A Simple Mantel-Haenszel Type Test for Non-Inferiority

Kallappa M. Koti
12815 Twinbrook Pkwy, MD 20851

Abstract

Randomized trials designed to establish non-inferiority of an experimental therapy as compared to a standard one as measured by binomial proportions have been widely used. Randomization is often stratified by prognostic factors. We propose a Mantel-Haenszel type test to demonstrate non-inferiority when the non-inferiority margin is not necessarily uniform in all strata. Derivation of the new test originates from Wittes and Wallenstein (1987). We provide an easy to calculate formula for sample size. The new test may be an alternative to Yanagawa et al. (1994). A SAS code that calculates the p-value is provided.

Key Words: Hyper-geometric distribution, good outcome, model-free approach, unconditional power, historical data, W-square test.

1. Introduction

Non-inferiority and equivalence trials aim to show that the experimental therapy is not clinically worse than (non-inferiority) or clinically similar to (equivalence) an active control therapy. We focus on non-inferiority in terms of binomial proportions in the comparative binomial (CB) design setting. Statistical exercise is to test the null hypothesis of non-zero differences in population proportions. One may refer to Tunes da Silva et al. (2008) who have presented a good introduction to non-inferiority testing from clinical trials perspective in un-stratified data analysis.

In most confirmatory clinical trials, stratified randomization is adopted. That is, subjects are grouped according to covariate values prior to randomization and subjects are then randomized within strata. The Mantel-Haenszel (MH) is most commonly used for testing the classical null hypothesis that the odds ratios with each stratum are equal to 1. In this article, we state the null hypothesis in terms of the strata proportions. The null and alternative hypotheses and the corresponding MH test are described in Section 2 below.

Many practitioners use logistic regression model based Wald's $(1-2\alpha)\%$ confidence intervals to demonstrate non-inferiority of an experimental therapy compared to an active control. Miettinen and Nurminen (1985), and Farrington and Manning (1990) stress the need for the restricted maximum likelihood estimates (RMLE) based testing hypothesis of non-zero differences in proportions. Yanagawa et al. (1994) proposed MH-type test for non-inferiority analysis of binomial data from stratified randomized trials. Yanagawa et al. obtained the test statistic using the Neyman's $C(\alpha)$ test criterion (see Neyman, 1979). Further details on their test are provided in the Appendix A at the end of this article. Wittes and Wallenstein (1987) derived the unconditional power of the MH test. We propose a new simple Mantel-Haenszel type test that originates from Wittes and Wallenstein (1987). We call the new test as *W-square test*.

2. The Mantel-Haenszel Test

The Mantel-Haenszel (MH) test is widely used to analyze categorical data from comparative binomial clinical trials. We borrow the terminology and notation from Wittes and Wallenstein (1987) and briefly describe the MH test. The data consist of $T\,(\geq 2)$ sets of 2×2 contingency tables. Let n_{ijk} denote the (i, j, k) th cell size, where i represents the table ($i = 1, 2, \cdot\cdot, T$), j represents the treatment [$j=1$ (experimental); $j=2$ (control)], and k represents the outcome [$k=1$ (success); $k=2$ (failure)]. Each contingency table constitutes a stratum.

Table 1: The ith 2×2 contingency table

	Outcome:		
Treatment	*Success*	*Failure*	*Total*
1	n_{i11}	n_{i12}	$n_{i1\cdot}$
2	n_{i21}	n_{i22}	$n_{i2\cdot}$
Total	$n_{i\cdot 1}$	$n_{i\cdot 2}$	$n_{i\cdot\cdot}$

In Table 1 above, under a comparative binomial design setting $n_{i\cdot\cdot}$ is the preselected number of individuals in the i th stratum. The row totals $n_{i1\cdot}$ and $n_{i2\cdot}$ represent the numbers of subjects randomized to the experimental treatment arm and the control arm, respectively. Often in a controlled randomized trial $n_{i1\cdot} = n_{i2\cdot}$. Let $N = \sum_1^T n_{i\cdot\cdot}$ be the trial size. Let $\rho_i = n_{i1\cdot} / n_{i\cdot\cdot}$ be the proportion of subjects receiving the experimental treatment in the i th table, and $\lambda_i = n_{i\cdot\cdot} / N$ be the proportion of subjects in the i th table. Let π_{ij} ($j = 1, 2; i = 1, 2, \cdot\cdot, T$) denote the probability of success of the j th treatment in the i th table. The π_{ij} s are the population parameters of interest. Consider testing the null hypothesis of no treatment difference

$$H_0 : \pi_{i1} - \pi_{i2} = 0 \text{, for all } i = 1,2, \cdot\cdot, T \qquad (2.1)$$

against the alternative hypothesis

$$H_A : \pi_{i1} - \pi_{i2} \geq 0 \text{, for all } i \text{, and } \pi_{i1} - \pi_{i2} > 0 \text{, for some } i \qquad (2.2)$$

