Published in final edited form as:
Biom J. 2014 January ; 56(1): . doi:10.1002/bimj.201300048.

Stratified Fisher's Exact Test and its Sample Size Calculation

Sin-Ho Jung[1]
Department of Biotatistics and Bioinformatics Duke University Durham, NC 27710, U.S.A.

Summary

Chi-squared test has been a popular approach to the analysis of a 2 × 2 table when the sample sizes for the four cells are large. When the large sample assumption does not hold, however, we need an exact testing method such as Fisher's test. When the study population is heterogeneous, we often partition the subjects into multiple strata, so that each stratum consists of homogeneous subjects and hence the stratified analysis has an improved testing power. While Mantel-Haenszel test has been widely used as an extension of the chi-squared test to test on stratified 2×2 tables with a large-sample approximation, we have been lacking an extension of Fisher's test for stratified exact testing. In this paper, we discuss an exact testing method for stratified 2 × 2 tables which is simplified to the standard Fisher's test in single 2 × 2 table cases, and propose its sample size calculation method that can be useful for designing a study with rare cell frequencies.

Keywords

Conditional type I error; Exact test; Hypergeometric distribution; Many 2 × 2 tables; Odds ratio

1 Introduction

In this paper, we discuss an exact test for stratified 2 × 2 tables with rare cell frequencies. Since the stratified exact test is simplified to the standard Fisher's (1935) exact test in single 2 × 2 table cases, we call it stratified Fisher's test.

Suppose that we want to compare the response probabilities between two groups, experimental (or case) and control. Oftentimes in a two group comparison, the characteristics of study subjects may be heterogeneous. In this case, the heterogeneity is characterized by some stratification factors, and a stratified method is applied in the final analysis. When the distribution of the stratification factors is identical between two groups, an unstratified testing ignoring the population heterogeneity controls the type I error rate but loses the efficiency. If the distribution of the stratification factors is different between two groups, however, an unstratified testing does not control the type I error rate. We want to test if the two groups have equal response probabilities or not while accounting for heterogeneity of the population defined by strata.

Multiple asymptotic testing methods have been proposed for testing on stratified 2 × 2 tables. Under the assumption that the odds ratios are identical among strata, Cochran (1954) proposes an asymptotic method for testing if the common odds ratio is 1 or not. Under the assumption of constant risk ratios across strata, Gart (1985) proposes an asymptotic method for testing if the common risk ratio is 1 or not. Woolson et al. (1986) and Nam (1992)

[1] Contact information: sinho.jung@duke.edu; Phone: 919-668-8658; Fax: 919-668-5888; Postal address: 2424 Erwin Road, Suit 1102, DUMC Box 2721, Durham, NC 27710.

propose sample size calculation methods for Cochran test, and Nam (1998) proposes a sample size method for Gart test.

In order to use these tests for testing on many 2 × 2 tables, we have to check the assumptions of common odds ratios or risk ratios in advance. For testing the common odds ratio assumption, Zelen (1971) proposes an exact method, which is implemented by StatXact, and Breslow and Day (1980) propose an asymptotic method.

If these assumptions do not seem to be valid, we need a robust test requiring no assumptions on the primary parameters for testing. Mantel and Haenszel (1959) propose an asymptotic test for testing if two groups have equal response probabilities without any assumption of common odds ratio or common risk ratio. Jung et al. (2007) propose a sample size calculation method for Mantel-Haenszel test. In this paper, we extend Fisher's exact test for testing stratified 2 × 2 tables with rare cell frequencies, and propose its sample size calculation method. These methods can be used in designing and analyzing small case-control studies or clinical trials. The input parameters to be specified for the sample size calculation of stratified Fisher's exact test are exactly the same as those for the sample size calculation of Mantel-Haenszel test. We will compare the performance of the proposed test with that of the asymptotic Mantel-Haenszel test and the standard Fisher's exact test ignoring strata under some practical settings.

2 Stratified Fisher's Exact Test

Suppose that there are J strata. Let N denote the total sample size, and n_j the sample size in stratum $j\left(\sum_{j=1}^{J} n_j=N\right)$. Among n_j subjects in stratum $j(= 1, ..., J)$, m_j are allocated to group 1 (case or experimental) and $\bar{m}_j$ to group 2 (control). For stratum j, group 1 has a response probability p_j and group 2 has a response probability q_j. Let $\bar{p}_j = 1 - p_j$, $\bar{q}_j = 1 - q_j$, and $\theta_j = p_j\bar{q}_j/(q_j\bar{p}_j)$ denote the odds ratio in stratum j. Suppose that we want to test

$$H_0{:}\theta_1=\cdots=\theta_J=1$$

against

$$H_1{:}\theta_j>1 \quad \text{for} \;\; \text{some} \quad j=1,\ldots,J.$$

For stratum $j(= 1, ..., J)$, let x_j and y_j denote the numbers of responders for groups 1 and 2, respectively, and $z_j = x_j + y_j$ denote the total number of responses. The frequency data in stratum j can be described as in Table 1.

