Inferiority Index and the Behrens-Fisher Problem for Non-inferiority Trials

George Y.H. Chi

Janssen Research & Development, 920 Route 202S, Raritan, NJ 08869

Abstract

The classical Behrens-Fisher problem poses the question regarding the distribution of the test statistics for the null hypothesis defined by the mean difference when the variances are not equal under normality. The Behrens-Fisher distribution for the test statistic that is defined as the observed mean difference divided by the square root of the sum of the sample variances divided by their respective sample size has been derived [see Kim and Cohen (1998)]. Welch (1938) proposed an approximate *t*-test and derived its distribution. Dannenberg et al (1994) derived the corresponding extended Behrens-Fisher distribution for the test statistic under heterogeneity of variances for the equivalence hypothesis with a pre-specified equivalence margin for bioequivalence trials. In this paper, we will apply the theory of inferiority index under normal distributions to derive an extended Behrens-Fisher distribution assuming heterogeneity of variances under the inferiority null of a non-inferiority trial for the relative difference measure where the non-inferiority margin is actually a function of the standard deviation of the control distribution defined at a specified level of the inferiority index. Based on this extended Behrens-Fisher distribution, we can then derive the corresponding Welch approximate *t*-test and its associated distribution. Further improvement of this extended Welch approximate *t*-test can be made analogous to the improvement made for the classical Welch distribution by Bhoj (1993).

Key Words: Behrens-Fisher problem, extended Behrens-Fisher distribution, extended Welch approximate *t*-test, heterogeneity of variances, inferiority index, non-inferiority trial, and margin function.

1. Introduction

The Behrens-Fisher problem is posed here within the context of a comparative trial, where a new treatment T is compared to a control C relative to an outcome X of interest. Furthermore, it is assumed that $X_T \sim N(\mu_T, \sigma_T^2)$ and $X_C \sim N(\mu_C, \sigma_C^2)$ are normally distributed and a larger value of X represents a better outcome. Let $\delta_{RD} = \mu_T - \mu_C$, then the standard superiority trial considers testing the following hypothesis:

$$H_o: \delta_{RD} \leq 0 \quad vs. \quad H_a: \delta_{RD} \leq 0 \tag{1}$$

The standard *t*-test is given by

$$\hat{t} = \frac{\delta_{RD}}{\sqrt{\frac{(n_T+n_C)}{n_T n_C (n_T+n_C-2)}(n_T S_{X_T}^2 + n_C S_{X_C}^2)}} \qquad (2)$$

where $\hat{\delta}_{RD} = (\hat{\mu}_T - \hat{\mu}_C) = \hat{\bar{X}}_T - \hat{\bar{X}}_C$ represents the sample mean difference, and $S_{X_T}^2$ and $S_{X_C}^2$ are the sample variances of X_T and X_C respectively.

The denominator of $\hat{t}$ in (2) is the pooled estimate of the common variance under assumption of homogeneity of variances.

The Behrens-Fisher problem poses the following question: What happens to the test statistic $\hat{t}$ and its distribution when variances are not assumed to be homogeneous?

Under the assumption of heterogeneity of variances, the following test statistic has been considered:

$$T_{BF(\theta)} = \frac{\delta_{RD} - \theta}{\sqrt{(\frac{S_{X_T}^2}{n_T} + \frac{S_{X_C}^2}{n_C})}} \qquad (3)$$

The distribution of (3) has been termed the Behrens-Fisher distribution [Kim and Cohen (1998)] and it is given by

$$F_{T_{BF(\theta)}}(t) = \xi^b F_{t_{2(a+b)},\theta}\left(t\sqrt{\frac{2(a+b)\xi}{\eta}}\right) +$$

$$\sum_{k=1}^{\infty} \frac{\xi^b \Gamma(b+k)(1-\xi)^k}{\Gamma(b)k!} \times F_{t_{2(a+b+k)},\theta}\left(t\sqrt{\frac{2(a+b+k)\xi}{\eta}}\right) \qquad (4)$$

where $\sigma^2 = \frac{\sigma_T^2}{\sigma_C^2}$, $\eta = n_T(n_T - 1)\left(\frac{1}{\sigma^2}\frac{1}{n_C} + \frac{1}{n_T}\right)$, $\xi = \frac{n_T(n_T-1)}{n_C(n_C-1)}\frac{1}{\sigma^2}$, $a = \frac{n_T-1}{2}$, $b = \frac{n_C-1}{2}$ and

