if the sampling structure in a stratified randomized trial is a prestratification case, we recommend our approach to control the type I error at the nominal level. Especially, when the number of patients included in a stratum is less than 25, we strongly recommend our approach.

Table 3: Results of the 50,000 simulated randomized trials in the case of post-stratification, and the stratum size $K = 2$. The type I errors for the Cochran test and proposed tests when π is known and unknown are shown. The significance level is set to $\alpha = 0.05$.

$[\pi_1, \pi_2]$	N_T	$\alpha_{\chi^2_C}$	$\alpha_{\chi^2_{pre}}$ (π **known**)	$\alpha_{\chi^2_{pre}}$ (π **unknown**)
[0.5, 0.5]	10	5.93	5.03	4.87
	20	5.40	4.78	4.76
	50	5.27	5.02	5.04
	100	5.07	4.95	4.95
	200	5.19	5.13	5.12
[0.7, 0.3]	10	5.75	4.70	4.47
	20	5.46	4.80	4.78
	50	5.27	5.02	5.04
	100	5.03	4.92	4.92
	200	5.06	5.00	5.00
[0.1, 0.9]	50	5.05	4.81	4.81
[0.2, 0.8]	50	5.19	4.99	4.99
[0.3, 0.7]	50	5.18	4.96	4.95
[0.4, 0.6]	50	5.18	4.93	4.94
[0.6, 0.4]	50	5.27	4.99	4.99
[0.7, 0.3]	50	5.27	5.02	5.04
[0.8, 0.2]	50	5.27	4.98	4.98
[0.9, 0.1]	50	5.19	4.91	4.88

Note: See note in Table 1.

Table 4: Results of the 50,000 simulated randomized trials in the case of poststratification and the stratum size $K = 5$ and 10. The type I errors for the Cochran test and proposed tests when π is known and unknown are shown. The significance level is set to $\alpha = 0.05$.

K	$[\pi_1, \cdots, \pi_K]$	N_T	$\alpha_{\chi^2_C}$	$\alpha_{\chi^2_{pre}}$ (π **known**)	$\alpha_{\chi^2_{pre}}$ (π **unknown**)
5	[0.2, 0.2, 0.2, 0.2, 0.2]	25	6.09	4.09	4.05
		50	5.54	4.57	4.55
		125	4.99	4.67	4.67
		250	5.05	4.86	4.85
		500	5.00	4.91	4.92
5	[0.1, 0.1, 0.2, 0.3, 0.3]	25	6.02	3.20	3.18
		50	5.56	4.49	4.47
		125	5.04	4.65	4.65
		250	5.02	4.84	4.83

		500	5.04	4.97	4.97
10	$[0.1, 0.1, \cdots, 0.1]$	50	6.31	3.71	3.66
		100	5.58	4.47	4.46
		250	5.14	4.75	4.74
		500	5.06	4.88	4.88
		1,000	5.13	5.02	5.02

Note: See note in Table 1.

Table 3 and Table 4 show the results for the case of post-stratification with the same settings shown in Table 1 and Table 2, respectively. Essentially, the results were the same as observed in the corresponding pre-stratification cases. However, in the post-stratification analyses, the tendency of over rejection of the null hypothesis by the Cochran test was stronger. The proposed approach showed better performance than the traditional approach in all situations, especially for the results shown in Table 3. In accordance with these results, we recommend the proposed approach when the sampling structure is the poststratification framework. In particular, when the number of patients included in a stratum is less than 25, we strongly recommend our approach as in the case of the prestratification.

These simulation studies confirmed that the Cochran approach has a risk of obtaining a higher type I error than the nominal level in both pre-stratification and post-stratification situations. In particular, when the number of patients is small in a post-stratification analysis, this risk becomes higher. On the other hand, as expected, our proposed approach could control the type I error well. Although the proposed approach tends to become slightly conservative, it can be considered to more appropriate than the Cochran test with respect to consideration of the significance of a type I error in clinical trials. Of course as the trade-off of increasing the estimation value of the variance of $\hat{\delta}$, our approach has lower power than the traditional approach. However, as previously mentioned, the type I error has to be controlled by the nominal significance level first in almost all situations.

