Advances and Applications in Statistics

Published Online: August 2016
http://dx.doi.org/10.17654/AS049010067
Volume 49, Number 1, 2016, Pages 67-86 **ISSN: 0972-3617**

ADJUSTING NOMINAL SIGNIFICANCE LEVELS AND TEST SIZES WHEN USING AN ASYMPTOTIC NON-INFERIORITY TEST

Félix Almendra-Arao[1], Hortensia Reyes-Cervantes[2] and José Juan Castro-Alva[2]

[1]UPIITA del Instituto Politécnico Nacional
Av. Instituto Politécnico Nacional 2580
Col. Laguna Ticomán, 07340, México, D. F.
México

[2]Facultad de Ciencias Físico Matemáticas
Benemérita Universidad Autónoma de Puebla
México

Abstract

In a recent article, Almendra-Arao et al. [5] shown that the classical asymptotic non-inferiority test for two independent proportions behaves in a liberal form, that is, the type I error is inflated and this happens inclusively for sample sizes as large as 1000, moreover the inflation is severe. This problem is not an exception of this test, but it is a common weakness of asymptotic tests. Therefore, it is recommended to have a way of adjusting the nominal level to specify to apply the asymptotic statistical test. In this research, we use the binary search algorithm to adjust both, the nominal significance level

Received: May 12, 2016; Accepted: June 30, 2016
2010 Mathematics Subject Classification: 62F05.
Keywords and phrases: non-inferiority test, Blackwelder test, significance level, binomial proportions, hypothesis test.
Communicated by K. K. Azad

and the test size of any asymptotical non-inferiority test in a conservative form, that is, to obtain an adjusted test size less than or equal to the nominal significance level. The method is motivated by using the classical asymptotic non-inferiority test for two independent proportions to conduct the presentation; however the scope for the application of this method is wide and includes any asymptotic non-inferiority or superiority test for two independent proportions. Calculations were carried out in a computational program in C++ written by the authors pursuing that objective.

Introduction

It is currently a common practice to use active-controlled trials in place of placebo-controlled trials to support marketing authorization of new medical products. Frequently active-controlled trials are based on non-inferiority trials, especially since the appearance of several regulatory guidelines that now recommend the use of this design in active-controlled trials. Non-inferiority statistical tests are used to demonstrate that a new therapy (usually with less secondary effects, easier application or less cost) is not substantially inferior in efficacy to the standard one.

Among several non-inferiority tests for two proportions (Almendra-Arao [1]; Almendra-Arao [4]; Dunett and Gent [12]; Blackwelder [8]; Miettinen and Nurminen [18]; Hauck and Anderson [14]; Farrington and Manning [13]; Chan [9]; Chen et al. [11]; Tu [21]; Martin and Herranz [16]; Martin and Herranz [17]; and Li and Chuang-Stein [15]), the classical asymptotic test or Blackwelder test has an outstanding role because it is frequently used in practice, mainly due to its simplicity. However, as usual for asymptotical tests, the Blackwelder test has the disadvantage of being liberal in that the test sizes are greater than the required nominal significance level (α).

Li and Chuang-Stein [15] made an evaluation of the performance of two very often used statistical procedures, the classical asymptotic normal approximation and the same method with the Hauck-Anderson continuity correction, their evaluation was based on simulation to estimate type I error and power.

Almendra-Arao [1] continued this investigation, but did an exact calculation of type I errors and test sizes instead of estimation by simulation. The main conclusion of the work done in Almendra-Arao [1] was that the test is indeed liberal, as test sizes were greater than the nominal level α, considering under study configurations $30 \leq n_1 = n_2 \leq 300$, for non-inferiority margin 0.10 and 0.15 and nominal significance level 0.025 and 0.05.

As currently understood, in the clinical context non-inferiority tests are often applied for sample sizes greater than 300. It is natural, then, to ask if for these larger sample sizes the behavior is even acceptable.

Thus, in continuing their research, Almendra-Arao et al. [5] developed a numerical study of the behavior of test sizes for the Blackwelder test. This analysis was based on sample sizes $n_1 = n_2 = 30(10), ..., 1000$; nominal significance levels 0.025 and 0.05 and non-inferiority margins 0.05, 0.10, 0.15, 0.20; also were considered unbalanced designs with sample sizes $(n_1, n_2 = 1.5n_1)$ with $n_1 = 50(50), ..., 1000$; and $(n_1 = 1.5n_2, n_2)$ with $n_2 = 50(50), ..., 1000$. The main conclusion in this investigation was that although it is known theoretically that test sizes converge to the nominal significance level α, the test continues to remain liberal (test sizes are greater than α) for the studied configurations, even for sample sizes as large as 1000, for balanced designs, and for sample sizes as large as 1500, for unbalanced designs. For all configurations studied in Almendra-Arao et al. [5], the test sizes were greater than α.