$F_{t_{2(a+b+k)},\theta}$ is the non-central *t*-distribution with *2(a+b+k)*-degrees of freedom and non-centrality parameter θ. Under the null in a superiority trial, $\theta = 0$ [see Hu (2010)].

Welch (1938) proposed to approximate the Behrens-Fisher distribution (4) by a *t*-distribution, t_ν, which is accomplished by approximating the pooled variance

$$V = \frac{\frac{S_{X_T}^2}{n_T} + \frac{S_{X_C}^2}{n_C}}{\frac{\sigma_{X_T}^2}{n_T} + \frac{\sigma_{X_C}^2}{n_C}} \qquad (5)$$

by the distribution of a χ_ν^2 random variable with ν-degrees of freedom, where

$$\nu = \frac{\left(\frac{\sigma_T^2}{n_T}+\frac{\sigma_C^2}{n_C}\right)^2}{\frac{1}{n_T-1}\left(\frac{\sigma_T^2}{n_T}\right)^2+\frac{1}{n_C-1}\left(\frac{\sigma_C^2}{n_C}\right)^2} \qquad (6)$$

and ν may be estimated by replacing the unknown variances σ_T^2 and σ_C^2 by their respective sample variances $S_{X_T}^2$ and $S_{X_C}^2$ [see Hu (2010)].

The performance of Welch's approximate *t*-test has been investigated by Bhoj (1993) and appears to have reasonable performance.

Now the inferiority hypothesis in a non-inferiority trial for relative difference measure is usually stated as follows:

$$H_o: \delta_{RD} \leq \delta_{RD,o} \quad vs. \quad H_a: \delta_{RD} \leq \delta_{RD,o} \qquad (7)$$

where $\delta_{RD,o} < 0$ is a fixed non-inferiority margin determined by the experimenter.

The corresponding Behrens-Fisher problem raises the following question: What would be the appropriate test statistic for testing (7) and what would be its distribution when variances are not assumed to be equal?

This paper discusses an extension of the Behrens-Fisher distribution for the test statistic testing the inferiority null of a non-inferiority trial through an application of the theory of inferiority index developed in Li and Chi (2011) and also derives the corresponding Welch approximate *t*-test.

2. Theory of Inferiority Index under Normal Distributions

Let F_T and F_C denote the distribution function of X_T and X_C respectively. The inferiority index between the distributions F_T and F_C is defined in Li and Chi (2011) as:

$$\rho = \rho(F_T, F_C) = Sup_{-\infty<u<\infty}[F_T(u) - F_C(u)] \qquad (8)$$

The inferiority index ρ is simply the one-sided version of the Kolmogorov-Smirnov distance metric between two sample distributions. It simply measures how much worse off T is compared to C. The inferiority index ρ is a probability and hence can be used as an index for measuring the degree of stringency of an inferiority margin for certain effect measure defined by the population parameters. It should be noted that as a measure of distributional differences, the inferiority index ρ does not require assumption of homogeneity of variances.

Under normal distributions, there is a function g linking the scaled relative difference $\delta_{SRD} = \frac{\mu_T - \mu_C}{\sigma_C}$ and the variance ratio $\sigma^2 = \frac{\sigma_T^2}{\sigma_C^2}$ to the inferiority index ρ. That is, for each pair of values (δ_{SRD}, σ), where $\delta_{SRD} < 0$ and $\sigma^2 > 0$, there is a unique inferiority index ρ, such that

$\rho = g(\delta_{SRD},\sigma)$.