6 Examples

We here present two examples of the analysis of randomized clinical trial data using the proposed methods.

6.1 Data from esophagitis patients

As the first example, we analyzed the data of patients with reflux esophagitis. This study was first reported by Vigneri et al. [15] and then subsequently taken up by Berger et al. [16]. The data are shown in Table 5. First, the patients were stratified according to the initial grade of esophagitis (Grade 1 or Grade 2). Then, the patients were randomly assigned to the treatment with cisapride or omeprazole at the same ratio. Therefore, this study can be viewed as a prestratification trial at $k_T/k_C = 1$.

Table 5: Esophagitis patient data.

Grade	Treatment	Response Recurrence	No recurrence
1	Cisapride	3	12
	Omeprazole	0	15
2	Cisapride	4	11
	Omeprazole	2	13

The p-value of the Cochran test for the significance of the overall treatment difference was about 0.0679, and if we choose the significance level as 0.05, there would be no reason to reject the hypothesis that the probabilities of recurrence under the two treatments are equal. In our approach, we used χ^2_{pre} because (π_1, π_2) is unknown, and the p-value was approximately 0.0685. Although the conclusions from the two tests are the same, the p-value of the proposed test was slightly greater than that of the Cochran test, as expected.

6.2 Data of patients with duodenal ulcers

As the second example, we analyzed the hypothetical clinical trial data of patients with duodenal ulcers. The details of these data are described in Blum [17] and the data were analyzed in Lachin [1]. Although the data were obtained from two hypothetical separate studies described in Blum [17], we focused on one side only, as in Lachin [1]. In the clinical trial, 200 patients with any of three ulcer types were randomized to either a drug or placebo group without consideration of ulcer type. Therefore, this simple randomized trial can be viewed as a poststratification trial at $N_T = N_C = 100$. The results of this trial are given in Table 6.

Table 6: Hypothetical duodenal ulcers patient data.

Group	Treatment	Response	
		Succes	Failuer
Drug-dependent	Drug	16	26
	Placebo	20	27
Acid-dependent	Drug	9	3
	Placebo	4	5
Intermediate	Drug	28	18
	Placebo	16	28

To estimate the overall treatment effect, we used the methods discussed in [10]. First, the unadjusted estimate of the risk difference, which is merely difference of the sample means in drug and placebo groups, was 0.13. Second, the doubly-robust weighted least squares (DR-WLS) estimate, which uses weighted logistic regression model to estimate the conditional probability of response given treatment effect and ulcer type, was 0.1225. This estimator is attributed to Marshall Joffe by Robins et al. [18].

The PLEASE ("precise, locally, augmented, simple estimator") estimate, which was first introduced by Colantuoni and Rosenblum [10], was 0.1225. The value is exactly same as that of the DR-WLS estimate in this setting. It is easy to understand why.

Like the DR-WLS estimate, the PLEASE uses a weighted logistic regression model to estimate the conditional probability of response. The only difference lies in the addition of new variables into the logistic regression model for the weight calculation. In this data setting, however, the newly added variables are linear combinations of the variables originally included in the model. Therefore, the coefficient estimates of these new variables in the logistic regression model equal 0, and the weights for calculating the PLEASE and the DR-WLS coincide. The R and SAS code for implementation of the estimators is given in the supplementary material of Colantuoni and Rosenblum [10].

The estimated standard errors of the unadjusted estimator and the DR-WLS estimator using the nonparametric bootstrap are both 0.0700. The p-values obtained by the Wald tests for the unadjusted and DR-WLS estimates were about 0.0632 and 0.0801, respectively.