As in a statistical test it is essential to control type I errors, it is also advantageous to have a way of guaranteeing that the test size of an asymptotic test is less than or equal to the nominal level (α) because otherwise one loses the control of the type I error and one would have the possibility of making a type I error greater than the avowed (α).

For this reason, the goal of this research is to introduce the binary search algorithm in this context to adjust both the nominal level and the test size

such that when the test is performed, it can be guaranteed that the test size is less than or equal to α.

Although the method is applicable to any asymptotic test, it is presented in a tangible form, through the use of the classical asymptotic non-inferiority test for two independent proportions. To carry out the necessary calculations for this test for large sample sizes in a reasonable time, first it was necessary to solve several numerical difficulties, as shown below. These difficulties were resolved by applying the recommendations in Almendra-Arao [3].

The Framework

Consider two binomial independent random variables X_1 and X_2 with parameters $(n_1,\ p_1)$ and $(n_2,\ p_2)$, respectively, where p_1 and p_2 represent true response probabilities of the standard drug and new drug, respectively, and consider the set of hypotheses

$$H_0 : p_2 \leq g(p_1) \text{ vs } H_1 : p_2 > g(p_1) \tag{1}$$

with $g : [0, 1] \to \mathbb{R}$, a nondecreasing function of class C^2 and $g(p_1) \leq p_1$ for all $p_1 \in [0, 1]$.

The function g can be represented in the form $g(p_1) = p_1 - \delta(p_1)$, the function δ is called the *margin function*. The domain of the function δ will be denoted by D_δ.

Naturally, the margin function must be a non-negative function, that is, $\delta(p_1) \geq 0$, for all $p_1 \in D_\delta$; the special case $\delta(p_1) \equiv 0$ leads to the superiority case, and when $\delta(p_1)$ is not identically zero on D_δ corresponds properly to the non-inferiority instance.

Let T denote a statistic for the hypothesis testing problem (1).

Thus, the joint likelihood function is

$$L(p_1,\ p_2;\ x_1,\ x_2) = \prod_{i=1}^{2} \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i}$$

and the power function is $\beta_T(p_1, p_2) = \sum_{(x_1, x_2) \in R_T(\alpha)} L(p_1, p_2; x_1, x_2)$, where $R_T(\alpha)$ is used to denote the critical region for the statistic *T*.

The corresponding sample space is $\chi = \{0, ..., n_1\} \times \{0, ..., n_2\}$ and the parameter space can be conveniently represented as $\Theta = [0, 1]^2$.

Therefore, the test size is given by $\sup_{\theta \in \Theta_0} \beta_T(p_1, p_2)$, where $\Theta_0 = \{(p_1, p_2) \in \Theta : p_2 \leq p_1 - \delta(p_1)\}$ is the null space.

An essential property for non-inferiority tests is that the critical region satisfies the two properties in the next definition.

A critical region R_T for a statistic *T* is a *Barnard convex set* if the following two properties are satisfied:

(a) $(x_1, x_2) \in R_T \Rightarrow (x_1 - 1, x_2) \in R_T, \ \forall\, 1 \leq x_1 \leq n_1, 0 \leq x_2 \leq n_2.$

(b) $(x_1, x_2) \in R_T \Rightarrow (x_1, x_2 + 1) \in R_T, \ \forall\, 0 \leq x_1 \leq n_1, 0 \leq x_2 \leq n_2 - 1.$

That the critical regions be Barnard convex sets is not only a matter of convenience to compute test sizes, but it is also a necessary requirement in order for the non-inferiority test to be coherent; see for example Almendra-Arao [2] and Almendra-Arao and Sotres-Ramos [7].

Another condition, to be fulfilled by non-inferiority critical regions for balanced designs, is defined as follows.

Let $n_1 = n_2 = n$. A critical region R_T for a statistic *T* is said to fulfill the *condition of symmetry in the same tail* (SST) if $(x_1, x_2) \in R_T \Rightarrow (n - x_2, n - x_1) \in R_T$.

Classical Asymptotic Non-inferiority Test

When in the expression (1) the difference between proportions as a dissimilarity measure is used, the hypotheses to contrast are

$$H_0 : p_1 - p_2 \geq d_0 \text{ vs } H_1 : p_1 - p_2 < d_0, \tag{2}$$

where d_0 is a positive constant, usually small.