We propose to reject H_0 in favor of H_1 if $S=\sum_{j=1}^{J} x_j$ is large. Under H_0, conditioning on the margin totals (z_j, m_j, n_j), x_j has the hypergeometric distribution

$$f_0\left(x_j|z_j,m_j,n_j\right)=\frac{\begin{pmatrix} m_j \\ x_j \end{pmatrix}\begin{pmatrix} \bar{m}_j \\ z_j-x_j \end{pmatrix}}{\sum_{i=m_{j-}}^{m_{j+}}\begin{pmatrix} m_j \\ i \end{pmatrix}\begin{pmatrix} \bar{m}_j \\ z_j-i \end{pmatrix}}$$

for $m_{j-} \le x_j \le m_{j+}$, where $m_{j-} = \max(0, z_j - \bar{m}_j)$ and $m_{j+} = \min(z_j, m_j)$. Let $\boldsymbol{z} = (z_1, ..., z_J)$, $\boldsymbol{m} = (m_1, ..., m_J)$ and $\boldsymbol{n} = (n_1, ..., n_J)$. Given $(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n})$, the conditional p-value for $s = \sum_{j=1}^{J} x_j$, pv = pv$(s|\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n})$, is obtained by

$$\text{pv}=P\left(S \ge s|z,\boldsymbol{m},\boldsymbol{n},H_0\right)=\sum_{i_1=m_{1-}}^{m_{1+}} \cdots \sum_{i_J=m_{J-}}^{m_{J+}} I\left(\sum_{j=1}^{J} i_j \ge s\right)\prod_{j=1}^{J} f_0\left(i_j|z_j,m_j,n_j\right).$$

Given type I error rate α^*, we reject H_0 if pv < α^*.

Similarly for the other one-sided alternative hypothesis

$$H_2{:}\theta_j<1 \quad \text{for} \quad \text{some} \quad j=1,\ldots,J,$$

the conditional p-value given $(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n})$ is obtained by

$$\text{pv}=P\left(S \le s|z,\boldsymbol{m},\boldsymbol{n},H_0\right)=\sum_{i_1=m_{1-}}^{m_{1+}} \cdots \sum_{i_J=m_{J-}}^{m_{J+}} I\left(\sum_{j=1}^{J} i_j \le s\right)\prod_{j=1}^{J} f_0\left(i_j|z_j,m_j,n_j\right).$$

A two-sided p-value may be calculated as two times the minimum of the two one-sided p-values. Without loss of generality, we focus our discussions on the one-sided alternative hypothesis H_1 in our paper.

Note that Mantel-Haenszel test also rejects H_0 in favor of H_1 for a large value of S, and its p-value is calculated using the standardized test statistic

$$W=\frac{S-E}{\sqrt{V}}$$

which is asymptotically $N(0, 1)$ under H_0, where $E=\sum_{j=1}^{J} E_j$, $V=\sum_{j=1}^{J} V_j$, $E_j = z_j m_j / n_j$ and $V_j=z_j m_j \bar{m}_j \left(n_j - z_j\right)/\left\{n_j^2\left(n_j-1\right)\right\}$. Westfall, Zaykin and Young (2002) propose a permutation procedure for stratified Mantel-Haenszel test, which permutes the two-sample binary data within each stratum in the context of multiple testing. Their permutation maintains the margin totals for 2 × 2 tables, $\{(z_j, m_j, n_j), 1 \le j \le J\}$, and E_j and V_j depend on the margin totals only, so that the permutation-based Mantel-Haenszel test will be identical to our stratified Fisher's exact test if they go through all the possible $\prod_{j=1}^{J}\left(m_{j+}-m_{j-}+1\right)$ permutations. Their permutation test is implemented by SAS. Compared to our exact test, the permutation test requires a much longer computing time. Furthermore, a permutation test often randomly selects partial permutations to approximate the exact p-value. In this case, the resulting approximate p-value will be different depending on the selected seed number for random number generation or the number of permutations, while the exact method always provides a constant exact p-value.

A real data example is taken from Li et al. (1979), where the investigators are interested in whether thymosin (experimental), compared to placebo (control), has any effect in the treatment of bronchogenic carcinoma patients receiving radiotherapy. Table 2 summarizes the data for three strata. The one-sided p-values are 0.1563 by the stratified Fisher's exact test and 0.0760 by Mantel-Haenszel test. Stratified Fisher's test has a larger p-value than Mantel-Haenszel test because of its conservative type I error control as demonstrated in Section 4 or because of the very small numbers of failures across the strata that can lead to a biased p-value for the asymptotic Mantel-Haenszel test.

3 Power and Sample Size Calculation

Jung et al. (2007) propose a sample size calculation method for Mantel-Haenszel test. In this section, we derive a sample size formula for stratified Fisher's exact test by specifying the values of the same input parameters as those for Mantel-Haenszel test by Jung et al. (2007). Following are input parameters to be specified for a sample size calculation.

Input Parameters

- Type I and II error probabilities: (α^*, β^*)
- Success probabilities for group 2 (control): $(q_1, ..., q_J)$
- Odds ratios: $(\theta_1, ..., \theta_J)$ under H_1, where $\theta_j > 0$. Note that, given q_j and θ_j, the success probability for group 1 (experimental) is given as $p_j = \theta_j q_j/(\bar{q}_j + \theta_j q_j)$ in stratum $j(= 1, ..., J)$.
- Prevalence for each stratum: $(a_1, ..., a_J)$, where $a_j = E(n_j/N)$. Note that $a_j > 0$ and $\sum_{j=1}^{J} a_j = 1$.
- Allocation probability for group 1 (experimental) within each stratum, $(b_1, ..., b_J)$, where $b_j = E(m_j/n_j)$ with $0 < b_j < 1$.

3.1 When Group and Stratum Allocations are Random

In designing a study, N is fixed at a predetermined size corresponding to a specified power. At the moment, we assume that, given N, the strata sizes and the sample sizes for two groups within each stratum are randomly selected by the prevalence rate of each category in the population. Hence, given N, $\{(x_j, z_j, m_j, n_j), 1 \le j \le J\}$ are random variables with following marginal or conditional probability mass functions that are indexed by the above input parameters.