MH test is conditioned on all of the margins. Under H_0, the number of successes n_{i11} from the experimental treatment arm follows the hypergeometric distribution with mean

$n_{i1\cdot}n_{i\cdot1}/n_{i\cdot\cdot}$ and variance $V_i = n_{i1\cdot}\,n_{i\cdot1}\,n_{i2\cdot}\,n_{i\cdot2}/\{n_{i\cdot\cdot}^2\,(n_{i\cdot\cdot}-1)\}$. The *uncorrected* Mantel-Haenszel test statistic M_U is given by

$$M_U = \sum_1^T g_i/(\sum_1^T V_i)^{1/2}, \tag{2.3}$$

where $g_i = n_{i11} - n_{i1\cdot}n_{i\cdot1}/n_{i\cdot\cdot}$ and V_i is given above. Note that g_i can be written as

$$g_i = N\lambda_i\,\rho_i(1-\rho_i)(p_{1i}-p_{2i}),$$

where for each i, $p_{ij} = n_{ij1}/n_{ij\cdot}$. Also note that V_i can be written as

$$V_i = \; n_{i1\cdot}\,n_{i2\cdot}\,\overline{p}_i\,(1-\overline{p}_i)/(n_{i\cdot\cdot}-1)\;,$$

where $\overline{p}_i = n_{i\cdot1}/n_{i\cdot\cdot}$ is the observed proportion of successes in both treatment arms combined in the i th table. For large N, the Mantel-Haenszel test statistic M_U given by (2.3) approaches the standard normal distribution under H_0. For the one-sided alternative hypothesis H_A, the null hypothesis H_0 is rejected at level α if $M_U > z_{1-\alpha}$, where z_γ is the γ-percentile of the standard normal distribution. Cochran (1954) proposed a test statistic similar to (2.3). He treated the rows in each 2×2 table as two independent binomials rather than a hypergeometric. The statistic M_U is popularly called the *Cochran-Mantel-Haenszel* (CMH) *statistic* (see Agresti, 2002). Stokes et al. (1995) provide SAS codes for the CMH test.

3. The Wittes-Wallenstein Power

Birch (1964), Levin (1982) and Muñoz and Rosner (1984) have discussed the calculation of conditional power of M_U. Wittes and Wallenstein (1987) advise against using the conditional power in planning a study. Berger, Wittes, and Gold (1979) used Neyman's $C(\alpha)$ method to derive the unconditional power of the MH statistic- as a function of the odds ratios. Wittes and Wallenstein (1987) derived the unconditional power of the MH test for H_0 when under H_A, $\pi_{i1}-\pi_{i2}=\delta_i$, for $i = 1, 2, \cdots, T$. They derived the power of MH test in *two settings*. In the first *setting*, T is fixed, $n_{i\cdot\cdot}\to\infty$ in such a way that $\sqrt{N}\delta_i$ approaches a nonzero constant γ_i, and π_{ij} approaches the common value π_i under H_0 for $j = 1,2$. By Lindeberg's condition for the central limit theorem, $\sum_1^T g_i/\sqrt{N}$ has an asymptotic normal distribution with mean

$$\mu = \sqrt{N}\sum_1^T \lambda_i\,\rho_i(1-\rho_i)\,\delta_i$$

and variance

$$\sigma^2 = \sum_1^T \lambda_i\,\rho_i(1-\rho_i)\,\pi_i(1-\pi_i)\,.$$

As the π_is are unknown, the variance σ^2 is replaced by

$$\hat{\sigma}^2 = \sum_1^T \lambda_i \rho_i (1-\rho_i) \bar{\pi}_i (1-\bar{\pi}_i), \tag{3.1}$$

where

$$\bar{\pi}_i = \rho_i \pi_{i1} + (1-\rho_i)\pi_{i2}. \tag{3.2}$$

Wittes and Wallenstein calculate the power of an upper-tailed test:

$$P(M_U \geq z_{1-\alpha}) = \Phi\left[\sqrt{N}\left(\sum_1^T \lambda_i \rho_i (1-\rho_i)\delta_i / \hat{\sigma} - z_{1-\alpha}\right)\right], \tag{3.3}$$

where Φ is the standard normal distribution function.