Conversely, there is an inverse function g^{-1} such that for each $\rho, 0 < \rho < 1$, there is a restricted interval $(\sigma_1(\rho), \sigma_2(\rho))$ such that for each $\sigma \epsilon\ (\sigma_1(\rho), \sigma_2(\rho))$, $g^{-1}(\rho, \sigma) = \delta_{SRD}$.

The pair of link functions g and g^{-1} are very useful in margin specification, since for any inferiority margin δ_{SRD}, however it was derived, the inferiority index $\rho = g(\delta_{SRD},\sigma^2)$ provides a measure of the degree of stringency of δ_{SRD} at a given variance ratio σ^2. Conversely, for a given inferiority index level ρ and a variance ratio σ^2 in the restricted interval $\sigma \epsilon\ (\sigma_1(\rho), \sigma_2(\rho))$, $g^{-1}(\rho, \sigma^2) = \delta_{SRD}$ defines a unique inferiority margin δ_{SRD} with the desired degree of stringency ρ.

Hence, in designing a non-inferiority trial, one can pre-specify the degree of stringency ρ_o that is desired, and if one has some knowledge of the variance ratio σ^2, then one can derive the non-inferiority margin $g^{-1}(\rho_o,\sigma^2) = \delta_{SRD,o}$ with the desired degree of stringency. One can then define the corresponding non-inferiority hypothesis as follows:

$$H_{SRD,o}: \delta_{SRD} \le \delta_{SRD,o} = \mathrm{g}^{-1}(\rho_o,\sigma^2) \quad vs. \quad H_{SRD,o}: \delta_{SRD} > \delta_{SRD,o} = \mathrm{g}^{-1}(\rho_o,\sigma^2) \qquad (9)$$

Let $\hat{\delta}_{SRD} = \frac{\hat{\mu}_T - \hat{\mu}_C}{\hat{\sigma}_C}$ be the estimate for the scaled relative difference measure δ_{SRD}. Then, Li and Chi (2011) proved the following theorem.

<u>Theorem 1</u>: Assuming equal sample size, the statistic $\sqrt{n}\left(\hat{\delta}_{SRD} - \delta_{SRD,o}\right) \sim N(0, \Sigma^2_{SRD,o})$ under the inferiority null in (9), where the asymptotic variance $\Sigma^2_{SRD,o}$ is given by

$$\Sigma^2_{SRD,o} = (1 + \sigma^2) + \frac{\delta^2_{SRD,o}}{4} \qquad (10)$$

That is, the test statistic

$$T_{SRD} = \frac{\sqrt{n}(\delta_{SRD} - \delta_{SRD,o})}{\sqrt{(1+\sigma^2) + \frac{\delta^2_{SRD,o}}{4}}} \sim N(0,1) \qquad (11)$$

Now the link functions between the inferiority index ρ and δ_{SRD} and σ^2 does not extend to the relative difference measure δ_{RD}, which is the measure of interest in most applications and is the focus of the Behrens-Fisher problem.

However, let us note that if ρ_o is a desired inferiority index level, and the variance ratio σ^2 is known, then one can define the inferiority margin function for the relative difference measure by

$$\delta_{RD,o}(\sigma_C) = \sigma_C\ \delta_{SRD,o} = \sigma_C\ g^{-1}(\rho_o,\sigma^2) \qquad (12)$$

It should be pointed out that for any given σ_C, the margin $\delta_{RD,o}(\sigma_C)$ is defined with the same desired degree of stringency ρ_o, since it is derived from $\delta_{SRD,o} = g^{-1}(\rho_o, \sigma^2)$. Indeed, the entire margin function as defined in (12) has the desired degree of stringency ρ_o. This is important, since we normally wouldn't know the variance σ_C^2, and hence the inferiority margin $\delta_{RD,o}(\sigma_C)$ would be unknown.