Distribution Functions

- Conditional distribution of x_j given (z_j, m_j, n_j):

$$f_j(x_j|z_j, m_j, n_j) = \frac{\binom{m_j}{x_j}\binom{\bar{m}_j}{z_j - x_j}\theta_j^{x_j}}{\sum_{i=m_{j-}}^{m_{j+}}\binom{m_j}{i}\binom{\bar{m}_j}{z_j - i}\theta_j^{i}}$$

  for $m_{j-} \le x_j \le m_{j+}$, where $m_{j-} = \max(0, z_j - \bar{m}_j$, $m_{j+} = \min(z_j, m_j)$ and $j = 1, ..., J$. Under H_0, this is simplified to $f_0(x_j|z_j, m_j, n_j)$.

- Conditional distribution of z_j given (m_j, n_j): Given (m_j, n_j), $x_j \sim B(m_j, p_j)$ and $y_j \sim B(\bar{m}_j, q_j)$ are independent, so that the conditional probability mass function of $z_j = x_j + y_j$ is expressed as

$$g_j\left(z_j|m_j,n_j\right)=\sum_{x=m_{j-}}^{m_{j+}}\left(\begin{array}{c}m_j\\x\end{array}\right)p_j^x\bar{p}_j^{m_j-x}\left(\begin{array}{c}\bar{m}_j\\z_j-x\end{array}\right)q_j^{z_j-x}\bar{q}_j^{\bar{m}_j-z_j+x}$$

for $z = 0, 1, ..., n_j$ and $j = 1, ..., J$, where $B(m, p)$ denotes the binomial distribution with number of trials m and success probability p. Under H_0, this is simplified to

$$g_{0j}\left(z_j|m_j,n_j\right)=q_j^{z_j}\bar{q}_j^{n_j-z_j}\sum_{x=m_{j-}}^{m_{j+}}\left(\begin{array}{c}m_j\\x\end{array}\right)\left(\begin{array}{c}\bar{m}_j\\z_j-x\end{array}\right).$$

Note that $\left(\begin{array}{c}0\\0\end{array}\right)p^0(1-p)^0=1$ for $p \in (0, 1)$.

- Conditional distribution of m_j given n_j: At the moment, we assume that, given a total sample size n_j of stratum j, the sample size of group 1 m_j is a binomial random variable with probability mass function

  $$h_j\left(m_j|n_j\right)=\left(\begin{array}{c}n_j\\m_j\end{array}\right)b_j^{m_j}(1-b_j)^{n_j-m_j}$$

  for $0 \le m_j \le n_j$ and $j = 1, ..., J$.

- Conditional distribution of $(n_1, ..., n_J)$ given N is multinomial with probability mass function

  $$l_N\left(n_1,\cdots,n_J\right)=\frac{N!}{\prod_{j=1}^J n_j!}\prod_{j=1}^J a_j^{n_j}$$

  for $0 \le n_1 \le N, ..., 0 \le n_J \le N$ and $\sum_{j=1}^J n_j=N$.

We first derive the power function for a given sample size N using these distribution functions. Given $(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n})$ and type I error rate α^*, the critical value $c_{\alpha^*} = c_{\alpha^*}(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n})$ is the smallest integer c satisfying

$$P\left(S\ge c|\boldsymbol{z},\boldsymbol{m},\boldsymbol{n},H_0\right)=\sum_{i_1=m_{1-}}^{m_{1+}}\cdots\sum_{i_J=m_{J-}}^{m_{J+}}I\left(\sum_{j=1}^J i_j\ge c\right)\prod_{j=1}^J f_0\left(i_j|z_j,m_j,n_j\right)\ge\alpha^*.$$

Note that $s \ge c_{\alpha^*}(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n})$ if and only if $\mathrm{pv}(s|\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n}) \le \alpha^*$. We call $\alpha(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n}) = P(S \ge c_{\alpha^*}|\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n}, H_0)$ the conditional type I error rate given $(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n})$. Similarly, the conditional power $1 - \beta(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n})$ given $(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{n})$ is obtained by

$$P\left(S\ge c_{\alpha^*}|\boldsymbol{z},\boldsymbol{m},\boldsymbol{n},H_1\right)=\sum_{i_1=m_{1-}}^{m_{1+}}\cdots\sum_{i_J=m_{J-}}^{m_{J+}}I\left(\sum_{j=1}^J i_j\ge c_{\alpha^*}\right)\prod_{j=1}^J f_j\left(i_j|z_j,m_j\right).$$

