Theory and Practice of Equivalence and Non-Inferiority Analyses

Kallappa M. Koti
Food and Drug Administration (Retired)
Kallappa.koti@gmail.com

Abstract

In this presentation, we point out some instances where statistical methods, which are advocated in some publications, are inferior to those which are thought to be technically more accurate and appropriate. Here are two examples. (i) The delta method based confidence interval approach for demonstrating non-inferiority in terms of the ratio of means is inaccurate and inappropriate. The delta-method contradicts theoretical results. Confidence interval is computed using a very crude variance estimate. Fieller-Hinkley distribution based analysis is a better option. (ii) The phrase "Mantel-Haenszel average risk difference (MHARD)" is meaningless in a non-inferiority setting. The phrase is a name for a clumsy estimand, which is data-dependent and leads to several shortcomings in the non-inferiority analysis. The MH type test by Yanagawa et al. is a valid approach and accommodates varying non-inferiority margins. The W-square test, which is a recently proposed MH type test, provides asymptotic unconditional inference. We provide further details and some related references.

Key Words: Statistical practice, Taylor series expansion, biased estimate, Fieller's theorem, 2×2 contingency tables, hypergeometric distribution, CMH test.

1. Introduction

The null hypothesis in a non-inferiority study states that the endpoint for the experimental treatment is worse than that for the positive control treatment by a pre-specified margin, and rejection of the null hypothesis at a pre-specified level of significance is used to support a claim that permits a conclusion of non-inferiority (Mauri and D'Agostino, 2017). D'Agostino Sr. et al. (2003) explain how to write the standard null and alternative hypotheses for proving non-inferiority. Let T and 'Test' represent the value of the efficacy variable for the new (experimental) treatment. Similarly let C and 'Control' represent the value of the efficacy variable for the active control. The null and alternative hypotheses are

$$\begin{aligned} &H_0\text{: } C - T \geq M \text{ (C is superior to T)} \\ &H_1\text{: } C - T < M \text{ (T is not inferior to C)} \end{aligned} \tag{1.1}$$

Here C and T are some parameters; $C - T$ is a parametric function, and M is the non-inferiority margin. The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgement. We don't discuss how to find M. Rejection of the null hypothesis is needed to conclude non-inferiority. In order to assess if non-inferiority is met one can perform a one-sided hypothesis test at α level of significance. Equivalently, we can compute a $100(1-2\alpha)$ per cent two-sided confidence interval (CI) for the difference $(C-T)$. If the confidence interval's upper bound is less than M, then with $100(1-2\alpha)$ per cent confidence, it is concluded that the active control is more efficacious than the experimental product by no more than M, hence one can claim non-inferiority of the experimental product as compared to the active control at an α level of significance (D'Agostino Sr. et al., 2003).

The efficacy analysis of non-inferiority trials is not simple and straight forward. The parametric function to be used in the hypotheses need not be a difference. It may be a ratio. It is argued that the ratio-based evaluation of equivalence and non-inferiority often reflects clinical rationale. However, it is perceived that assessing non-inferiority and equivalence when defined in terms of the ratio of parameters is more difficult than the problem defined in terms of the difference of parameters. Reference is made to Hauschke and Hothorn (2006), and Berger and Hsu (1996) for further details. We have two objectives in this article. Our first objective is to discuss non-inferiority evaluation in terms of the ratio of means when the efficacy endpoint is a continuous variable. Specifically, we conclude that the delta method based confidence interval approach for demonstrating non-inferiority in terms of the ratio of means, which is proposed in Rothmann et al. (2012), is unscientific and inappropriate.

Stratification is often carried out in randomized clinical trials. It is an important method used to adjust for prognostic factors. Mantel-Haenszel (1959) proposed a model-free test to compare binomial proportions between treatments in superiority trials when randomization is stratified. Non-inferiority testing also arises in stratified analysis of binary data. Yanagawa et al. (1994) proposed a Mantel-Haenszel-type statistic for testing whether a new treatment is at least as effective as the standard treatment in comparative binomial trials. Our second objective is to discuss non-inferiority evaluation when the efficacy endpoint is a binary outcome and randomization is stratified. In particular, we argue that the phrase "Mantel-Haenszel average risk difference", which is advocated in Rothmann et al. (2012), is meaningless and the associated analysis is inferior in proving non-inferiority.

Several authors have written on the two topics that are discussed and criticized in this article. We have cited Rothmann et al. (2012) for discussion. We sincerely look for opposing arguments on our views from practising statisticians.

2. Ratio of the sample means

In Section 12.3, Rothmann et al. (2012) consider the situation where larger outcomes are more desired than smaller outcomes. The objective is to test the null hypothesis H_0 versus the alternative H_A stated as follows.

$$H_0 : \mu_E/\mu_C \leq \delta \text{ vs. } H_A : \mu_E/\mu_C > \delta, \;\; \mu_E, \; \mu_C > 0, \text{and } 0 < \delta < 1. \tag{2.1}$$

A parallel group design setting is assumed. In the following, $X_1, X_2, \cdot\cdot, X_{n_C}$ and $Y_1, Y_2, \cdot\cdot, Y_{n_E}$ denote independent random samples from the control arm and the experimental arm, respectively. In the following, μ_C and μ_E denote the population means of the control arm and the experimental arm, respectively. The population variances are denoted by σ_C^2 and σ_E^2, respectively. As usual, $\overline{X}$ and $\overline{Y}$ denote the sample means. And S_E^2 and S_C^2 denote the sample variances. No distributional assumptions on Xs and Ys are made.