The p-value obtained by the Cochran test was about 0.0807, and the null hypothesis was not rejected based on a significance level of 0.05. On the other hand, the p-value obtained by the proposed approach χ^2_{post}, where (π_1, π_2, π_3) is unknown, was about 0.0848. As expected, the proposed approach once again gave a slightly more conservative result.

7 Conclusion

In this paper, we focused on the variability in the number of patients in different strata of randomized clinical trials with a binary endpoint. In traditional approaches to test the significance of the overall treatment effect, the number of patients in each stratum is assumed to be fixed. However, in several situations of comparative trials, the patient number in different strata is allowed to vary. Therefore, using traditional approaches in such situations comes with a potential risk of underestimating the variability of the estimated treatment effect, which consequently increases the type I error over the nominal level. This problem can be illustrated with the results of simulation studies.

To deal with this problem, we calculated the new variance of the estimated treatment effect defined by the risk difference in both pre-stratification and post-stratification situations. We assumed that the number of patients follows a multinomial distribution. As seen in eqs (10), (14), and the results of the simulation studies, the proposed approach could effectively control the type I error in almost all situations. Furthermore, in examples of patient data, our approaches clearly resulted in greater p-values than those yielded by the Cochran test.

Although we focused on the test of the significance of the overall treatment effect in this study, our approach can easily be extended to the calculation of the confidence interval of the treatment effect. Of course, in such cases, the interval constructed from our approach would be expected to include the true overall treatment effect with a nominal level, although the length of the interval would become wider than that of traditional approaches.

From the perspective of increasing the variance, our approach can be considered to be similar to an approach that considers a random effect of the treatment between strata. However, the assumptions of the two approaches are essentially different. In fact, the rejection probabilities were almost 0 for all situations modeled in our simulation when using DerSimonian's approach [13], which is a typical method used in meta-analyses to consider random effects with a binary endpoint.

Finally, our approach can be extended to applications with many methods, including other criteria for evaluating a treatment effect such as relative risk or odds ratio, other sampling structures such as a case-control study, and other endpoints such as a continuous variable or inclusion of censored cases. In several cases, it is difficult to calculate the variances exactly, and the obtained formula would be complex. However, it is worth carefully considering the variation in each of these situations from the point of view of aiming to achieve precise control over the type I error.

References

[1] Lachin JM. Biostatistical methods: the assessment of relative risks, 2nd ed New York: John Wiley & Sons, 2011.
[2] Agresti A. Categorical data analysis, 3rd ed Hoboken, NJ: John Wiley & Sons, 2013.
[3] Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions, 3rd ed New York: John Wiley & Sons, 2003.
[4] Mehrotra DV, Railkar R. Minimum risk weights for comparing treatments in stratified binomial trials. Stat Med. 2000;19:811–25.
[5] Cochran WG. Some methods for strengthening the tests. Biometrics. 1954;10:417–51.
[6] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J National Cancer Inst. 1959;22:719–48.
[7] Sánchez-Meca J, Marín-Martńez F. Testing the significance of a common risk difference in meta-analysis. Comput Stat & Data Anal. 2000;33:299–313.
[8] Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis. 1985;27:335–71.
[9] Mehrotra DV. Stratification issues with binary endpoints. Drug Inf J. 2001;35:1343–50.
[10] Colantuoni E, Rosenblum M. Leveraging prognostic baseline variables to gain precision in randomized trials. Stat Med. 2015;34:2602–15.
[11] Ganju J, Zhou K. The benefit of stratification in clinical trials revisited. Stat Med. 2011;30:2881–9.
[12] Ganju J, Mehrotra DV. Stratified experiments reexamined with emphasis on multicenter trials. Controlled Clin Trials. 2003;24:167–81.
[13] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clin Trials. 1986;7:177–88.
[14] Emerson JD, Hoaglin DC, Mosteller F. Simple robust procedures for combining risk differences in sets of 2×2 tables. Stat Med. 1996;15:1465–88.
[15] Vigneri S, Termini R, Leandro G, Badalamenti S, Pantalena M, Savarino V, et al. A comparison of five maintenance therapies for reflux esophagitis. N Engl J Med. 1995;333:1106–10.
[16] Berger VW, Stefanescu C, Zhou YY. The analysis of stratified 2×2 contingency tables. Biom J. 2006;48:992–1007.
[17] Blum AL. Principles for selection and exclusion. In: Tygstrup N, Lachin JM, Juhl E, editors. The randomized clinical trial and therapeutic decisions. New York: Marcel Dekker, 1982:43–58.
[18] Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. Stat Sci. 2007;22:544–59.
[19] Dieters MJ, White TL, Littell RC, Hodge GR. Applicaton of approximate variances of variance components and their ratios in genetic tests. Theor Appl Genet. 1995;91:15–24.