In the second *setting*, $n_{i..}$ is bounded, $2 < n_{i..} < B$, and $T \to \infty$, so that consequently $N \to \infty$. In addition, it is assumed that $\sqrt{N}\delta_i$ approaches a nonzero constant γ_i. Next, they have argued that a more precise approximation would reasonably be based upon the finite sample moments $E(g_i)$, $\text{var}(g_i)$, and $E(V_i)$, rather than upon their asymptotic values. Wittes and Wallenstein (1987) derive the following finite sample moments for the comparative binomial setting.

$$E(g_i) = N\lambda_i \rho_i (1-\rho_i)\delta_i, \tag{3.4}$$

$$\text{var}(g_i) = N\lambda_i \rho_i (1-\rho_i)[(1-\rho_i)\pi_{i1}(1-\pi_{i1}) + \rho_i \pi_{i2}(1-\pi_{i2})] \tag{3.5}$$

$$E(V_i) = N\lambda_i \rho_i (1-\lambda_i)[\bar{\pi}_i(1-\bar{\pi}_i) + \delta_i^2 \rho_i (1-\rho_i)/(N\lambda_i - 1)] \tag{3.6}$$

By Lindeberg's condition for the central limit theorem, $\sum_1^T g_i / \sqrt{N}$ has an asymptotic normal distribution with mean

$$\begin{aligned} \mu &= \lim_{T\to\infty} \sum_1^T E(g_i)/\sqrt{N} \\ &= \sqrt{N} \sum_1^T \lambda_i \rho_i (1-\rho_i)\delta_i, \end{aligned} \tag{3.7}$$

and variance

$$\begin{aligned} \sigma_{CB}^2 &= \lim_{T\to\infty} \text{var}\left(\sum_1^T g_i\right)/N \\ &= \sum_1^T \lambda_i \rho_i (1-\rho_i)[(1-\rho_i)\pi_{i1}(1-\pi_{i1}) + \rho_i \pi_{i2}(1-\pi_{i2})] \end{aligned} \tag{3.8}$$

The subscript *CB* in σ_{CB}^2 of (3.8) is used to distinguish it from the σ^2 in (3.1). Next, Wittes and Wallenstein point out that when $n_{i..}$s are bounded, as $T \to \infty$ and therefore, $N \to \infty$, the events "$M_U > Z_{1-\alpha}$" and "$\sum g_i / \sqrt{N} > Z_{1-\alpha}\sqrt{W}$" are equivalent, where

$$\begin{aligned} W &= \lim_{T\to\infty} \sum_1^T E(V_i)/N \\ &= \sum_1^T \lambda_i \rho_i (1-\rho_i)[\bar{\pi}_i(1-\bar{\pi}_i) + \delta_i^2 \rho_i (1-\rho_i)/(N\lambda_i - 1)], \end{aligned}$$

and $\bar{\pi}_i$ is given by (3.2). The term $\delta_i^2 \rho_i (1-\rho_i)/(N\lambda_i - 1)$ in W above is of $O(\delta_i^2)/n_{i..}$ and is dropped in the following. That is,

$$W = \sum_1^T \lambda_i \rho_i (1-\rho_i) \bar{\pi}_i (1-\bar{\pi}_i) \tag{3.9}$$

The Wittes-Wallenstein's improved one-tailed power of MH test is

$$P(M_U > z_{1-\alpha}) = \Phi[(\mu - z_{1-\alpha} \times \sqrt{W})/\sigma_{CB}], \tag{3.10}$$

where μ, σ_{CB}^2, and W are given by (3.7), (3.8) and (3.9), respectively. The derivation of the power in (3.10) is based on more accurate approximation of the expected value and variance of the numerator and denominator of the MH test statistic. Wittes and Wallenstein claim that the new approximations for power are closer to the exact values than previously obtained formulas. We are now ready to derive the new test for establishing non-inferiority.