2.1 The delta method and the confidence interval

To discuss asymptotic normality of the ratio of sample means, Govindarajulu (1988) assumed that $\overline{X}$ and $\overline{Y}$ are statistically independent, and noted that

$$\frac{\overline{Y}}{\overline{X}} - \frac{\mu_E}{\mu_C} = \frac{\overline{Y}\,\mu_C - \overline{X}\,\mu_E}{\overline{X}\,\mu_C} = \frac{\mu_C\,(\overline{Y} - \mu_E) - \mu_E\,(\overline{X} - \mu_C)}{\overline{X}\,\mu_C},$$

and used the approximation:

$$\sqrt{n_E}\,\left(\frac{\overline{Y}}{\overline{X}} - \frac{\mu_E}{\mu_C}\right) \approx \sqrt{n_E}\left(\frac{\overline{Y} - \mu_E}{\mu_C}\right) - \sqrt{n_E}\,\frac{\mu_E}{{\mu_C}^2}\,(\overline{X} - \mu_C\,)\,. \tag{2.2}$$

Note that the expectation of the right hand side of (2.2) is 0. The variance of the right hand side of (2.2), which is an approximation variance of $\overline{Y}/\overline{X}$, is

$$var\left(\frac{\overline{Y}}{\overline{X}}\right) \cong \frac{\sigma_E^2}{\mu_C^2} + \frac{n_E}{n_C}\frac{\mu_E^2\sigma_C^2}{\mu_C^4} = (\frac{\mu_E}{\mu_C})^2\left[\frac{\sigma_E^2}{\mu_E^2} + \frac{n_E}{n_C}\frac{\sigma_C^2}{\mu_C^2}\right], \mu_C \neq 0$$

Then they claimed that

$$\sqrt{n_E}\,\left(\frac{\overline{Y}}{\overline{X}} - \frac{\mu_E}{\mu_C}\right) \sim AN\left(0\,,\left(\frac{\mu_E}{\mu_C}\right)^2\left[\frac{\sigma_E^2}{\mu_E^2} + \frac{n_E}{n_C}\frac{\sigma_C^2}{\mu_C^2}\right]\right)$$

Rothmann et al. (2012) have the same concept. They state:

$$\sqrt{n}\,\left(\frac{\overline{Y}}{\overline{X}} - \frac{\mu_E}{\mu_C}\right) \xrightarrow{d} N\left(0,\ \frac{\sigma_E^2/n_E}{\mu_C^2} + \frac{\mu_E^2\,\sigma_C^2/n_C}{\mu_C^4}\right)$$

Note that

$$var\left(\frac{\overline{Y}}{\overline{X}}\right) \cong \left(\frac{\mu_E}{\mu_C}\right)^2\left(\frac{\sigma_E^2/n_E}{\mu_E^2} + \frac{\sigma_C^2/n_C}{\mu_C^2}\right) \tag{2.3}$$

The unrestricted delta-method estimator of the variance of the ratio of sample means $\overline{Y}/\overline{X}$, which is used in Rothmann et al. (2012), is

$$\widehat{var}\left(\frac{\overline{Y}}{\overline{X}}\right) = \left(\frac{\overline{Y}}{\overline{X}}\right)^2\left(\frac{S_E^2/n_E}{\overline{Y}^2} + \frac{S_C^2/n_C}{\overline{X}^2}\right) \tag{2.4}$$

They use it to define the test statistic W_R (subscript R stands for Rothmann et al.):

$$W_R = \left(\frac{\overline{Y}}{\overline{X}} - \delta\right)\Big/\sqrt{\left(\frac{\overline{Y}}{\overline{X}}\right)^2\left(\frac{S_E^2/n_E}{\overline{Y}^2} + \frac{S_C^2/n_C}{\overline{X}^2}\right)} \tag{2.5}$$

The delta-method approximate $100(1-\alpha)\%$ confidence interval for the ratio μ_E/μ_C of population means, as given in Rothmann et al. (2012, p. 338) is as follows.

$$\overline{y}/\overline{x} \pm z_{\propto/2}\sqrt{(\overline{y}/\overline{x})^2\,(\frac{S_E^2/n_E}{\overline{y}^2}+\frac{S_C^2/n_C}{\overline{x}^2})} \qquad (2.6)$$

They intend to reject the null hypothesis H_0 in (2.1) if the lower confidence limit exceeds δ. Rothmann et al. (2012) have also stated that the test statistic in (2.5) can be modified to use an estimator of the standard error that is restricted to $\mu_E/\mu_C=\delta$.