The Blackwelder's or classical statistic to contrast these hypotheses is

$$T_0(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\hat{\sigma}},$$

where $\hat{p}_i = \frac{X_i}{n_i}$ is the maximum likelihood estimator of p_i for $i = 1, 2$ and $\hat{\sigma}$ is the following estimator of the standard deviation of $\hat{d} = \hat{p}_1 - \hat{p}_2$, $\hat{\sigma} = \left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2} \right)^{1/2}$.

It is known that the statistic T_0 has normal standard asymptotic distribution.

For a given nominal significance level α, the critical region of the asymptotic test is given by

$$R_{T_0}(\alpha) = \{(x_1, x_2) \in \chi : T_0(x_1, x_2) < -z_\alpha\},$$

where z_α is the upper quantile α of the standard normal distribution, in other words, $P(Z > z_\alpha) = \alpha$. Notation for the critical region $R_{T_0}(\alpha)$ very often will be simplified to R_{T_0}, $R_0(\alpha)$ or R_0.

$C = \frac{1}{2 \min(n_1, n_2)}$ is known as Hauck-Anderson continuity correction and will be used in what follows.

Thus, we have the other test which considers this continuity correction; its statistic is given by

$$T_1(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0 + C}{\hat{\sigma}}.$$

For the test T_1, we will use a similar notation as used for T_0.

Since, as by definition

$$\hat{\sigma} = \hat{\sigma}(X_1,\ X_2) = \left(\frac{\frac{X_1}{n_1}\left(1 - \frac{X_1}{n_1}\right)}{n_1} + \frac{\frac{X_2}{n_2}\left(1 - \frac{X_2}{n_2}\right)}{n_2} \right)^{1/2},$$

it is clear that $\hat{\sigma}$ is equal to zero in four points and in these points T_0 and T_1 remain undefined. To be able to calculate both tests in these points, Almendra-Arao [1] suggested a redefinition of $\hat{\sigma}$, which will be used in what follows.

As it is known, when the critical region for this non-inferiority test is a Barnard convex set, the test size is given by

test size

$$= \max_{p \in [d_0, 1]} \left\{ \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \binom{n_1}{x_1} p^{x_1} (1-p)^{n_1 - x_1} \binom{n_2}{x_2} (p - d_0)^{x_2} (1 - p + d_0)^{n_2 - x_2} \times I_{[(x_1, x_2) \in R_T(\alpha)]} \right\} \tag{3}$$

see for example Almendra-Arao [1], Röhmel and Mansmann [20] and Almendra-Arao and Sotres-Ramos [6].

As was shown by Almendra-Arao et al. [5] and Almendra-Arao [1], the above redefinition of $\hat{\sigma}$ is essential in order for the critical regions of the statistical tests T_0 and T_1 to meet the conditions for the Barnard convex set definition and SST.

Furthermore, if the critical region fulfills SST, then the expression (3) can be reduced to

test size

$$= \max_{p \in [d_0, (1+d_0)/2]} \left\{ \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \binom{n_1}{x_1} p^{x_1} (1-p)^{n_1 - x_1} \times \binom{n_2}{x_2} (p - d_0)^{x_2} (1 - p + d_0)^{n_2 - x_2} I_{[(x_1, x_2) \in R_T(\alpha)]} \right\} \tag{4}$$

see for example Almendra-Arao [1].

Hereafter, we will use $ts(\alpha)$ *to denote the test size for a given nominal level* α.

Adjusting Nominal Levels and Test Sizes

Prior to presenting the technique to adjust the nominal significance levels and the test sizes, we will begin this section by explaining the procedure necessary to calculate test sizes.

When the critical region for a non-inferiority test is a Barnard convex set, Almendra-Arao [3] proved that the power function has manageable representations. Therefore, under such assumption it is practical to apply Newton's method. In the following, we describe how to calculate test sizes by using Newton's method.

Procedure used to calculate test sizes

For this procedure we have fixed n_1, n_2, d_0, α.

1. Verify that the critical region is a Barnard convex set or not.

2. Verify that the critical region fulfills SST.

3. If the critical region is both, a Barnard convex set and fulfills SST, then use (4), discretizing the interval $[d_0, (1+d_0)/2]$ with a step Δ, in this investigation we took $\Delta = 0.01$ for the increments of p in (4) to obtain an initial approximation α_0 to the test size. We denote the corresponding value of the parameter p by p_0.

4. If the critical region is a Barnard convex set and does not fulfill SST, then use (3), discretizing the interval $[d_0, 1]$ with a step Δ, we took $\Delta = 0.01$ for the increments of p in (3) to obtain an initial approximation to the test size. We denote the corresponding value of the parameter p by p_0.