Let us define the non-inferiority hypothesis for the relative difference measure δ_{RD} at a given inferiority index level ρ_o and variance ratio σ^2 by the margin function $\delta_{RD,o}(\sigma_C)$ in (12) as follows:

$$H_{RD,o}: \delta_{RD} \leq \delta_{RD,o}(\sigma_C) \quad vs. \quad H_{RD,o}: \delta_{RD} > \delta_{RD,o}(\sigma_C) \tag{13}$$

Now, we will derive the theorem corresponding to Theorem 1 for the relative difference measure δ_{RD}.

<u>Theorem 2</u>: If the variance of the control σ_C^2 were known, then the margin $\delta_{RD,o}(\sigma_C)$ in (13) would be considered as fixed. Then, assuming equal sample size, the statistic $\sqrt{n}\left(\hat{\delta}_{RD} - \delta_{RD,o}(\sigma_C)\right) \sim N(0, \Sigma^2_{RD,o})$ under the inferiority null in (13), where the asymptotic variance $\Sigma^2_{RD,o}$ is given by

$$\Sigma^2_{RD,o} = \sigma_C^2 \left(+ \frac{\delta^2_{SRD,o}}{4} \right) + \sigma_T^2 \tag{14}$$

That is, the test statistic

$$T_{RD} = \frac{\sqrt{n}\;[\delta_{RD} - \delta_{RD,o}(\sigma_C)]}{\sqrt{\sigma_C^2\left(1 + \frac{\delta^2_{SRD,o}}{4}\right) + \sigma_T^2}} \sim N(0,1) \tag{15}$$

where $\hat{\delta}_{RD} = \hat{\mu}_T - \hat{\mu}_C$ and the variance σ_T^2 may be replaced by its sample variance estimates.

<u>Proof</u>: Essentially, from the functional relation, $\delta_{RD} = f(\delta_{SRD}, \sigma_C) = \sigma_C \delta_{SRD}$, one can apply the multivariate Taylor transformation and obtain the approximation

$$\sqrt{n}(\hat{\delta}_{RD} - \delta_{RD,o}(\sigma_C)) \approx \frac{\partial f}{\partial \delta_{SRD}}\left[\sqrt{n}\,(\hat{\delta}_{SRD} - \delta_{SRD,o})\right] + \frac{\partial f}{\partial \sigma_C}\left[\sqrt{n}\,(\hat{\sigma}_C - \sigma_C)\right] \tag{16}$$

One can show that the asymptotic variance of $\sqrt{n}(\hat{\delta}_{RD} - \delta_{RD,o})$ is given by

$$\mathrm{var}\left(\sqrt{n}\left[\hat{\delta}_{RD} - \delta_{RD,o}(\sigma_C)\right]\right) = \Sigma^2_{RD,o} = \sigma_C^2\left(1 + \frac{\delta^2_{SRD,o}}{4}\right) + \sigma_T^2 \tag{17}$$

However, when the variance of the control σ_C^2 is not known, then one would like to substitute the unknown variance σ_C^2 by its estimate both in the numerator and the denominator of the test statistic in (15). However, if this is done, then the variance in the denominator of the test statistic in (15) would need a slight adjustment as shown in the next theorem.

<u>Theorem 3</u>: If the variance of the control σ_C^2 were considered as unknown, then the margin $\delta_{RD,o}(\sigma_C)$ in the hypothesis (13) would be considered as a fixed *margin function*. Then, assuming equal sample size, the statistic $\sqrt{n}\ [\hat{\delta}_{RD} - \delta_{RD,o}(\hat{\sigma}_C)]$ is asymptotically normal with mean 0 and variance $\Sigma_{RD,o}^{*2}$ under the inferiority null of (13), where

$$\Sigma_{RD,o}^{*2} = \sigma_C^2\ \left(1 + \frac{\delta^2_{SRD,o}}{2}\right) +\ \sigma_T^2$$

That is, the test statistic

$$T_{RD}^* = \frac{\sqrt{n}\ [\delta_{RD} - \delta_{RD,o}(\hat{\sigma}_C)]}{\sqrt{\hat{\sigma}_C^2\left(1+\frac{\delta^2_{SRD,o}}{2}\right)+\hat{\sigma}_T^2}} \sim N(0,1) \qquad (18)$$

where $\hat{\delta}_{RD} = \hat{\mu}_T - \hat{\mu}_C$ and $\hat{\sigma}_C^2$ and $\hat{\sigma}_T^2$ are their sample variance estimates.