2.2 Author's comments on the delta method approach

Rothmann et al. (2012, p. 334) state: The advantage of using $\mu_E-\delta\mu_C$ in practice is that smaller sample sizes should be needed for $\overline{Y}-\delta\overline{X}$ to be approximately normally distributed than the required sample sizes needed for $\overline{Y}/\overline{X}$ to be approximately normally distributed. This statement is subject to criticism: If X_i and Y_i are normally distributed, the statistic $\overline{Y}-\delta\overline{X}$ is normally distributed when sample sizes are as small as 2 in each arm. If X_i and Y_i are not normally distributed, one needs samples of sizes of 30 or more in each arm for $\overline{Y}-\delta\overline{X}$ to follow approximate normal distribution. On the other hand, the ratio $\overline{Y}/\overline{X}$ is not normally distributed, no matter what sample sizes you choose.

Readers may refer to Hauschke et al. (1999) who studied sample size determination for proving equivalence based on the ratio of two means for normally distributed data.

When a statistician is asked about the distribution of a ratio, the first thing he or she thinks about is the following. "If $X_1$and X_2 are i.i.d. from $N(0,1)$, then the ratio $U=X_1/X_2$ is Cauchy with probability density function:

$$f(u)=\pi^{-1}/(1+u^2),\ u\in R.$$

The mean of U, i.e., $E(U)$ does not exist." It does not need a reference.

As stated in Section 3 below, Hinkley (1969) derived the exact distribution function of the ratio of two random variables- having a bivariate normal distribution. He also derived an approximation to the distribution function of the ratio when the denominator is a positive valued random variable. These distribution functions are given in the next section in (3.1) and (3.2), respectively. Clearly, the ratio W in (3.1) or in (3.2) in Section 3 below is not normally distributed.

When the exact distribution of a statistic is given, why anyone wants a delta method based approximation? Knowing that the ratio of sample means is not normally distributed, it is irrational and awkward to look for a normal approximation. Delta-method is unnecessary, inaccurate, and contradicts theoretical results in this case.

The sample ratio $\overline{Y}/\overline{X}$ is not an unbiased estimator of the ratio μ_E/μ_C of the population means. In Fieller's theorem, the ratio $\overline{y}/\overline{x}$ of sample means is considered as a point estimate of the ratio μ_E/μ_C of population means.

Several comments are in order: One may recall that $E\left(\overline{Y}^2\right) = var(\overline{Y}) + \mu_E^2$. Therefore, replacing μ_E^2 in the unrestricted variance estimate in (2.5) by $\overline{Y}^2$ is not a thoughtful step. Similar comment applies for substituting $\overline{X}^2$ in place of μ_C^2 . We would not assume that variance of $\overline{X}$ is zero or close to zero. The following expected value of the ratio of the sample means given in (2.7) is a second order Taylor series expansion approximation. It is an improved version. One may refer to Govindarajulu (1988) and in some websites including www.stat.cmu.edu/~hseltman/files/ratio.pdf.

$$E\left(\frac{\overline{Y}}{\overline{X}}\right) \cong \frac{\mu_E}{\mu_C}\left(1 + \frac{\sigma_C^2}{n_C\,\mu_C^2}\right) \tag{2.7}$$

The website by Seltman also contains approximate expected value and variance of the ratio $\overline{Y}/\overline{X}$ when $\overline{Y}$ and $\overline{X}$ are not independent.

Rothmann et al. conveniently ignore the improved expected value given by (2.7) in defining the test statistic W_R of (2.5). Substitution of $(\overline{Y}/\overline{X})^2$ in place of $(\mu_E/\mu_C)^2$ makes the unrestricted variance estimator given by (2.5) unscientific and unreliable. Rothmann et al. (2012, p. 337-338) state: "Alternatively, the denominator in the test statistic W_R can be modified to use an estimator of the standard error that is restricted to $\mu_E/\mu_C = \delta$." No, it does not help. The variance of $\overline{Y}/\overline{X}$ still contains μ_E^2 and μ_C^2, and the denominator of W_R still ends up with $\overline{Y}^2$ and $\overline{X}^2$. Either way, they hit a dead end.

The test statistic W_R of (2.5) is a complicated function of $\overline{Y}, \overline{X}, S_E^2$, and S_C^2. A question: is $W_R \sim N(0,1)$? The denominator of the right hand side of W_R in (2.5) contains, for example, S_E^2 and $\overline{Y}^2$. We point out that, for example, if Ys are normally distributed, $(n_E - 1)S_E^2/\sigma_E^2 \sim \chi^2(n_E - 1)$, and $n_E\overline{Y}^2/\sigma_E^2 \sim \chi^2(1,\tau)$, where $\tau = \mu_E^2/\sigma_C^2$ is the noncentrality parameter. Rothmann et al. (2012) did not provide proper justification for $W_R \sim N(0,1)$. The W_R based p-value has no value in practice.