5. Use the value p_0 as seed or initial value to apply Newton's method.

6. Apply Newton's method and denote the approximation obtained in the step i, for $i \geq 1$, as α_i. If the obtained succession p_i converges, then there

exists a pair of consecutive approximations, denoted by p_t and p_{t+1}, with $t \geq 0$, such that $|\, p_t - p_{t+1} \,| < \varepsilon$; in this research, we took, $\varepsilon = 0.00001$.

7. If Newton's method fails to converge, then use the exhaustive method with a refinement Δ', in this work, we took $\Delta' = 0.1\Delta$. If the critical region fulfills SST, this new application of exhaustive method is on the interval $[p_0 - 5\Delta,\, p_0 + 5\Delta] \bigcap [d_0, (1+d_0)/2]$, whereas that if the critical region does not fulfill SST, it is on the interval $[p_0 - 5\Delta,\ p_0 + 5\Delta] \bigcap [d_0,\, 1]$.

For more details about the above procedure, please consult Almendra-Arao [3].

Note that the previous description depends on the statistic only when using (3) or (4), these formulae can be easily generalized and therefore can be applied to any statistical test.

In all cases, we have studied in the present investigation, the critical regions were Barnard convex sets. However, if some critical region is not a Barnard convex set, it is possible to apply the procedure described above to the Barnard convex hull, see Almendra-Arao [2].

In all balanced cases (equal sample sizes) that we have analyzed in this research, SST property was fulfilled.

Hereafter, we will be using the following notation.

α : nominal significance level, that is, the required nominal level, usually 0.025 or 0.05.

α_{adj} : adjusted nominal significance level that is necessary to specify, such that, when used with the testing procedure $(T_0$ or $T_1)$ then the corresponding test size is less than or equal to α.

ts_{adj} : adjusted test size, this is the test size obtained when the nominal level α_{adj} is used with the test procedure, in other words, this is the real test size corresponding to the nominal test size α_{adj}.

Next we describe the procedure to search α_{adj}, this procedure is based on the computational binary search algorithm. In all the processes we fixed the values of n_1, n_2, d_0 so what will change is the nominal level to apply the procedure above.

For a concise description of the method in general, first we present a useful notation.

Given a closed interval $I_i = [a_i, b_i]$, the following notation will be used:

m_i : middle point of I_i, that is, $m_i = (a_i + b_i)/2$,

L_{i+1} : the left half subinterval of I_i, that is, $L_{i+1} = [a_i, (a_i + b_i)/2]$,

R_{i+1} : the right half subinterval of I_i, that is, $R_{i+1} = [(a_i + b_i)/2, b_i]$.

The method we are going to present is a direct application of the binary search algorithm to determine a suitable test size less than or equal to the nominal value and close to it.

The proposed method

(1) Calculate the test size using the above procedure with the original α, that is, calculate $ts(\alpha)$. If $ts(\alpha) \leq \alpha$, then no adjustment is necessary because the test size is less than or equal to the nominal level (α). Thus, the process is finished and moreover $\alpha_{adj} = \alpha, ts_{adj} = ts(\alpha)$.

(2) If $ts(\alpha) > \alpha$, take $I_1 = [0, \alpha]$ and compute $ts(m_1)$.

(3) If $ts(m_1) \leq \alpha$, take $I_2 = R_2$. If $ts(m_1) > \alpha$, take $I_2 = L_2$. Compute $ts(m_2)$.

(4) Repeat the point (3) k times with $k \geq 1$. Note that in k iterations of (3), we have computed $ts(m_{k+1})$ and $|m_k - m_{k+1}| = 1/2^{k+1}$.

(5) If $ts(m_i) > \alpha$ for all $i = 1, ..., k+1$, the test sizes are always greater than the nominal value (α) and it is not possible to adjust the value of the significance level to have test size less than or equal to α.

(6) If for some i with $1 \leq i \leq k+1$, $ts(m_i) \leq \alpha$, then there are two cases. If $ts(m_{k+1}) > \alpha$, then $ts_{adj} = ts(m_k)$ and $\alpha_{adj} = m_k$. Whereas that if $ts(m_{k+1}) \leq \alpha$, then $ts_{adj} = ts(m_{k+1})$ and $\alpha_{adj} = m_{k+1}$.

Note that in the description above there is no particular restriction to any specific statistic, so this method can be applied to any asymptotical statistical test.

In the program, we fixed $k = 8$ because the difference between two consecutive adjusted nominal levels is $|m_8 - m_9| = 1/2^9 = 0.00195313$, sufficient to have a good approximation.