<u>Proof</u>: The proof follows from Theorem 2 and an application of the delta method.

This result is analogous to the asymptotic results established by Zhang (2008) for (variable) margin functions that satisfy certain regularity conditions for non-inferiority trials with binary outcomes. However, it is worth emphasizing that the inferiority *margin function* $\delta_{RD,o}(\sigma_C)$ as defined in (12) provides the same degree of stringency ρ_o for all values of σ_C.

3. The Extended Behrens-Fisher Distribution and the Extended Welch Approximate *t*-Test

From the form of the asymptotic variance (18), we can derive the extended Behrens-Fisher distribution. This is given by the next theorem.

<u>Theorem 4</u>: For the non-inferiority hypothesis (13), the following test statistic

$$T_{BF}^*(\theta) = \frac{[\hat{\delta}_{RD} - \delta_{RD,o}(\hat{\sigma}_C)] + \theta}{\sqrt{\hat{\sigma}_C^2\left(1+\frac{\delta^2_{SRD,o}}{2}\right)+\hat{\sigma}_T^2}} \qquad (19)$$

which has the Behrens-Fisher distribution $F_{BF}(\theta)$ given below.

$$F_{T^*_{BF}(\theta)}(t) = \xi^2 F_{t_{2(a+b)},\theta}\left(t\sqrt{\frac{2(a+b)\xi}{\eta}}\right) +$$

$$\sum_{k=1}^{\infty} \frac{(-1)^k \xi^{b+k} \Gamma(b+k)(1-\xi^{-1})^k}{\Gamma(b)k!} \times F_{t_{2(a+b+k)},\theta}\left(t\sqrt{\frac{2(a+b+k)\xi}{\eta}}\right) \qquad (20)$$

where

$\eta = n_T(n_T - 1)\left[\frac{\gamma}{\sigma^2}\frac{1}{n_C} + \frac{1}{n_T}\right]$, $\sigma^2 = \frac{\sigma_T^2}{\sigma_C^2}$, $\xi = \frac{n_T(n_T-1)}{n_C(n_C-1)}\frac{\gamma}{\sigma^2}$, $a = \frac{n_T-1}{2}$, $b = \frac{n_C-1}{2}$, and

$\gamma = \left(1 + \frac{\delta_{SRD,o}}{2}\right)$.

Again, $F_{t_{2(a+b+k)},\theta}$ is the non-central *t*-distribution with *2(a+b+k)*-degrees of freedom and non-centrality parameter θ. Note that here under the inferiority null of (13), $\theta = 0$. The variances σ_T^2 and σ_C^2 may be estimated by their sample variances.

Similarly, one can derive the extended Welch approximate *t*-test for the non-inferiority hypothesis (13). The extended Welch *t*-distribution is given by the following theorem.

Theorem 5: The extended Behrens-Fisher distribution $F_{T^*_{BF}(0)}(t)$ as given in (20) for the test statistics

$$T^*_{BF}(0) = \frac{\left[\hat{\delta}_{RD} - \delta_{RD,o}(\hat{\sigma}_C)\right]}{\sqrt{\left(\hat{\sigma}_C^2\left(1 + \frac{\delta^2_{SRD,o}}{2}\right) + \hat{\sigma}_T^2\right)}}$$

under the inferiority null of (13) can be approximated by an extended Welch approximate *t*-test, t^*_υ, where the degrees of freedom υ is given by:

$$\upsilon = \frac{\left[\frac{\gamma\,\sigma_C^2}{n_C} + \frac{\sigma_T^2}{n_T}\right]^2}{\frac{\gamma^2\,\sigma_C^4}{n_C^2(n_C-1)} + \frac{\sigma_T^4}{n_T^2(n_T-1)}}$$

where $\gamma = \left(1 + \frac{\delta^2_{SRD,o}}{2}\right)$ and the unknown variances σ_C^2 and σ_T^2 may be estimated by their sample variances S_T^2 and S_C^2.