Appendix A

The derivation of eq. (8) is given in this appendix. The notations and assumptions are the same as described in Section 2 and Section 3. The variance of $\hat{\delta}_i$ conditional on $n_{Ti} = m_{Ti}$ is given by:

$$\begin{aligned} Var\left(\hat{\delta}_i \middle| n_{Ti} = m_{Ti}\right) &= Var\left(\hat{p}_{Ti} - \hat{p}_{Ci} \middle| n_{Ti} = m_{Ti}\right) \\ &= \frac{1}{m_{Ti}^2} Var\left(a_i \middle| n_{Ti} = m_{Ti}\right) + \frac{1}{m_{Ci}^2} Var\left(b_i \middle| n_{Ti} = m_{Ti}\right) \\ &= \frac{1}{k_C m_{Ti}} \left\{k_C p_{Ti}(1-p_{Ti}) + k_T p_{Ci}(1-p_{Ci})\right\}. \end{aligned} \tag{15}$$

The expected value of $\hat{\delta}_i$ conditional on $n_{Ti} = m_{Ti}$ is given by:

$$\begin{aligned} E\left(\hat{\delta}_i \middle| n_{Ti} = m_{Ti}\right) &= E\left(\hat{p}_{Ti} - \hat{p}_{Ci} \middle| n_{Ti} = m_{Ti}\right) \\ &= \frac{1}{m_{Ti}} E\left(a_i \middle| n_{Ti} = m_{Ti}\right) + \frac{1}{m_{Ci}} E\left(b_i \middle| n_{Ti} = m_{Ti}\right) \\ &= \delta_i. \end{aligned} \tag{16}$$

Under the assumption of pre-stratification, the harmonic mean in eq. (3) can be expanded as

$$w_i = \frac{k_C}{k_T + k_C} n_{Ti}.$$

Then, the variance of $\hat{\delta}$ can be extended as follows:

$$\begin{aligned} Var_{pre}(\hat{\delta}) &= Var\left(\frac{k_C}{k_T + k_C} \sum_i n_{Ti} \hat{\delta}_i\right) \\ &= E\left[\left(\frac{k_C}{k_T + k_C}\right)^2 \sum_i m_{Ti}^2 Var\left(\hat{\delta}_i \middle| n_{Ti} = m_{Ti}\right)\right] \\ &\quad + Var\left[\frac{k_C}{k_T + k_C} \sum_i m_{Ti} E\left(\hat{\delta}_i \middle| n_{Ti} = m_{Ti}\right)\right]. \end{aligned} \tag{17}$$