For a chosen *N*, the marginal type I error rate and power are given as

$$\begin{aligned}\alpha_{N} &\equiv E\left\{\alpha\left(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n}\right)|H_0\right\}\\ &=E^{\boldsymbol{n}}\left(E^{\boldsymbol{m}}\left[E^{\boldsymbol{z}}\left\{\alpha\left(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n}\right)|\boldsymbol{m},\boldsymbol{n},H_0\right\}\mid n\right]\right)\\ &=\sum_{\boldsymbol{n}\in\mathscr{D}_N}\sum_{m_1=0}^{n_1}\cdots\sum_{m_J=0}^{n_J}\sum_{z_1=m_{1-}}^{m_{1+}}\cdots\sum_{z_J=m_{J-}}^{m_{1+}}\alpha\left(z_1,\cdots,z_J;m_1,\cdots,m_J;n_1\cdots,n_J\right)\\ &\quad\times\left\{\prod_{j=1}^{J}g_{0j}\left(z_j|m_j,n_j\right)\right\}\left\{\prod_{j=1}^{J}h_j\left(m_j|n_j\right)\right\}l_N\left(n_1,\ldots,n_J\right)\end{aligned} \tag{1}$$

and

$$\begin{aligned}1-\beta_{N} &\equiv E\left\{1-\beta\left(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n}\right)|H_1\right\}\\ &=E^{\boldsymbol{n}}\left(^{\boldsymbol{m}}\left[E^{\boldsymbol{z}}\left\{1-\beta\left(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n}\right)|\boldsymbol{m},\boldsymbol{n},H_1\right\}\mid\boldsymbol{n}\right]\right)\\ &=\sum_{\boldsymbol{n}\in\mathscr{D}_N}\sum_{m_1=0}^{n_1}\cdots\sum_{m_J=0}^{n_J}\sum_{z_1=m_{1-}}^{m_{1+}}\cdots\sum_{z_J=m_{J-}}^{m_{J+}}\left\{1-\beta\left(z_1,\ldots,z_J;m_1,\ldots,m_J;n_1,\ldots,n_J\right)\right\}\\ &\quad\times\left\{\prod_{j=1}^{J}g_i\left(z_j|m_j,n_j\right)\right\}\left\{\prod_{j=1}^{J}h_j\left(m_j|n_j\right)\right\}l_N\left(n_1,\ldots,n_J\right).\end{aligned} \tag{2}$$

respectively, where $\mathscr{D}_N=\left\{\left(n_1,\ldots,n_J\right):0\le n_1\le N,\ldots,0\le n_J\le N,\sum_{j=1}^{J}n_j=N\right\}$ and $E^{\boldsymbol{w}}(\cdot)$ denotes the expected value with respect to a random vector $\boldsymbol{w}$.

Since $\alpha(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n})\le\alpha^*$ for all $(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n})$, we have $\alpha_N\le\alpha^*$. Given power $1-\beta^*$, the required sample size is chosen by the smallest integer *N* satisfying $1-\beta_N\ge 1-\beta^*$. In other words, while the statistical testing controls the conditional type I error $\alpha(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n})$, the sample size is determined to guarantee a specified level of marginal power. In summary, a sample size is calculated as follows.

Sample Size Calculation

- **A.** Specify input parameters: J, (α^*,β^*), $(q_1,\ldots,q_J)$, $(\theta_1,\ldots,\theta_J)$, $(a_1,\ldots,a_J)$, $(b_1,\ldots,b_J)$.
- **B.** Starting from the sample size for Mantel-Haenszel test N_{MH}, do following by increasing *N* by 1,
  - **B1** For $j=1,\ldots,J$, $z_j\in[0,n_j]$, $m_j\in[0,n_j]$, $n_j\in[0,N]$, and $\sum_{j=1}^{J}n_j=N$,
    1. Find $c_{\alpha^*}=c_{\alpha^*}(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n})$.
    2. Calculate $1-\beta(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n})=P(S\ge c_{\alpha^*}|\boldsymbol{z},\boldsymbol{m},\boldsymbol{n},H_1)$
  - **B2** Calculate $1-\beta_N=E\{1-\beta(\boldsymbol{z},\boldsymbol{m},\boldsymbol{n})|H_1\}$.
- **C.** Stop (B) if $1-\beta_N\ge 1-\beta^*$. This *N* is the required sample size.

3.2 When Stratum Allocation is Fixed

In a case-control study or a clinical trial, one may want to assign a fixed proportion of subjects to stratum *j*, say $100a_j$%, regardless of its prevalence. In this case, in calculating α_N and $1-\beta_N$, n_j are fixed at Na_j and the step to calculate the expectations with respect to $\boldsymbol{n}$ are omitted. That is, given *N*, we set $n_j=\left[Na_j\right],\ldots,n_{J-1}=\left[Na_{J-1}\right],n_J=N-\sum_{j=1}^{J-1}n_j$, and calculate (1) and (2) by

$$\alpha_{N}=\sum_{m_1=0}^{n_1}\cdots\sum_{m_J=0}^{n_J}\sum_{z_1=m_{1-}}^{m_{1+}}\cdots\sum_{z_J=m_{J-}}^{m_{J+}}\alpha\left(z_1,\ldots,z_J;m_1,\ldots,m_J;n_1\ldots,n_J\right)\times\left\{\prod_{j=1}^{J}g_{0j}\left(z_j|m_j,n_j\right)\right\}\left\{\prod_{j=1}^{J}h_j\left(m_j|n_j\right)\right\}$$

and

$$1-\beta_{N}=\sum_{m_1=0}^{n_1}\cdots\sum_{m_J=0}^{n_J}\sum_{z_1=m_{1-}}^{m_{1+}}\cdots\sum_{z_J=m_{J-}}^{m_{J+}}\left\{1-\beta\left(z_1,\ldots,z_J;m_1,\ldots,m_J;n_1\ldots,n_J\right)\right\}$$
$$\times\left\{\prod_{j=1}^{J}g_j\left(z_j|m_j,n_j\right)\right\}\left\{\prod_{j=1}^{J}h_j\left(m_j|n_j\right)\right\},$$

where [*a*] is the round-off of *a*.