3. Fieller-Hinkley distribution based analysis

Fieller (1932) derived the probability distribution function of the ratio of two correlated normal random variables with nonzero means. Hinkley (1969) examined the exact distribution of the ratio and the standard approximation based on assuming the denominator is a positive valued random variable. We state it as it is in Hinkley (1969). Let $X_1$and X_2 be normally distributed random variables with means θ_i and σ_i^2 $(i = 1,2)$ and correlation coefficient ρ, and let $W = X_1/X_2$. The exact distribution of W and the standard approximation based on assuming $X_2 > 0$ are examined in some detail. The exact cumulative distribution function $F(w)$ of W is:

$$F(w) = L\left\{\frac{\theta_1-\theta_2 w}{\sigma_1\sigma_2\, a(w)}, -\frac{\theta_2}{\sigma_2}, \frac{\sigma_2\, w-\rho\sigma_1}{\sigma_1\sigma_2\, a(w)}\right\} + L\left\{\frac{\theta_2 w-\theta_1}{\sigma_1\sigma_2\, a(w)}, \frac{\theta_2}{\sigma_2}, \frac{\sigma_2\, w-\rho\sigma_1}{\sigma_1\sigma_2\, a(w)}\right\}, \tag{3.1}$$

where

$$a(w) = (\ w^2/\sigma_1^2 - 2\rho w/\sigma_1\sigma_2\ + 1/\sigma_2^2\ )^{1/2},$$

and

$$L(h,k;\gamma) = \frac{1}{2\pi\sqrt{1-\gamma^2}} \int_h^\infty \int_k^\infty exp\left\{-\frac{x^2-2\gamma xy+y^2}{2(1-\gamma^2)}\right\} dx\, dy$$

is the standard bivariate normal integral. The distribution function $F(w)$ is easy to calculate (see Koti, 2007). The Fieller's pdf whose distribution function is given by (3.1) is not necessarily unimodal.

In addition, Hinkley (1969) has shown that as $\theta_2/\sigma_2 \rightarrow \infty$, i.e. as $P(X_2 > 0) \rightarrow 1$,

$$F(w) \rightarrow G(w) = \Phi\left[\frac{\theta_2 w - \theta_1}{\sigma_1\sigma_2\ a(w)}\right] \tag{3.2}$$

We apply Hinkley's approximation (3.2) for $W = \overline{Y}/\overline{X}$. We go back to the notation of Section 2. We note that $E(\overline{Y}) = \mu_E$, $var(\overline{Y}) = \sigma_1^2 = \sigma_E^2/n_E$, $E(\overline{X}) = \mu_C$, and $var(\overline{X}) = \sigma_2^2 = \sigma_C^2/n_C$. To be consistent with, Rothmann et al., we consider a parallel group design and set the correlation coefficient ρ equal to 0.We also assume that $\mu_C/\sigma_C \rightarrow \infty$, i.e., $P(\overline{X} > 0) \rightarrow 1$. The distribution function G of $W = \overline{Y}/\overline{X}$ is given by

$$G(w) = \Phi\left[\frac{\mu_C\ (w-\delta)}{\sigma_1\ \sigma_2\ a(w)}\right], \tag{3.3}$$

where $\delta = \mu_E/\mu_C$, and $a(w)$ is given above in (3.1) with modified subscripts. Let $\hat\sigma_1^2 = s_1^2/n_1$ and $\hat\sigma_2^2 = s_2^2/n_2$, where s_1^2 and s_2^2 are the observed values of the sample variances S_1^2 and S_2^2, respectively. Let

$$\hat{G}(w) = \Phi\left[\frac{\bar{x}\,w - \bar{y}}{\hat\sigma_1\hat\sigma_2\,\hat{a}(w)}\right],\ \hat{a}(w) = \left(\frac{w^2}{\hat\sigma_1^2} + \frac{1}{\hat\sigma_2^2}\right)^{1/2} \tag{3.4}$$

A $100(1-2\alpha)$ percent confidence interval (CI) for the ratio μ_E/μ_C is given by the interval

$$\text{CI} = \left(\hat{G}^{-1}(\alpha),\ \hat{G}^{-1}(1-\alpha)\right), \text{ for } \alpha < 1-\alpha. \tag{3.5}$$

It is simple to compute the CI of (3.5). Tabulate $\hat{G}$ to find the limits. That is, calculate $\hat{G}$, for example, for $w = 0.1, 0.11, \cdot\cdot, 1.09, 1.10$ and search for w for which $\hat{G}(w) = 0.05$ and $\hat{G}(w) = 0.95$. Alternatively, solve for w:

$$\hat{G}(w) = \alpha, \text{ and } \hat{G}(w) = 1-\alpha.$$

See Koti (2007) for details.

4. Mantel-Haenszel average risk difference

Rothmann et al., (2012, p. 299) have stated that there are many choices for how to do a stratified or adjusted non-inferiority analysis of a binary endpoint. This is one of them. This is an adjusted analysis, which uses the harmonic mean of the number of subjects in the experimental and control arms within a stratum as the stratum weight. Table 1 below contains some basic notation.