4. The Main Result: New Test for Non-Inferiority Demonstration

In the following we assume that higher probabilities $\{\pi_{ij}\}$ of success are preferred. Tunes da Silva et al. (2008) describe such π_{i1} and π_{i2} as the probabilities of *good* outcome. For example, complete response (CR) in an oncology trial is a good outcome. See Section 8 for an example of a bad outcome. We want to test the null hypothesis that the probability of success in the experimental treatment arm is at least an amount δ_i worse than the control arm to demonstrate that the experimental therapy is non-inferior. We assume that δ_i to be known constants and that $0 \le \delta_i < 1$ for $i = 1, 2, \cdot\cdot, T$. Our objective is to test a null hypothesis of non-zero differences in success probabilities

$$K_0 : \pi_{i1} - \pi_{i2} \le -\delta_i, \text{ for all } i = 1, 2, \cdot\cdot, T \tag{4.1}$$

against the alternative hypothesis

$$K_A : \pi_{i1} - \pi_{i2} \ge -\delta_i, \text{ for all } i, \text{ and } \pi_{i1} - \pi_{i2} > -\delta_i, \text{ for some } i \tag{4.2}$$

As we are considering an upper-tailed test of size α, the objective is to find a constant c_α such that

$$\lim_{T\to\infty} P(M_U > c_\alpha \mid K_0) = \alpha .$$

Needless to say, c_α defines the rejection region of the proposed test of K_0. From Section 3, it follows that, under the null hypothesis $K_0 : \pi_{i1} - \pi_{i2} = -\delta_i$, $\sum g_i / \sqrt{N}$ has an asymptotic normal distribution with mean μ and variance σ_{CB}^2, where

$$\mu = -\sqrt{N}\sum_1^T \lambda_i\, \rho_i(1-\rho_i)\delta_i\,, \tag{4.3}$$

and σ_{CB}^2 is given by (3.8). Following Wittes and Wallenstein, we claim that the events "$M_U > c_\alpha$" and "$\sum g_i/\sqrt{N} > c_\alpha\sqrt{W}$", where W is given in (3.9), are equivalent. Therefore, the constant c_α should be such that

$$\lim_{T\to\infty} P\big(\sum_1^T (g_i/\sqrt{N}) > c_\alpha\sqrt{W} \mid K_0\big) = \alpha.$$

That is, c_α satisfies:

$$\Phi[\frac{\sum g_i/\sqrt{N} - \mu}{\sigma_{CB}} > \frac{c_\alpha\sqrt{W} - \mu}{\sigma_{CB}}] = \alpha$$

Therefore, we set

$$\frac{c_\alpha\sqrt{W} - \mu}{\sigma_{CB}} = z_{1-\alpha}$$

It readily follows that

$$c_\alpha = (z_{1-\alpha}\,\sigma_{CB} + \mu)/\sqrt{W}\,, \tag{4.4}$$

where μ, σ_{CB} and W are given by (4.3), (3.8) and (3.9), respectively. We reject K_0 in favor of K_A at α level of significance if $M_U > (z_{1-\alpha}\,\sigma_{CB} + \mu)/\sqrt{W}$. There is one problem. The constant c_α depends on σ_{CB}^2 and W , which in turn depend on the unknown success probabilities π_{i1} and π_{i2}. However, as under K_0, $\pi_{i1} = \pi_{i2} - \delta_i$, and δ_i are pre-specified, we need to know only the success probabilities π_{i2} from the control arm in order to perform the test. There are three simple solutions to the problem: First, as π_{i2}s belong to the reference group whose efficacy has been well established, it is reasonable to assume that π_{i2} are reliably known from previous studies. Second, a simpler option is to use the sample proportions $p_{i2} = n_{i22}/n_{i2\cdot}$ from the current non-inferiority trial in place of π_{i2}. Third, one can express c_α in terms of the restricted maximum likelihood estimates (RMLE) $\tilde{\pi}_{i2}$ that are derived in Miettinen and Nurminen (1985) and referenced later in Farrington and Manning (1990) and Yanagawa et al. (1994). We discuss the RMLEs in Appendix A for the benefit of the reader.