3.3 When Both Group and Stratum Allocations are Fixed

In a more simplified study design, we may want to further prespecify the allocation proportion of group 1 within each stratum. In this case, given N, n_j and m_j are fixed at $[Na_j]$ and $[Na_jb_j]$, respectively, and the calculation of (1) and (2) is further simplified to

$$\alpha_{N}=\sum_{z_1=m_{1-}}^{m_{1+}}\cdots\sum_{z_J=m_{J-}}^{m_{J+}}\alpha\left(z_1,\ldots,z_J;m_1,\ldots,m_J;n_1,\ldots,n_J\right)\prod_{j=1}^{J}g_{0j}\left(z_j|m_j,n_j\right)$$

and

$$1-\beta_{N}=\sum_{z_1=m_{1-}}^{m_{1+}}\cdots\sum_{z_J=m_{J-}}^{m_{J+}}\left\{1-\beta\left(z_1,\ldots,z_J;m_1,\ldots,m_J;n_1,\ldots,n_J\right)\right\}\prod_{j=1}^{J}g_j\left(z_j|m_j,n_j\right).$$

4 Numerical Studies

We want to compare the small sample performance of stratified Fisher's test and Mantel-Haenszel test using simulations. We generate B = 10,000 simulation samples of size N = 25, 50 or 75 with J = 2 strata under a_1 = 0.25, 0.5 or 0.75; (b_1, b_2) = (1/4, 3/4), (1/2, 1/2) or (3/4, 1/4); q_1 = 0.1, q_2 = 0.3 or 0.7; (θ_1, θ_2) = (1, 1), (5, 10), (7.5, 7.5) or (10, 5). Stratified Fisher's test, the standard (unstratified) Fisher's test and Mantel-Haenszel test are applied to each simulation sample, and empirical power for each test is calculated as the proportion of simulation samples rejecting H_0 with one-sided α^* = 0.05. The exact type I error rate and power for stratified Fisher's test can be calculated by using the methods in Section 3, but through simulations we want to compare the performance of the these testing methods applied to the same data sets. We consider large odds ratios to investigate the performance of Fisher's tests and Mantel-Haenszel test with small sample sizes.

Table 3 summarizes the simulation results. Under $H_0 : \theta_1 = \theta_2 = 1$, stratified Fisher's test is conservative overall. With 10,000 simulations and α^* = 0.05, the 95% confidence limits for the empirical type I error rate are 0.05 ± 0.004. Due to the discreteness of the exact tests and the conservative control of conditional type I error at all possible outcomes, stratified Fisher's test is always conservative as expected, especially with a small sample size (N =

25). Unstratified test has a similar type I error rate to stratified Fisher's test when allocation proportions are identical between two strata (i.e. $b_1 = b_2 = 1/2$). However, if more patients are allocated in the stratum with a higher response probabilities (i.e. $b_1 = 1/4$ and $b_2 = 3/4$), then unstratified Fisher's test becomes becomes anticonservative. On the other hand, if more patients are allocated to the stratum with a smaller response probabilities (i.e., $b_1 = 3/4$ and $b_2 = 1/4$), then unstratified Fisher's test becomes very conservative. In this sense, a testing ignoring the strata can be biased unless the allocation proportions are identical across strata. With $N = 25$ or 50, Mantel-Haenszel test is anti-conservative with $q_2 = 0.7$ (i.e. when two strata have very different response rates) or with $a_1 = 0.75$ (i.e. when a small number of subjects are allocated to the stratum with large response probabilities). The anti-conservativeness diminishes as N increases, but is still of some issue with $a_1 = 0.75$ and $N = 75$.

When allocation proportions are equal across the strata, ignoring the strata results in a slight loss of statistical power. Stratified Fisher's test is less powerful than Mantel-Haenszel test, but the difference in power decreases in N. For all three testing methods, the power increases when more subjects are allocated to the stratum with the larger odds ratio, e.g. $\theta_1 < \theta_2$ and $a_1 < a_2$.

Table 4 reports sample sizes for Mantel-Haenszel test and stratified Fisher's test. Also reported are sample sizes for stratified Fisher's test by fixing (***m, n***) or only ***n*** at their expected values. The design parameters are set at one-sided $\alpha^* = 0.05$; $1 - \beta^* = 0.9$; $J = 2$ strata; $a_1 = 0.25$, 0.5 or 0.75; $(b_1, b_2) = (0.25, 0.25)$, (0.25,0.75), (0.5,0.5), (0.75,0.25) or (0.75,0.75); $(q_1, q_2) = (0.1, 0.3)$; $(\theta_1, \theta_2) = (5, 10)$, (7.5, 7.5) or (10, 5). For stratified Fisher's test, fixing (***m, n***) at their expected values reduces N, while fixing only ***n*** requires almost the same N compared to the case with random (***m, n***). The sample sizes are minimized with a balanced allocation, i.e. $b_1 = b_2 = 1/2$. We also observe that the cases of (b_1, b_2), $(1–b_1, b_2)$, $(b_1, 1–b_2)$ and $(1–b_1, 1–b_2)$ require similar sample sizes. That is, when the allocation between two groups is unbalanced, the required sample size does not much depend on whether the larger group is control or experimental across the different strata.