Table 1: *The ith contingency table in Rothmann et al. (2012, p. 301)*

	Outcome:		
Treatment	*Success*	*Failure*	*Total*
E	$x_{E,i}$	$n_{E,i} - x_{E,i}$	$n_{E,i}$
C	$x_{C,i}$	$n_{C,i} - x_{C,i}$	$n_{C,i}$
*Total**	n_{i+1}		N_i

* Rothmann et al. did not show n_{i+1} in the table.

They denote the sample proportions of successes in the ith ($i = 1, 2, \, .. \, , k$) stratum as $\hat{p}_{E,i}$ and $\hat{p}_{C,i}$ for the experimental arm and the control arm, respectively. Rothmann et al. (2012, p. 301) define the so called the Mantel-Haenszel estimator of the **common** risk difference across strata by

$$\hat{\Delta}_{MH} = \frac{\Sigma\ (n_{C,i}\, x_{E,i} - n_{E,i}\, x_{C,i})}{\Sigma\ (n_{E,i} \times n_{C,i}/N_i)} = \sum w_i\, \hat{\Delta}_i / \sum w_i \ , \tag{4.1}$$

where $\hat{\Delta}_i = \hat{p}_{E,i} - \hat{p}_{C,i}$ is the difference in the sample proportions in stratum i, and $w_i = n_{E,i} \times n_{C,i}/N_i$. Note that summation is over $i = 1, 2, \, .. \, , k$. They state that the weight w_i for a given stratum can be considered as the harmonic mean of the within-stratum sizes for the experimental and control arms.

The estimators

$$\hat{p}_{E,MH} = \frac{\Sigma(n_{E,i} \times n_{C,i}/N_i)\, \hat{p}_{E,i}}{\Sigma(n_{E,i} \times n_{C,i}/N_i)} \quad \text{and} \quad \hat{p}_{C,MH} = \frac{\Sigma(n_{E,i} \times n_{C,i}/N_i)\, \hat{p}_{C,i}}{\Sigma(n_{E,i} \times n_{C,i}/N_i)} \tag{4.2}$$

are regarded as the Mantel-Haenszel estimators of p_E and p_C , respectively. We have explicitly defined them in Section 4.1 below. Rothmann et al. say that the parameters p_E and p_C are analogous weighted averages of the respective strata probabilities of success. They point out that

$$\hat{\Delta}_{MH} = \hat{p}_{E,MH} - \hat{p}_{C,MH} \tag{4.3}$$

We think that $\hat{p}_{E,MH}$ and $\hat{p}_{C,MH}$ may be alternatively better written as

$$\hat{p}_{E,MH} = \textstyle\sum_1^k w_i' \hat{p}_{E,i}, \text{ and } \hat{p}_{C,MH} = \sum_1^k w_i' \hat{p}_{C,i}, \tag{4.4}$$

where

$$w_i' = \frac{n_{C,i} \times n_{E,i}/(n_{C,i} + n_{E,i})}{\sum_1^k [n_{C,i} \times n_{E,i}/(n_{C,i} + n_{E,i})]} \tag{4.5}$$

It follows that

$$\hat{\Delta}_{MH} = \textstyle\sum_1^k w_i' (\hat{p}_{E,i} - \hat{p}_{C,i}) \tag{4.6}$$

The estimand

$$\Delta_{MH} = p_{E,MH} - p_{C,MH} \ , \tag{4.7}$$

where

$$p_{E,MH} = \textstyle\sum_1^k w_i' p_{E,i} \text{ and } p_{C,MH} = \sum_1^k w_i' p_{C,i}, \tag{4.8}$$

is called the Mantel-Haenszel average risk difference (MHARD). Rothmann et al. (2012) state that when $\hat{p}_{E,MH}$ and $\hat{p}_{C,MH}$ have approximate normal distributions, **confidence intervals** can be found for the Mantel-Haenszel average risk difference Δ_{MH} of (4.7).

4.1 Author's comments on MHARD

In this section, we briefly refer to Stokes et al. (1995), Nam (1992) and Radhakrishna (1965) and borrow as much information as needed, to save reader's time. Note that, by definition, p_E and p_C are:

$$p_E = \sum_1^k w_i' p_{E,i} \text{ and } p_C = \sum_1^k w_i' p_{C,i} \qquad (4.9)$$

Rothmann et al. (2012) left out MH in the subscripts of the parameters p_E and p_C. As seen from (4.8), we have included MH in the subscripts of p_E and p_C. The estimand Δ_{MH} proposed in Rothmann et al. has several pitfalls and shortcomings. First, we point out that the estimand Δ_{MH} is mathematically intractable. Now

$$E\left(\hat{\Delta}_{MH}\right) = E\left[\sum_1^k w_i'\left(\hat{p}_{E,i} - \hat{p}_{C,i}\right)\right] \qquad (4.10)$$

We know that $E\left(\hat{p}_{E,i} - \hat{p}_{C,i}\right) = p_{E,i} - p_{C,i}$. How do we handle the weights w_i' in (4.5) in computing the expectation on the right hand side of (4.6)? It is a legitimate question because the MHARD is the pivotal parametric function that also depends on the weights that are data-dependent. As seen from (4.13) below, the Cochran-Mantel-Haenszel (CMH) test uses the weights $\{w_i\}$, which equal to the inverse of the harmonic mean of $n_{C,i}$ and $n_{E,i}$. Radhakrishna (1965), who discussed the Cochran's (1954) method of combining the results from several 2×2 contingency tables, identified the problem. He stated that the weights $\{w_i\}$ are assumed to have zero variances. The weights $\{w_i'\}$ lead to unconventional null and alternative hypotheses in a non-inferiority trial. See the hypotheses stated in (4.16) below. The investigator does not know the weights at the trial planning stage. The investigator will have a hard time in determining a suitable non-inferiority margin.