5. Power and p-Value

We discuss the power and p-value of the new test of K_0 against K_A. By definition, the power is the probability of rejecting a false null hypothesis. We calculate the power of the test given by (4.4) when all δ_i $(i = 1, 2, \cdots, T)$ are set equal to 0. The power of the test is

$$\psi = P(M_U > c_\alpha \mid \delta_i = 0 \text{ for all } i)$$

As seen from Section 2, the test statistic M_U has standard normal distribution when all δ_i s are equal to 0. Therefore, the power of the test is given by

$$\psi = 1 - \Phi(c_\alpha), \tag{5.1}$$

where c_α is given by (4.4).

We next derive the p-value. By definition, the p-value is

$$\begin{aligned} \mathrm{p} &= P(M_U > m_u | K_0) \\ &= \lim\nolimits_{T\to\infty} P(\textstyle\sum_1^T g_i / \sqrt{N} > m_u \sqrt{W}) \end{aligned}$$

where m_u denotes an observed value of M_U. It follows that

$$\mathrm{p} \ = 1 - \Phi[(m_u \sqrt{W} - \mu) / \sigma_{CB}], \tag{5.2}$$

where μ, W and σ_{CB} are given by (4.3), (3.9) and (3.8), respectively.

6. Sample Size Calculation

We discuss sample size determination for testing K_0 against K_A. The type 2 error rate β is the probability of not rejecting the null hypothesis K_0 when it is false. From (5.1), it follows that $\beta = \Phi(c_\alpha)$. We find the sample size by setting $c_\alpha = z_\beta$, where c_α is given by (4.4). That is, we set

$$z_{1-\alpha}\, \sigma_{CB} + \mu = z_\beta \sqrt{W} \ ,$$

where μ, σ_{CB}^2 and W are given by (3.7), (3.8) and (3.9), respectively. This gives

$$N = (z_\beta \sqrt{W} - z_{1-\alpha}\, \sigma_{CB})^2 / [\textstyle\sum_1^T \lambda_i\, \rho_i (1-\rho_i) \delta_i\,]^2 , \tag{6.1}$$

where all δ_i s are pre-specified and it is assumed that the success probabilities π_{i2} of the reference therapy are made available from previous studies. We provide some sample sizes for test with $\alpha = 0.05$ and $\beta = 0.2$ in Table 2 below.

Table 2: Sample sizes from (6.1)

Trial Parameters	δ_i, $i = 1, 2, \cdots, T$	N
$T = 3$, $\rho_i = 0.67$, $\lambda_i = 0.4, 0.3, 0.3$	0.1, 0.1, 0.1	673
π_{i2} = 0.5, 0.6, 0.7	0.15, 0.15, 0.15	299
$T = 3$, $\rho_i = 0.67$, $\lambda_i = 0.4, 0.3, 0.3$	0.1, 0.1, 0.1	581
π_{i2} = 0.7, 0.75, 0.8	0.15, 0.15, 0.15	266
	0.2, 0.2, 0.2	153

Suppose that both treatment arms are to be allocated with equal number of subjects in each stratum and all strata are to be of equal sizes. That is, $\rho_i = 1/2$ and $\lambda_i = 1/T$ for all $i = 1, 2, \cdots, T$. We have

$$\sigma_{CB}^2 = \sum_1^T [\pi_{i1}(1-\pi_{i1}) + \pi_{i2}(1-\pi_{i2})]/(8T),$$

and

$$W = \sum_1^T \bar{\pi}_i (1-\bar{\pi}_i)/(4T).$$

Substituting these in (6.1), the required sample size is

$$N = \frac{2T}{(\sum_1^T \delta_i)^2} [z_\beta \sqrt{\sum_1^T 2\bar{\pi}_i (1-\bar{\pi}_i)} - z_{1-\alpha} \sqrt{\sum_1^T \{\pi_{i1}(1-\pi_{i1}) + \pi_{i2}(1-\pi_{i2})\}}\,]^2 \,, \quad (6.2)$$

where $\pi_{i1} = \pi_{i2} - \delta_i$ and $\bar{\pi}_i = (2\pi_{i2} - \delta_i)/2$ with $\bar{\pi}_i < 1$.

7. Simulation

In this section, we provide some simulation results that compare the performance of the new test with Yanagawa et al. (1994).