4. Discussion

Theorem 3 will be useful for most applications involving mean difference. In such applications, the inferiority margin function $\delta_{RD,o}(\sigma_C)$ will be defined at a pre-specified index level ρ_o. The test statistic T^*_{RD} in (18) or the Welch's approximate *t*-test t^*_υ in Theorem 5 may be used to test the non-inferiority hypothesis (13).

Dannenberg et al. (1994) considered the extension of the Behrens-Fisher problem to bioequivalence studies, where it is assumed that the equivalence margin Δ has been pre-specified in some manner, though not linked to an inferiority index as presented in this paper for non-inferiority trials, which is very crucial in the formulation of the problem. More generally, the theory of inferiority index appears to be the proper framework for developing the statistical framework for addressing the bioequivalence problem for highly variable drugs. Work has been done for cross-over design in this regard.

The performance of the Welch's approximate *t*-distribution t^*_υ in Theorem 5 needs to be investigated as previously performed by other authors Golhar (1972), Bhoj (1993), Reed (2003) and Best (2012). It can be somewhat improved using the method of Bhoj (1993) by matching higher moments. One can also investigate the optimal sample size allocation when the variances are not equal for example as done by Dette & Munk (1997) and Dann & Koch (2008).

For NI trials with binary outcomes where the heterogeneity of variances is always true under the inferiority null, the Behrens-Fisher problem has a satisfactory resolution. The results will be reported elsewhere.

References

Best. DJ amd Rayner, JCW (2012). Welch's approximate solution for the Behrens-Fisher problem. *Technometrics* 29 (2): 206-210.

Bhoj, DS (1993). An approximate solution to the Behrens-Fisher problem. *Biometrical Journal* 35 (5): 635-640.

Dann, RS and Koch, GG (2008). Methods for one-sided testing of the difference between proportions and sample size considerations related to non-inferiority clinical trials. *Pharmaceutical Statistics* 7: 130-141.

Dannenberg, O, Dette, H and Munk, A (1994). An extension of Welch's approximate *t*-solution to comparative bioequivalence trials. *Biometrika* 81(1): 91-101.

Dette, H and Munk, A (1997). Optimum allocation of treatments for Welch's test in equivalence assessment. *Biometrics* 53: 1143-1150

Golhar, MB (1972). The errors of first and second kinds in Welch-Aspin's solution of the Behrens-Fisher problem. *Journal of Statistical Computation and Simulations* 1: 209-224.

Hu, FJ (2010). Methods for comparing two means with application in adaptive clinical trials. Master of Science Thesis, Department of Mathematics, Georgia Southern University.

Kim, SH and Cohen, AS (1998). On the Behrens-Fisher Problem: A Review, Journal of Educational and Behavioral Statistics. 23 (4): 356-377.

Li, Gang & Chi, GYH (2011). Inferiority index and margin in NI trials. *Statistics in Biopharmaceutical Research* 3(2): 288 – 301.

Reed, JF III (2003). Solutions to the Behrens-Fisher problem. *Computer Methods and Programs in Biomedicine* 70: 259-263.

Welch, BL (1936). The specification of rules for rejecting too variable a product, with particular reference to an electric lamp problem. *Journal of the Royal Statistical Society, Supplement III* 1: 29-47.

Welch, BL (1938). The significance of the difference between means when the population variances are unequal. *Biometrika* 29: 350-362.

Zhang, Z (2006). Non-inferiority testing with a variable margin. *Biometrical Journal* 48 (6): 948-965.