Next, we explain why the adjective Mantel-Haenszel in the characterization of the estimand Δ_{MH} is not justified. In superiority trial, under H_0, the null hypothesis of no treatment difference, $x_{E,i}$ has a hypergeometric distribution and its expected value is (Stokes et al., 1995; page 40):

$$E\left(x_{E,i} \mid H_0 : p_{i1} - p_{i2} = 0\right) = \frac{n_{i+1} \times n_{E,i}}{n_{C,i} + n_{E,i}} \left(= m_{i11}\right), \qquad (4.11)$$

where n_{i+1}, as shown in Table 1, is the total number of successes. Stokes et al. (1995) denote this expected value of $x_{E,i}$ as m_{i11} on page 40 i.e., in Section 3.2 of their book. It is used in the numerator of the Mantel-Haenszel statistic Q_{MH}. The statistic Q_{MH}, in their (Stokes et al.) notation, is as follows.

$$Q_{MH} = \{\sum n_{h11} - \sum m_{h11}\}^2 / \sum v_{h11}$$

$$Q_{MH} = \left\{\sum(n_{h1+} n_{h2+} / n_h)(p_{h11} - p_{h21})\right\}^2 / \sum v_{h11}, \qquad (4.12)$$

where

$$v_{h11} = V\{n_{h11} \mid H_0\} = \frac{n_{h1+}\, n_{h2+}\, n_{h+1}\, n_{h+2}}{n_h^2(n_h - 1)} \quad ,$$

and p_{h11}, p_{h21} are the observed proportions of favorable response. The summation on the right-hand-side of Q_{MH} is extended over all strata.

Radhakrishna (1965) writes: "Cochran's method consists of calculating a weighted average of the difference in efficacy between the two treatments in the various 2×2 tables, the weights being based on some consideration of optimality". It may be pointed out that the optimality is obtained under the null hypothesis of no treatment difference.

Nam (1992) derived a sample size formula for Cochran's statistic with continuity correction which guarantees that the actual Type I error rate of the test does not exceed the nominal level. Cochran considers the following null and alternative hypotheses (Nam, 1992):

H_0' : The common odds ratio is 1 ($\psi = 1$)
H_1' : The common odds ratio is greater than 1 ($\psi > 1$)

We point out that H_0' states that there is no treatment difference. Cochran's test statistic is based on the sum of the weighted difference

$$U = \sum w_i(\hat{p}_{E,i} - \hat{p}_{C,i}), \text{ where } w_i = n_{C,i} \times n_{E,i}/N_i \qquad (4.13)$$

The test statistic, as shown in Nam (1992), is

$$z_c = (U - \Delta'/2)/\{var_{est}(U)_0\}^{1/2} ,$$

where $var_{est}(U)_0 = \sum w_i\ P_i(1 - P_i)$ with $P_i = (x_{E,i} + x_{C,i})/N_i$, and $\Delta' = 1$ or 0 . Nam (1992) quotes Radhakrishna (1965), "The test is optimal under a logistic model and nearly efficient under a wide range of other models."

What is the difference between $\hat{\Delta}_{MH}$ of (4.1) and U of (4.13)? If you divide U by the sum of weights, $\sum w_i$, you get the estimator $\hat{\Delta}_{MH}$. That is how the qualifier "average" comes from. However, as explained earlier, these weights become irrelevant in the non-inferiority setting.

The weight $w_i = (n_{C,i} \times n_{E,i})/N_i$ mentioned in Rothmann et al. (2012) on page 301 does appear in the numerator of the Mantel-Haenszel statistic of (4.8) above. The expression

$$(n_{h1+}n_{h2+}/n_h)(p_{h11} - p_{h21})$$

shown in (4.12) is legitimately derived under the null hypothesis of no treatment difference. Rothmann et al. have expressed a null hypothesis in terms of an odds ratio on page 290. They don't state a null hypothesis in terms of the MHARD. If the null hypothesis of no treatment difference is not true,

$$E\big(\, x_{E,i} \mid H_0 : p_{i1} - p_{i2} \neq 0\big) \neq n_{C,i} \times n_{E,i}/N_i\, .$$

Then the equation (4.12) is not true. Then the adjective “Mantel-Haenszel” in the phrase MHARD is meaningless and irrelevant in the non-inferiority setting. It happens to be just a gimmick.