First, we consider a case with $T = 2$. Let $\delta_1 = 0.15$, and $\delta_2 = 0.13$. We intend to test the null hypothesis

$$K_0 : \pi_{11} - \pi_{12} \le -0.15, \text{ and } \pi_{21} - \pi_{22} \le -0.13$$

against the alternative hypothesis K_A that K_0 is false. We arbitrarily assume that $\pi_{12} = 0.6$, $\pi_{22} = 0.5$. We set $n_{1j.} = 15, 15$, and $n_{2j.} = 15, 15$ for $j = 1,2$. This means that the table totals $n_{1..} = n_{2..} = 30$, and the trial total $N = 60$. This also means that $\lambda_1 = \lambda_2 = 0.5$, and $\rho_1 = \rho_2 = 0.5$. We label this scenario as Case 1 in Table 3 below. We choose π_{i1} such that simulated data may not possibly reject the null hypothesis K_0 in favor of K_A. We conjecture that, for example, the data simulated with $\pi_{i1} = \pi_{i2} - \delta_i - 0.1$, should induce higher p-value and should not reject the null hypothesis K_0. We call the SAS subroutine *streaminit* (132435), and use:

$$n_{ijk} = \text{rand}(\text{"binomial"}, \pi_{ij}, n_{ij.})$$

to generate the data for the i th ($i = 1, 2$) contingency table. One may refer to Wicklin (2013) for further details on simulation using SAS. We calculate the p-value given by (5.2). We repeat the whole thing 10,000 times. That is, we generate 10,000 p-values. We calculate the average p-value ($\bar{p}$). We also calculate the average p-value for the

Yanagawa's test, which is described in Appendix A below. In generating Yanagawa's p-values, we noticed that RMLE $\tilde{\pi}_{i2}$ of π_{i2} did not exist in a small number of simulations. The results are shown in Table 3 below.

Table 3: Simulated average p-values

	Average p-value $\bar{p}$:	
Trial Parameters	*W-square test*	*Yanagawa's test*
Case 1: $T=2,\ \pi_{12}=0.6,\ \ \pi_{22}=0.5,\ n_{11.}=15,\ n_{12.}=15,\ N=60$ $\delta_1=\ 0.15\ \ \delta_2=\ 0.13,\ \pi_{i1}=\pi_{i2}-\delta_i-0.1$	0.7153	0.3101
Case 2: $T=2,\ \pi_{12}=0.6,\ \ \pi_{22}=0.5,\ n_{11.}=15,\ n_{12.}=15\ ,\ N=60$ $\delta_1=\ 0.15\ \ \delta_2=\ 0.13,\ \pi_{i1}=\pi_{i2}-\delta_i-0.05$	0.6098	0.1726
Case 3: $T=3,\ \pi_{12}=0.7,\ \pi_{22}=0.6,\ \pi_{32}=0.5,$ $n_{11.}=15,\ n_{12.}=15,\ n_{21.}=15,\ n_{22.}=15,$ $n_{31.}=15,\ n_{32.}=15\ ;\ N\ =\ 90$ $\delta_1=\ 0.12,\ \delta_2=0.11,\ \delta_3=0.10;\ \pi_{i1}\ =\pi_{i2}-\delta_i-0.1$	0.7501	0.2323
Case 4: $T=3,\ \pi_{12}=0.7,\ \pi_{22}=0.6,\ \pi_{32}=0.5,$ $n_{11.}=15,\ n_{12.}=15,\ n_{21.}=15,\ n_{22.}=15,$ $n_{31.}=15,\ n_{32.}=15\ ;\ N\ =\ 90$ $\delta_1=\ 0.12,\ \delta_2=0.11,\ \delta_3=0.10;\ \pi_{i1}\ =\pi_{i2}-\delta_i-0.05$	0.6318	0.104

8. Concluding Remarks

For example, Graft-versus-host disease (GVHD) burden in a study to compare peripheral blood cells (PBSC) versus bone narrow (BM) as a cell source is a bad outcome (Tunes da Silva et al., 2008). If π_{i1} and π_{i2} are the success probabilities of *bad* outcome, then smaller $\{\pi_{ij}\}$ are preferred. We want to consider the null hypothesis $K_0'\colon \pi_{i1}-\pi_{i2}\geq\delta_i$ for all i against the alternative hypothesis: $K_A'\colon \pi_{i1}-\pi_{i2}\leq\delta_i$, for all i, and $\pi_{i1}-\pi_{i2}<\delta_i$, for some i, where $0\leq\delta_i<1$. We reject K_0' in favor of K_A' at α level of significance if $M_U<(z_\alpha\,\sigma_{CB}+\mu)/\sqrt{W}$. We leave further details to the reader.