By substituting eqs (15) and (16) to eq. (17), $Var_{pre}(\hat{\delta})$ is extended as follows:

$$\begin{aligned} Var_{pre}(\hat{\delta}) &= \left(\frac{k_C}{k_T + k_C}\right)^2 \frac{1}{k_C} \sum_i \left\{k_C p_{Ti}(1-p_{Ti}) + k_T p_{Ci}(1-p_{Ci})\right\} E(m_{Ti}) \\ &\quad + \left(\frac{k_C}{k_T + k_C}\right)^2 Var\left(\sum_i \delta_i m_{Ti}\right) \\ &= \left(\frac{k_C}{k_T + k_C}\right)^2 \frac{1}{k_C} \sum_i \left\{k_C p_{Ti}(1-p_{Ti}) + k_T p_{Ci}(1-p_{Ci})\right\} E(m_{Ti}) \\ &\quad + \left(\frac{k_C}{k_T + k_C}\right)^2 \left\{\sum_i \delta_i^2 Var(m_{Ti}) + 2 \sum_{i<} \sum_j \delta_i \delta_j Cov(m_{Ti}, m_{Tj})\right\}. \end{aligned}$$

From the assumption that the number of patients follows a multinomial distribution, $E(m_{Ti}) = N_T \pi_i$, $Var(m_{Ti}) = N_T \pi_i (1-\pi_i)$, and $Cov(m_{Ti}, m_{Tj}) = -N_T \pi_i \pi_j$ for $i \neq j$. Then, $Var_{pre}(\hat{\delta})$ is given by

$$\begin{aligned} Var_{pre}(\hat{\delta}) &= N_T \left(\frac{k_C}{k_T + k_C}\right)^2 \left[\frac{1}{k_C} \sum_i \left\{k_C p_{Ti}(1-p_{Ti}) + k_T p_{Ci}(1-p_{Ci})\right\} \pi_i \right. \\ &\quad \left. + \left\{\sum_i \delta_i^2 \pi_i (1-\pi_i) - 2 \sum_{i<} \sum_j \delta_i \delta_j \pi_i \pi_j\right\}\right]. \end{aligned} \tag{18}$$

Since $\sum \sum_{i<j} \delta_i \delta_j \pi_i \pi_j = (\sum_i \sum_j \delta_i \delta_j \pi_i \pi_j - \sum_i \delta_i^2 \pi_i^2)/2$ and $\sum_i \pi_i = 1$, the second term of eq. (18) is expanded as

$$\begin{aligned}\sum_i \delta_i^2 \pi_i(1-\pi_i) - 2\sum_{i<}\sum_j \delta_i\delta_j\pi_i\pi_j &= \sum_i \delta_i^2\pi_i - \sum_i\sum_j \delta_i\delta_j\pi_i\pi_j \\ &= \sum_i \delta_i^2\pi_i - \left(\sum_i \delta_i\pi_i\right)^2 \\ &= \sum_i \pi_i(\delta_i-\delta)^2. \end{aligned} \tag{19}$$

By substituting eq. (19) to eq. (18), we can obtain the formula of the variance of $\hat{\delta}$ given in the form of eq. (8).

Appendix B

The derivation of eq. (12) is given in this appendix. The notations and assumptions are same as those described in Section 2 and Section 4. In the post-stratification case, the variance and expected value of $\hat{\delta}_i = \hat{p}_{Ti} - \hat{p}_{Ci}$ are given by

$$Var\left(\hat{\delta}_i | n_{Ti} = m_{Ti}, n_{Ci} = m_{Ci}\right) = \frac{1}{m_{Ti}} p_{Ti}(1-p_{Ti}) + \frac{1}{m_{Ci}} p_{Ci}(1-p_{Ci}) \tag{20}$$

and

$$E\left(\hat{\delta}_i \middle| n_{Ti} = m_{Ti}, n_{Ci} = m_{Ci}\right) = \delta_i, \tag{21}$$

respectively.