Yanagawa et al. (1994) proposed Mantel-Haenszel-type statistics for testing whether a new treatment is at least as effective as the standard treatment in comparative binomial trials. They considered the following hypotheses (superscript Y stands for Yanagawa)

$$H_0^Y\colon \pi_{i1} - \pi_{i2} = -\delta_i,\ \ i = 1, 2,\ .., K \qquad (4.14)$$
$$H_A^Y\colon \pi_{i1} - \pi_{i2} > -\delta_i,\ \ i = 1, 2,\ .., K\ ;$$

$0 \leq \delta_i < 1.$ Note that the hypotheses in (4.14) accommodate non-uniform non-inferiority margins. In fact, Miettinen and Nurminen (1985) and Yanagawa et al. (1994) have indicated that there is no need to introduce a ***common*** **risk difference** in the CMH setting. Rothmann et al. have proposed to compute a confidence interval on Δ_{MH} to demonstrate non-inferiority. We have seen their null and alternative hypotheses- stated in equation (11.23) on page 290. We imagine that they want to use the CI on Δ_{MH} to test the following null and alternative hypotheses (superscript R stands for Rothmann)

$$K_0^R\colon \Delta_{MH} \leq \Delta_0 \text{ vs. } K_A^R : \Delta_{MH} > \Delta_0\,. \qquad (4.15)$$

We write the hypotheses that are stated in (4.15) in expanded form:

$$K_0^R\colon \frac{\sum (n_{E,i} \times n_{C,i}/N_i)(p_{E,i} - p_{C,i})}{\sum (n_{E,i} \times n_{C,i}/N_i)} \leq \Delta_0 \text{ and } K_A^R\colon \frac{\sum (n_{E,i} \times n_{C,i}/N_i)(p_{E,i} - p_{C,i})}{\sum (n_{E,i} \times n_{C,i}/N_i)} > \Delta_0 \qquad (4.16)$$

The null and alternative hypotheses of (4.16) are unconventional and not easily interpretable. It is difficult to implement the design and analysis of the non-inferiority trial. Rothmann et al. (2012) did not elaborate on how they calculate the CI on Δ_{MH}. Farrington and Manning (1990) discussed the analysis of binary data from a non-inferiority trial. They state that when it is required to establish a materially significant difference between two treatments, or, alternatively, to show that two treatments are equivalent, standard test statistics and sample size formulae based on a null hypothesis of no difference no longer apply. The Wald-type tests do not conform to the first principle, which requires using the variance estimate constrained by the null hypothesis (Miettinen and Nurminen, 1985). Yanagawa (1994) used the restricted maximum likelihood variance estimate to test the null hypothesis in (4.14). This makes their (Yanagawa et al.) non-inferiority analysis statistically rigorous and acceptable. Rothmann et al. confidence interval approach to test K_0^R in (4.15) lacks rigor and clarity. Longford and Nelder (1999) wrote: “Even though much statistical practice has its origins in sound scientific concepts and principles, it is all too easily reduced to conventions and prescriptions (procedures). “ This is very true in this (MHARD) case.

5. The W-square test

Recently, Koti (2017) proposed a new Mantel-Haenszel (MH) type test to demonstrate non-inferiority in terms of the differences in success proportions and when the non-inferiority margin is not necessarily uniform in all strata. Derivation of this new test originates from Wittes and Wallenstein (1987), who have derived asymptotic unconditional power and sample size for the MH test in the comparative binomial (CB) design setting. The new test is called the W-square test.

The Table 2 below represents a typical 2×2 contingency table that contains the basic notations. Let π_{ij} denote the proportion of responders in the ith stratum and receiving treatment j $(= 1, 2)$. Assume that δ_i to be known constants and that $0 \leq \delta_i < 1$; $i = 1, 2, \cdot\cdot, T$, where T is the number of strata (2×2 tables). The $\{\delta_i\}$ represent the strata non-inferiority margins. Let $N = \sum n_{i\cdot\cdot}$, $\rho_i = n_{i1\cdot}/n_{i\cdot\cdot}$, $\lambda_i = n_{i\cdot\cdot}/N$, and $\bar{\pi}_i = \rho_i \pi_{i1} + (1 - \rho_i)\pi_{i2}$ .The summation is extended over all strata.

Table 2: *The ith 2×2 contingency table in Koti (2017)*

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

Koti (2017) considered the following null and alternative hypotheses K_0 and K_A .

$$K_0: \pi_{i1} - \pi_{i2} \leq -\delta_i, \text{ for all } i = 1, 2, \ .., T \tag{5.1}$$

$$K_A: \pi_{i1} - \pi_{i2} \geq -\delta_i, \text{ for all } i, \text{ and } \pi_{i1} - \pi_{i2} > -\delta_i, \text{ for some } i, \text{ where } \delta_i > 0 \text{ for all } i = 1, 2, \ .., T. \tag{5.2}$$