If the number of strata T is small and the strata totals $n_{i..}$s are large, a test based on the Wittes-Wallenstein power given in (3.3) above is easily derived. Note that the power approximation in (3.3) is a simplification of the approximation in (3.10). The critical region for testing K_0 against K_A is obtained by replacing σ_{CB} in (4.4) by $\hat{\sigma}$ of (3.1). We leave further details to the reader.

The proposed W-square test is very easy to implement. It can be performed with a desk calculator. However, we provide a SAS code in Appendix B, which refreshes all the steps involved in the test. The SAS code analyzes the Neriproct suppository data from Table 2 of Section 4 in Yanagawa et al. (1994). In this code, $T = 3$. Input variables are self-explanatory. The p-value of the MH test was 0.19. Yanagawa et al. have used the non-inferiority margins- $\delta_1 = \delta_2 = \delta_3 = 0.05$. We calculated the restricted maximum likelihood estimates: $\tilde{\pi}_{2\,1} = 0.5394$, $\tilde{\pi}_{2\,2} = 0.5588$ and $\tilde{\pi}_{2\,3} = 0.4869$. The Yanagawa's test has a p-value of 0.021. They infer that Neriproct suppository is at least as effective as Tribenoside. We have used the sample proportions from the control arm to calculate σ_{CB} and W. The SAS code calculates the MH statistic m_U, the critical region c_α, the p-value, and the power of the W-square test. We noted that $c_\alpha = 0.8726$ and $m_U = 1.3102$. We get a p-value of 0.0186 for the W-square test.

The W-square test is an unconditional test. It is conservative compared to the Yanagawa et al. test in that the W-square test, in general, yields higher p-values. The W-square test does not need each stratum total $n_{i..}$ to be large.

Appendix A

Here we provide formulas for the restricted maximum likelihood estimator $\tilde{\pi}_2$ of π_2 that are provided in Farrington and Manning (1990). It is assumed that n_{i11} and n_{i21} follow binomial distributions. Likelihood of n_{i11} and n_{i21} from the ith table is expressed as a function of π_{i2}, where $\pi_{i1} = \pi_{i2} - \delta_i$. We state them without the subscript i. Note that $\tilde{\pi}_2$ is the unique solution in $(\delta, 1)$ of the maximum likelihood equation:

$$a\,x^3 + b\,x^2 + cx + d = 0$$

with

$$a = 1 + n_{1.}/n_{2.},$$
$$b = -[1 + (n_{1.}/n_{2.}) + p_2 + (n_{1.}/n_{2.})p_1 + \delta(n_{1.}/n_{2.} + 2)]\,,$$
$$c = \delta^2 + \delta(2p_2 + n_{1.}/n_{2.} + 1) + p_2 + (n_{1.}/n_{2.})p_1,$$
$$d = -\,p_2\delta(1+\delta)\,,$$

where p_1 and p_2 are the sample proportions. The solution is

$$\tilde{\pi}_2 = 2u\cos(w) - b/3a\,,\ \tilde{\pi}_1 = \tilde{\pi}_2 - \delta$$

where

$$w = [\pi + \cos^{-1}(v/u^3)]/3,$$
$$v = b^3/(3a)^3 - bc/6a^2 + d/2a,$$
$$u = sign(v)[(b^2/(3a)^2 - c/3a]^{1/2}.$$

Yanagawa et al. (1994) state that $\tilde{\pi}_2 \in (\delta, 1)$ is $\sqrt{n}$ -consistent.

The test statistic in Yanagawa et al. is given by

$$Z_{diff} = \sum\nolimits_1^T [n_{i\,11} - n_{i1.}(\tilde{\pi}_{i2} - \delta_i)] \times \{\sum\nolimits_1^T \frac{n_{i1.}\, n_{i2.}(\tilde{\pi}_{i2} - \delta_i)^2\,(1 - \tilde{\pi}_{i2} + \delta_i)^2}{n_{i1.}\,\tilde{\pi}_{i2}\,(1 - \tilde{\pi}_{i2}) + n_{i2.}(\tilde{\pi}_{i2} - \delta_i)(1 - \tilde{\pi}_{i2} + \delta_i)}\}^{-1/2}$$

Asymptotically, it has a standard normal distribution under K_0.