By using eq. (20), the variance of $\hat{\delta} = \sum_i w_i \hat{\delta}_i$, where w_i is defined by eq. (3), conditional on n_{Ti} and n_{Ci}, is given by

$$\begin{aligned} &Var\left(\hat{\delta} \middle| n_{Ti} = m_{Ti}, n_{Ci} = m_{Ci}, \forall i\right) \\ &= \sum_i \frac{m_{Ti}m_{Ci}}{(m_{Ti}+m_{Ci})^2}\left\{m_{Ci}p_{Ti}(1-p_{Ti}) + m_{Ti}p_{Ci}(1-p_{Ci})\right\}. \end{aligned} \tag{22}$$

Similarly, by using eq. (21), the conditional expectation of $\hat{\delta}$ is given by

$$E\left(\hat{\delta} \middle| n_{Ti} = m_{Ti}, n_{Ci} = m_{Ci}, \forall i\right) = \sum_i \frac{m_{Ti}m_{Ci}}{m_{Ti}+m_{Ci}}\delta_i. \tag{23}$$

Since these formula are a complex function of $m_{T1}, m_{T2}, \cdots, m_{TK}$ and $m_{C1}, m_{C2}, \cdots, m_{CK}$, the exact expressions of the expectation of eq. (22) and variance of eq. (23) are difficult to derive. Therefore, we calculate the approximation formulas by using the Taylor expansion. The same approximation expansion for the variances of a multivariate function was reported by Dieters et al. [19]. For any $f(x_1, x_2, \cdots, x_n)$, the multivariate first-order Taylor expansion for $\theta = (\theta_1, \theta_2, \cdots, \theta_n)$ is

$$f(X_1, X_2, \cdots, X_n) = f(\theta) + \sum_i \frac{\partial f(\theta)}{\partial x_i}(X_i - \theta_i) + O(n^{-r}),$$

where $O(n^{-r})$ is a remainder term. By replacing $\theta = (E(X_1), E(X_2), \cdots, E(X_n))$, the expectation for $f(X_1, X_2, \cdots, X_n)$ is given by

$$E\left[f(X_1, X_2, \cdots, X_n)\right] = f(\theta) + E\left[O(n^{-r})\right]. \tag{24}$$

If we assume that the expectations of all the second-order and higher-order terms in $O(n^{-r})$ are negligible, then the expectation of $f(X_1, X_2, \cdots, X_n)$ is given by $f(E(X_1), E(X_2), \cdots, E(X_n))$.

Similarly, the variance of $f(X_1, X_2, \cdots, X_n)$ is derived as follows:

$$
\begin{aligned}
&Var\left[f(X_1, X_2, \cdots, X_n)\right] = E\left[\{f(X_1, X_2, \cdots, X_n) - E[f(X_1, X_2, \cdots, X_n)]\}^2\right] \\
&= E\left[\left\{\sum_i \frac{\partial f(\theta)}{\partial x_i}(X_i - \theta_i)\right\}^2 + 2\left\{\sum_i \frac{\partial f(\theta)}{\partial x_i}(X_i - \theta_i)\right\}\{O(n^{-r}) - E\left[O(n^{-r})\right]\}\right] \\
&\quad + Var\left[O(n^{-r})\right] \qquad (25)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_i \left(\frac{\partial f(\theta)}{\partial x_i}\right)^2 Var(X_i) + 2\sum_{i<} \sum_j \left(\frac{\partial f(\theta)}{\partial x_i}\right)\left(\frac{\partial f(\theta)}{\partial x_j}\right) Cov(X_i, X_j) \\
&\quad + 2\sum_i \frac{\partial f(\theta)}{\partial x_i} Cov\left[X_i, O(n^{-r})\right] + Var\left[O(n^{-r})\right]. \qquad (26)
\end{aligned}
$$

Therefore, if we assume that the variance of $O(n^{-r})$ and the covariances between X_i's and $O(n^{-r})$ are negligible, then the variance of $f(X_1, X_2, \cdots, X_n)$ is expressed as the linear sum of $Var(X_i)$'s and $Cov(X_i, X_j)$'s. The assumption of the negligible covariances between X_i's and $O(n^{-r})$ can be regarded the same as the assumption that the difference between $O(n^{-r})$ and its expectation is negligible according to eq. (25).