Let M_U denote the standard uncorrected Mantel Haenszel test statistic given by

$$M_U = \ \sum g_i / (\sum V_i)^{1/2} \ , \tag{5.3}$$

where

$$g_i = n_{i11} - n_{i1\cdot}\, n_{i\cdot 1}/n_{i\cdot\cdot} \ \text{ and } \ V_i = n_{i1\cdot}\, n_{i\cdot 1} n_{i2\cdot}\, n_{i\cdot 2}/\{n_{i\cdot\cdot}^2 (n_{i\cdot\cdot} - 1)\}$$

The W-square test rejects K_0 of (5.1) in favor of K_A in (5.2) at α level of significance if $M_U > c_\alpha$, where M_U is given by (5.3) and

$$c_\alpha = (z_{1-\alpha}\sigma_{CB} + \mu_{CB})/\sqrt{W_{CB}}$$

with

$$\mu_{CB} = -\sqrt{N} \ \sum \lambda_i \rho_i (1 - \rho_i)\delta_i \ ,$$

$$\sigma_{CB}^2 = \sum \lambda_i \, \rho_i (1 - \rho_i)[(1 - \rho_i)\pi_{i1}(1 - \pi_{i1}) + \rho_i \pi_{i2}(1 - \pi_{i2})] \ , \tag{5.4}$$

$$W_{CB} = \sum \lambda_i \, \rho_i (1 - \rho_i) \, [\bar{\pi}_i (1 - \bar{\pi}_i) + \delta_i^2 \, \rho_i (1 - \rho_i)/(N\lambda_i - 1)].$$

The π_{i2}s in σ_{CB}^2 of (5.4) are supposed to be reliably known from previous studies. Note that, under K_0, $\pi_{i1} = \pi_{i2} - \delta_i$. The power, p-value and sample size calculation are discussed in Koti (2017). The test described above assumes that T is large so that N → ∞. Similar test is provided when T is fixed, and $n_{i\cdot\cdot}$ → ∞. See Koti (2017) for a SAS code and for other details. A referee from ***Statistics in Biopharmaceutical Research*** stated that in general, setting 2 (fixed T, and with relatively large $\{n_{i..}\}$) is more reasonable in practice. W-square test can be applied even when the non-inferiority margin is identical for all strata.

6. Concluding remarks

The delta-method approach contradicts theoretical results. It is unsophisticated, irrational, awkward, and inappropriate to use in practice. Koti's (2007) Fieller-Hinkley distribution based analysis is a better one in that there are no major technical flaws.

The phrase "Mantel-Haenszel average risk difference" is meaningless in the non-inferiority setting. The associated confidence interval method is unscientific and irrelevant to the objective of proving the non-inferiority of an experimental treatment. Koti's (2017) W-square test is a better alternative. One of the reasons is that the W-square test approach provides asymptotic unconditional inference.

Acknowledgements

The author is thankful to all the statisticians who read this manuscript in full. The author appreciates their passion for statistics.

References

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 11: 283-319.

Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometrics,* 10, 417-51.

D'Agostino, R. B. Sr., Massaro, J. M. and Sullivan, L. M. (2003). Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine* 22: 169-186.

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.

Fieller, E. C. (1932). The distribution of the index in a normal bivariate population. *Biometrika,* 24: 428-440.

Govindarajulu, Z. (1988). *Statistical techniques in bioassay*. Karger: Verlag.

Hauschke, D. and Hothorn, L. A. (2007). Letter to the Editor. *Statistics in Medicine* 26: 230-236.

Hauschke, D., Kieser, M., Diletti, E., Burke, M. (1999). Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Statistics in Medicine* 18: 93-105.

Hinkley, D. V. (1969). On the ratio of two correlated normal variables. *Biometrika,* 56: 635-639.

Koti, K. M. (2017). A simple Mantel-Haenszel type test for non-inferiority. *Joint Statistical Meetings Proceedings*: 1436-1448.

Koti, K. M. (2007). Use of the Fieller-Hinkley distribution of the ratio of random variables in testing for non-inferiority. *Journal of biopharmaceutical statistics* 17(2): 215-228.

Longford, N. T. and Nelder, J. A. (1999). Statistics versus statistical science in the regulatory process. *Statistics in Medicine* 17(2): 215-228.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute* 2: 719-748.

Mauri, L. M., and D'Agostino, R. B. Sr. (2017). Challenges in the design and interpretation of non-inferiority trials. *The New England Journal of Medicine* 377: 1357-67.

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

Nam, J. (1992). Combination Sample size determination for case-control studies and the comparison of stratified and un-stratified analyses. *Biometrics,* 48: 389-395.

Radhakrishna, S. (1965). Combination of results from several 2×2 contingency tables. *Biometrics,* 21, No. 1, pp. 86-98.

Rothmann, M. D., Weines, B. L. and Chan, I. S. F. (2012). *Design and analysis of non-inferiority trials.* CRC Press, New York.

Seltman, H. (2019). *Approximations for mean and variance of a ratio.* www.stat.cmu.edu/~hseltman/files/ratio.pdf

Stokes, M. E., Davis, C. S. and Koch, G. G. (1995). *Categorical data analysis using the SAS System.* Cary, NC: 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.