Under each setting, the sample size for stratified Fisher's test is about 30% larger than that of Mantel-Haenszel test. This difference results from the conservative type I error and power control of stratified Fisher's test. For example, from Table 3, with $(a_1, b_1, b_2, q_1, q_2) = (0.5, 0.25, 0.75, 0.1, 0.3)$, stratified Fisher's test controls the type I error at 0.0230 with $N = 75$ and has a power of 0.9042 at $(\theta_1, \theta_2) = (5, 10)$. Under this design setting, stratified Fisher's test requires a sample size of size of $N = 75$ with $(\alpha^*, 1 - \beta^*) = (0.05, 0.9)$ from Table 4. For Mantel-Haenszel test, the required sample size with $(\alpha^*, 1 - \beta^*) = (0.0230, 0.9042)$ under the same design setting is $N = 73$ which is close to $N = 75$ required for stratified Fisher's test. In other words, the conservativeness of the Fisher test results from the discreteness of the exact testing distributions. Mantel-Haenszel test approximates this exact distribution when the sample size is large. Crans and Schuster (2008) propose to conduct Fisher's test with a larger type I error $\alpha^* = \alpha + \epsilon$ $(\epsilon > 0)$ so that the maximal marginal type I error rate within the whole range [0, 1] of the response probability under H_0 becomes close to the intended α level.

Suppose that we want to design a study similar to that of Li et al. (1979). Since this is a balanced randomized study, we fix (n_1, n_2, n_3) at $(N/3, N/3, N/3)$ and (m_1, m_2, m_3) at $(N/6, N/6, N/6)$. We further assume that $(q_1, q_2, q_3) = (0.9, 0.75, 0.6)$, and $(\theta_1, \theta_2, \theta_3) = (1, 30, 30)$. (The estimates from Table 2 are $\hat{\theta}_1 = 0.833$ and $\hat{\theta}_2 = \hat{\theta}_3 = \infty$.) In order to control the one-sided conditional type I error at $\alpha^* = 0.1$ and the marginal power at $1 - \beta^* = 0.9$, we need $N = 83$. Under the design, this sample size provides marginal $\alpha_N = 0.0625$ and power $1 - \beta_N = 0.9087$.

5 Discussions

Numerous testing methods have been proposed to test on two binomial proportions adjusting for stratum effect based on different assumptions. For example, Cochran (1954) test assumes common odds ratios across strata and Gart (1985) assumes common relative risks. Mantel-Haenszel test makes no assumption on the parameters. These methods are based on large sample theories, so that their testing results may be distorted with a small sample size or sparse data.

In this paper, we propose to use an exact test extending Fisher's test to the analysis of many 2 × 2 tables together with its sample size calculation method. This test does not make any assumptions of large sample size or equal parameter values across strata, so that it does not require to check any assumptions before conducting a testing. The power and sample sizes are compared between the exact test and Mantel-Haenszel test using simulations and the proposed sample size formulas. While the type I error for Mantel-Haenszel test can be anti-conservative with a small sample size or sparse data, the exact test always controls the type I error below a specified level. When the effect size is so large that the required sample size is small (say, about N=70 or smaller), the exact test needs about 20% to 30% larger sample size than Mantel-Haenszel test. However, due to the small sample sizes, the increase in sample size in this case is not very large in absolute number (say, 10 to 20), so that, for robustness of the testing results, we propose to use the exact test by slightly increasing the sample size rather than obtaining a biased result by an asymptotic test.

If $J \geq 3$, the sample size calculation for stratified Fisher's test requires a long computing time. We found, in calculating the marginal type I error rate and power, that conditioning the sizes of strata $(n_1, ..., n_J)$ on their expected numbers provides very accurate sample sizes for the stratified Fisher's test even when $(n_1, ..., n_J)$ are random, while drastically saving the computing time.

REFERENCES

1. Breslow, NE.; Day, NE. The Analysis of Case-Control Studies. IARC Scientific Publications; No. 32, Lyon, France: 1980.
2. Cochran WC. Some methods of strengthening the common χ^2 tests. Biometrics. 1954; 10:417–451.
3. Crans GG, Schuster JJ. How conservative is Fisher's exact test? A quantitave evaluation of the two-sample comparative binomiial trial. Statistics in Medicine. 2008; 27:3598–3611. [PubMed: 18338319]
4. Fisher RA. The logic of inductive inference (with discussion). Journal of Royal Statistical Society. 1935; 98:39–82.
5. Gart JJ. Approximate tests and interval estimation of the common relative risk in the combination of 2 × 2 tables. Biometrika. 1985; 72:673–677.
6. Jung SH, Chow SC, Chi EM. A note on sample size calculation based on propensity analysis in nonrandomized trials. Journal of Biopharmaceutical Statistics. 2007; 17:35–41. [PubMed: 17219754]
7. Li SH, Simon RM, Gart JJ. Small sample properties of the Mantel-Haenszel test. Biometrika. 1979; 66:181–183.
8. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute. 1959; 22:719–748. [PubMed: 13655060]
9. Nam JM. Sample size determination for case-control studies and the comparison of stratified and unstratified analyses. Biometrics. 1992; 48:389–395. [PubMed: 1637968]
10. Nam JM. Power and sample size for stratified prospective studies using the score method for testing relative risk. Biometrics. 1998; 54:331–336. [PubMed: 9544526]

11. Westfall, PH.; Zaykin, DV.; Young, SS. Multiple tests for genetic effects in association studies.. In: Looney, Stephen, editor. Methods in Molecular Biology, vol. 184 Biostatistical Methods. Humana Press; Toloway, NJ: 2002. p. 143-168.