By using eq. (24) and the multinomial distribution assumption, $E(m_{Ti}) = N_T\pi_i$ and $E(m_{Ci}) = N_C\pi_i$, the approximation of the expectation of eq. (22) is given by

$$
\begin{aligned}
&E\left[Var\left(\hat{\delta}\middle| n_{Ti} = m_{Ti}, n_{Ci} = m_{Ci}, \forall i\right)\right] \\
&\approx \sum_i \frac{N_T N_C}{N^2}\left\{N_C\pi_i p_{Ti}(1 - p_{Ti}) + N_T\pi_i p_{Ci}(1 - p_{Ci})\right\}. \qquad (27)
\end{aligned}
$$

By deriving the first partial derivatives of eq. (23) with respect to m_{Ti} and m_{Ci} and evaluating them by $E(m_{Ti})$ and $E(m_{Ci})$, we can get

$$
\left.\frac{\partial E\left(\hat{\delta}\middle| n_{Ti} = m_{Ti}, n_{Ci} = m_{Ci}, \forall i\right)}{\partial m_{Ti}}\right|_{E(m_{Ti}), E(m_{Ci})} = \frac{N_C^2}{N^2}\delta_i, \qquad (28)
$$

and

$$
\left.\frac{\partial E\left(\hat{\delta}\middle| n_{Ti} = m_{Ti}, n_{Ci} = m_{Ci}, \forall i\right)}{\partial m_{Ci}}\right|_{E(m_{Ti}), E(m_{Ci})} = \frac{N_T^2}{N^2}\delta_i. \qquad (29)
$$

From eqs (26), (28), (29), the multinomial distribution assumption, $Var(m_{Ti}) = N_T\pi_i(1 - \pi_i)$, $Var(m_{Ci}) = N_C\pi_i(1 - \pi_i)$, $Cov(m_{Ti}, m_{Tj}) = -N_T\pi_i\pi_j$, $Cov(m_{Ci}, m_{Cj}) = -N_C\pi_i\pi_j$, and the independence of two groups $Cov(m_{Ti}, m_{Cj}) = 0$, the approximation of the variance of eq. (23) is given by

$$
\begin{aligned}
&Var\left[E\left(\hat{\delta}\middle| n_{Ti} = m_{Ti}, n_{Ci} = m_{Ci}, \forall i\right)\right] \\
&\approx \sum_i \left(\frac{N_C^2}{N^2}\delta_i\right)^2 N_T\pi_i(1 - \pi_i) + \sum_i \left(\frac{N_T^2}{N^2}\delta_i\right)^2 N_C\pi_i(1 - \pi_i) \\
&\quad - 2\sum_{i<} \sum_j \left(\frac{N_C^2}{N^2}\delta_i\right)\left(\frac{N_C^2}{N^2}\delta_j\right) N_T\pi_i\pi_j \\
&\quad - 2\sum_{i<} \sum_j \left(\frac{N_T^2}{N^2}\delta_i\right)\left(\frac{N_T^2}{N^2}\delta_j\right) N_C\pi_i\pi_j. \qquad (30)
\end{aligned}
$$

Moreover, by the same expansion used in eqs (19), (30) can be expressed as follows:

$$
\begin{aligned}
&Var\left[E\left(\hat{\delta}\middle| n_{Ti} = m_{Ti}, n_{Ci} = m_{Ci}, \forall i\right)\right] \\
&\approx \frac{N_T N_C^4}{N^4}\sum_i \pi_i\left(\delta_i - \sum_j \pi_j\delta_j\right)^2 \\
&\quad + \frac{N_T^4 N_C}{N^4}\sum_i \pi_i\left(\delta_i - \sum_j \pi_j\delta_j\right)^2. \qquad (31)
\end{aligned}
$$

Therefore, we can obtain the approximation formula (12) for the variance of $\hat{\delta}$ under post-stratification by taking the sum of eqs (27) and (31).