Appendix B

```
data Yanagawa  ;
input  Table  ni11  ni1dot  ni21  ni2dot  nidotdot  deltai ;
lines ;
   1   13   23   15   29   52    0.05
   2   30   50   27   45   95    0.05
   3   19   38   8    31   69    0.05
 ;
data Tango ; set Yanagawa ;
N = 216 ;   * N is the total of input variable nidotdot- in column 6   ;
pi2 = ni21/ni2dot  ;
pi1 = pi2-deltai ;
lambdai = nidotdot/N  ;
rhoi = ni1dot/nidotdot ;
pibar = rhoi*pi1+(1-rhoi)*pi2 ;
egi = - sqrt(N)*lambdai*rhoi*(1-rhoi)*deltai ;
evi = lambdai*rhoi*(1-rhoi)*(pibar*(1-pibar) +
        deltai*deltai*rhoi*(1-rhoi)/(nidotdot-1)) ;
sigma2 = lambdai*rhoi*(1-rhoi)*((1-rhoi)*pi1*(1-pi1)+rhoi*pi2*(1-pi2)) ;
nidot1 = ni11+ ni21 ;
nidot2 = (ni1dot-ni11)+(ni2dot-ni21) ;
gi = ni11-(ni1dot*nidot1/nidotdot) ;
Vi = ni1dot*nidot1*ni2dot*nidot2/(nidotdot*nidotdot*(nidotdot-1))  ;
run ;

PROC IML ;
use Tango ;
read all var{egi} into c11 ;  read all var{vargi} into c12 ;
read all var{evi} into c13 ;  read all var{sigma2} into c14 ;
read all var{gi} into c21 ;    read all var{Vi} into c22 ;
Mu = sum(c11) ;
sigd2 = sum (c12) ;
W = sum(c13);
```

```
sig2CB = sum(c14) ;
calpha = (Mu+1.645*sqrt(sig2CB))/sqrt(W) ;
power = 1-PROBNORM(calpha) ;
sumgi = sum(c21) ; sumVi = sum(c22) ;
MHmu = sumgi/sqrt(sumVi) ;
pvalue = 1-PROBNORM((MHmu*sqrt(W)-Mu)/sqrt(sig2CB)) ;
print  MHmu  calpha  pvalue  power ;
run ;
```

Acknowledgements

This article reflects the views of the author and should not be construed to represent FDA's views or policies.

References

Agresti, A. 2002. *Categorical data analysis.* John Wiley & Sons, Inc., Hoboken, NJ.

Berger, A., Wittes, J. T., and Gold, R. Z. (1979). On the power of the Cochran-Mantel-Haenszel test and other approximately optimal tests for partial association. Technical Report B-03. Columbia University school of Public Health, Division of Biostatistics.

Birch, M. W. 1964. The detection of partial association, I: The 2×2 Case. *Journal of the Royal Statistical Society*, Ser. B, 26, 313-324.

Cochran, W. G. 1954. Some methods of strengthening the common χ^2 tests. *Biometrics* 10: 417-451.

Farrington, P. C. and Manning, G. 1990. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9: 1447-1454.

Levin, B. 1982. On the accuracy of the normal approximation to the power of the Mantel-Haenszel procedure. *Journal of Statistics and Computer Simulations*, 14, 201-218.

Muñoz, A. and Rosner, B. 1984. Power and sample size for a collection of 2×2 Tables. *Biometrics,* 40, 995-1004.

Miettinen, O. and Nurminen, M. 1985. Comparative analysis of two rates. *Statistics in Medicine,* 4: 213-226.

Neyman, J. 1979. $C(\alpha)$ tests and their use. *Sankhya: The Indian Journal of Statistics,* Volume 41, Series A, pp. 1-21.

Stokes, M. E., Davis, C. S. and Koch G. G. 1995. *Categorical data analysis using SAS System.* SAS Institute Inc.

Tunes da Silva, G. T., Logan, B. R., and Klein, J. P. 2008. Methods for equivalence and non- inferiority testing. *Biol Blood Marrow Transplant* 15(1): 120-127.

Wicklin, R. 2013. Simulating data with SAS®, SAS Institute Inc.

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

Yanagawa, T., Tango, T., Hiejima, Y. 1994. Mantel-Haenszel-type testing equivalence or more than equivalence in comparative clinical trials. *Biometrics* 50: 859-864.