12. Woolson RF, Bean JA, Rojas PB. Sample size for case-control studies using Cochran's statistic. Biometrics. 1986; 42:927–932. [PubMed: 3814733]

13. Zelen M. The analyses of several 2×2 contingency tables. Biometrika. 1971; 58:129–137.

Table 1

Frequency data of 2 × 2 table for stratum $j(= 1, ..., J)$

	Group		
Response	**Case**	**Control**	**Total**
Yes	x_j	y_j	z_j
No	$m_j - x_j$	$\bar{m}_j - y_j$	$n_j - z_j$
Total	m_j	$\bar{m}_j$	n_j

Table 2

Response to thymosin in bronchogenic carcinoma patients (T=thymosin, P=placebo)

	Stratum 1			**Stratum 2**			**Stratum 3**		
	T	**P**		**T**	**P**		**T**	**P**	
Success	10	12	22	9	11	20	8	7	15
Failure	1	1	2	0	1	1	0	3	3
	11	13	24	9	12	21	8	10	18

Table 3

Empirical power of stratified Fisher's test/unstratified Fisher's test/Mantel-Haenszel test with one-sided α^* = 0.05, J = 2 strata, and q_1 = 0.1

a_1	(b_1, b_2)	q_2	(θ_1, θ_2) = (1, 1)	(5,10)	(7.5,7.5)	(10,5)
(a) N = 25						
.25	(1/4, 3/4)	.3	.0094/.0354/.0370	.4812/.7864/.6809	.4305/.7230/.6332	.3470/.5989/.5387
		.7	.0149/.1785/.0572	.2464/.7160/.5062	.2556/.7143/.5011	.2395/.6782/.4642
	(1/2, 1/2)	.3	.0179/.0176/.0522	.6164/.5878/.7595	.5764/.5762/.7179	.4668/.4855/.6261
		.7	.0161/.0189/.0493	.2608/.2398/.5003	.2909/.2693/.5202	.2803/.2671/.4854
	(3/4, 1/4)	.3	.0171/.0055/.0599	.4447/.2817/.6448	.4061/.3229/.6053	.3354/.3220/.5223
		.7	.0059/.0007/.0314	.0847/.0211/.2523	.1073/.0354/.2934	.1192/.0435/.3024
.5	(1/4, 3/4)	.3	.0101/.0560/.0476	.3750/.7984/.5898	.4091/.7852/.6219	.3862/.7073/.5919
		.7	.0157/.2987/.0668	.2285/.8140/.4737	.2847/.8356/.5371	.3156/.8318/.5563
	(1/2, 1/2)	.3	.0118/.0120/.0455	.4852/.4608/.6653	.5366/.5253/.6958	.5055/.5131/.6689
		.7	.0164/.0210/.0507	.2564/.2177/.4796	.3441/.2914/.5603	.3980/.3447/.6036
	(3/4, 1/4)	.3	.0130/.0032/.0542	.3260/.1740/.5389	.3627/.2698/.5717	.3467/.3266/.5583
		.7	.0052/.0001/.0255	.0951/.0146/.2684	.1547/.0308/.3673	.2019/.0511/.4286
.75	(1/4, 3/4)	.3	.0096/.0380/.0564	.2830/.6508/.5041	.3815/.6876/.5959	.4382/.6830/.6485
		.7	.0125/.2124/.0669	.2238/.6819/.4463	.3325/.7500/.5620	.4054/.7967/.6364
	(1/2, 1/2)	.3	.0104/.0106/.0425	.3572/.3681/.5478	.4739/.4814/.6543	.5553/.5570/.7143
		.7	.0097/.0166/.0478	.2514/.2274/.4532	.3981/.3510/.5950	.5014/.4460/.6863
	(3/4, 1/4)	.3	.0033/.0009/.0321	.1904/.1315/.3921	.2896/.2468/.5013	.3479/.3443/.5649
		.7	.0019/.0004/.0162	.0999/.0258/.2703	.1980/.0592/.4100	.2937/.1077/.5202
(b) N = 50						
.25	(1/4, 3/4)	.3	.0188/.0808/.0419	.8271/.9837/.9073	.7918/.9713/.8816	.6851/.9156/.7950
		.7	.0228/.3873/.0525	.5607/.9696/.7422	.5702/.9702/.7431	.5318/.9555/.7115
	(1/2, 1/2)	.3	.0287/.0272/.0540	.9343/.9174/.9657	.9064/.8958/.9463	.8343/.8353/.8983
		.7	.0219/.0278/.0460	.6610/.5045/.7930	.6856/.5609/.8058	.6426/.5581/.7657
	(3/4, 1/4)	.3	.0248/.0051/.0546	.8273/.5754/.9048	.7890/.6340/.8734	.6789/.6166/.7898
		.7	.0170/.0002/.0473	.4361/.0335/.6240	.4653/.0614/.6427	.4568/.0870/.6180
.5	(1/4, 3/4)	.3	.0192/.1345/.0522	.7341/.9864/.8457	.7670/.9848/.8652	.7416/.9670/.8373
		.7	.0229/.6002/.0586	.5288/.9872/.7099	.6113/.9911/.7734	.6658/.9930/.8047
	(1/2, 1/2)	.3	.0208/.0212/.0498	.8547/.8247/.9174	.8818/.8670/.9344	.8610/.8560/.9173
		.7	.0243/.0289/.0527	.6250/.4701/.7606	.7359/.5956/.8400	.7785/.6685/.8656
	(3/4, 1/4)	.3	.0199/.0014/.0507	.7192/.3740/.8281	.7542/.5538/.8476	.7333/.6580/.8377
		.7	.0142/.0000/.0413	.4254/.0163/.5971	.5495/.0442/.7021	.6067/.0854/.7388
.75	(1/4, 3/4)	.3	.0181/.1024/.0550	.6018/.9394/.7492	.7407/.9580/.8486	.7877/.9579/.8821
		.7	.0197/.4640/.0624	.4961/.9534/.6698	.6632/.9773/.8036	.7632/.9863/.8708
	(1/2, 1/2)	.3	.0208/.0211/.0525	.7304/.7025/.8277	.8521/.8385/.9153	.8978/.8935/.9425
		.7	.0206/.0250/.0521	.6033/.4931/.7383	.7891/.6826/.8764	.8757/.7958/.9325
	(3/4, 1/4)	.3	.0142/.0016/.0458	.5771/.3242/.7119	.7015/.5407/.8109	.7597/.6874/.8567
		.7	.0083/.0000/.0321	.4141/.0343/.5740	.6104/.1125/.7415	.7184/.2060/.8282

a **1**	**(b_1, b_2)**	***q*** **2**	**(θ_1, θ_2) = (1, 1)**	**(5,10)**	**(7.5,7.5)**	**(10,5)**
(c) *N* = 75						
.25	(1/4, 3/4)	.3	.0250/.1274/.0474	.9552/.9995/.9779	.9335/.9980/.9639	.8760/.9887/.9250
		.7	.0279/.5583/.0554	.7572/.9975/.8654	.7694/.9969/.8714	.7293/.9949/.8366
	(1/2, 1/2)	.3	.0262/.0279/.0472	.9917/.9872/.9961	.9844/.9806/.9919	.9561/.9538/.9726
		.7	.0291/.0316/.0519	.8660/.7010/.9223	.8730/.7500/.9236	.8442/.7538/.8998
	(3/4, 1/4)	.3	.0249/.0028/.0559	.9545/.7684/.9761	.9339/.8248/.9646	.8701/.8162/.9197
		.7	.0235/.0001/.0467	.7096/.0394/.8142	.7305/.0862/.8230	.7008/.1235/.7999
.5	(1/4, 3/4)	.3	.0230/.1974/.0472	.9042/.9993/.9472	.9214/.9995/.9586	.9059/.9976/.9490
		.7	.0266/.7866/.0547	.7246/.9999/.8375	.8066/.9996/.8946	.8395/.9999/.9145
	(1/2, 1/2)	.3	.0245/.0258/.0498	.9660/.9506/.9818	.9768/.9694/.9873	.9686/.9661/.9812
		.7	.0280/.0307/.0544	.8388/.6587/.9044	.9096/.7792/.9496	.9294/.8395/.9611
	(3/4, 1/4)	.3	.0234/.0013/.0487	.8984/.5448/.9407	.9153/.7408/.9519	.9025/.8300/.9408
		.7	.0212/.0000/.0477	.6646/.0130/.7738	.7788/.0473/.8618	.8244/.1121/.8891
.75	(1/4, 3/4)	.3	.0238/.1521/.0543	.8057/.9919/.8838	.9024/.9964/.9468	.9277/.9967/.9627
		.7	.0238/.6531/.0567	.6984/.9953/.8135	.8531/.9983/.9209	.9172/.9990/.9576
	(1/2, 1/2)	.3	.0236/.0237/.0491	.9092/.8914/.9491	.9672/.9591/.9837	.9830/.9786/.9907
		.7	.0249/.0317/.0527	.8070/.6657/.8778	.9352/.8521/.9631	.9751/.9312/.9873
	(3/4, 1/4)	.3	.0193/.0016/.0469	.7965/.4821/.8670	.8917/.7363/.9323	.9260/.8699/.9581
	.	.7	.0171/.0000/.0413	.6444/.0400/.7582	.8299/.1583/.8996	.9038/.3032/.9475

Table 4

Sample size for Mantel-Haenszel test/stratified Fisher's test with (***m***, ***n***) fixed/stratified Fisher's test with ***n*** fixed/stratified Fisher's test under J =2 strata, $(q_1, q_2) = (0.1, 0.3)$, one-sided $\alpha^* = 0.05$, and $1 - \beta^* = 0.9$

		(θ_1, θ_2)		
a_1	(b_1, b_2)	(5,10)	(7.5,7.5)	(10,5)
0.25	(0.25,0.25)	46/59/61/61	51/64/66/66	65/80/82/82
	(0.25,0.75)	45/53/60/61	50/62/66/66	65/79/82/82
	(0.5,0.5)	36/43/45/45	39/48/49/49	50/59/62/62
	(0.75,0.25)	46/59/61/61	51/64/66/67	65/80/83/83
	(0.75,0.75)	46/53/60/61	51/60/66/66	65/76/83/83
0.5	(0.25,0.25)	58/72/75/76	55/72/72/71	58/72/75/75
	(0.25,0.75)	58/65/75/75	54/65/70/71	58/72/75/75
	(0.5,0.5)	45/53/56/54	43/50/53/53	45/54/56/56
	(0.75,0.25)	59/72/75/76	56/66/71/71	59/72/76/78
	(0.75,0.75)	59/69/75/76	55/63/71/71	59/68/76/76
0.75	(0.25,0.25)	78/96/97/98	59/75/76/76	52/64/67/67
	(0.25,0.75)	77/89/97/97	59/69/76/76	52/64/67/67
	(0.5,0.5)	61/70/73/73	47/55/57/57	41/49/51/51
	(0.75,0.25)	80/96/98/99	61/75/77/77	53/65/69/69
	(0.75,0.75)	80/85/98/99	61/69/77/77	53/62/69/69