Analyzing the Behavior of Adjusted Values for the Classical Non-inferiority Test

To calculate the adjusted values, we can proceed in the following form:

1. Use the program written by the authors, introducing interactively the following values: sample sizes (n_1, n_2), non-inferiority margin (d_0) and the nominal significance level (α).

2. The program will show you on the screen the following results: the initial test size $(ts(\alpha))$, the adjusted nominal value (α_{adj}) and the adjusted test size (ts_{adj}) for both, T_0 and T_1.

3. Once you have obtained the value α_{adj}, apply the Blackwelder test by using the value α_{adj} as the nominal significance level, that is, instead of α. To reject or not reject the null hypotheses, it is important to note that the adjusted Blackwelder critical region is given by $\{(x_1, x_2) \in \chi : T_i(x_1, x_2) < -z_{\alpha_{adj}}\}$ for $i = 0, 1$.

To compute these adjusted values, the C++ program written by the authors can be obtained by request to the authors. This program includes numerical verification of both conditions in the Barnard convex set definition and SST condition in the case of balanced designs.

Balanced designs

To show the behavior of the test sizes, the adjusted test sizes and the adjusted nominal significance level, for balanced designs $(n_1 = n_2)$, we present two plots including the values of test size, ts_{adj} and α_{adj}. In the horizontal axis the sample size is plotted with $(n_1 = n_2)$.

Figures 1-2 correspond to T_0 and T_1, respectively, both for $d_0 = 0.10$, $\alpha = 0.05$, $n_1 = n_2$.

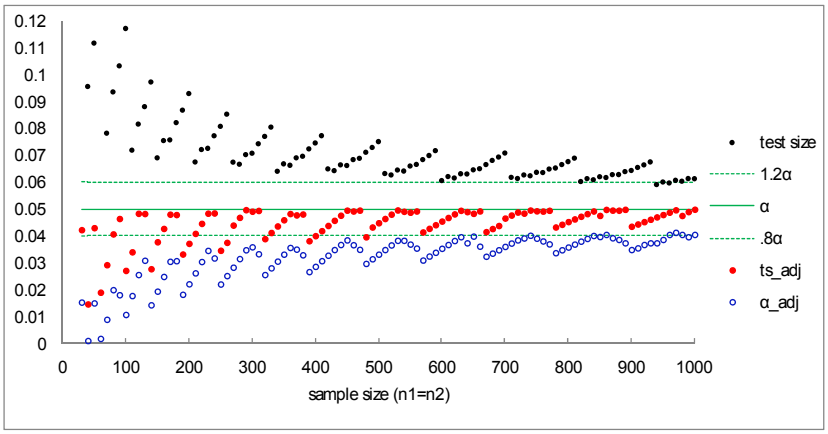

Figure 1. Test sizes, ts_{adj} and α_{adj}, for T_0, $n_1 = n_2$, $\alpha = 0.05$, $d_0 = 0.10$.

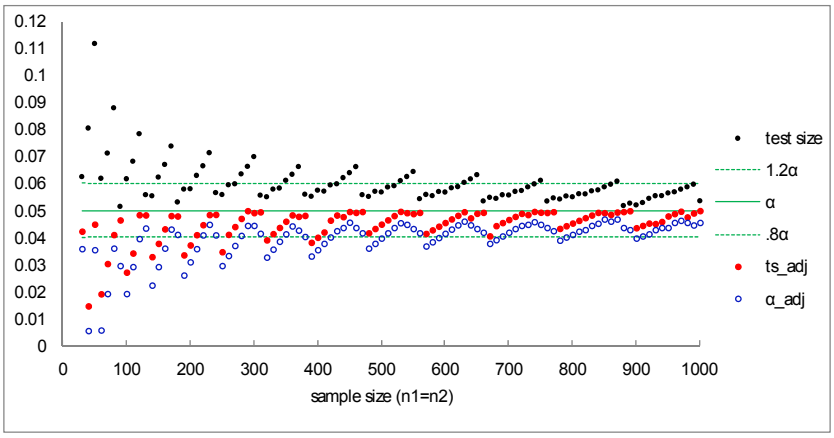

Figure 2. Test sizes, ts_{adj} and α_{adj}, for T_1, $n_1 = n_2$, $\alpha = 0.05$, $d_0 = 0.10$.

From Figures 1-2, we note that the original test size is always greater than the nominal test size and in general terms test sizes decrease when sample sizes increase; moreover it is noted that test sizes are quite inflated inclusive for large sample sizes. For example, for $n_1 = n_2 = 460$ the test sizes are 0.68399 and 0.066237 for T_0 and T_1, respectively; these values correspond to inflation of 36.80% and 32.47%, respectively. Additionally, for $n_1 = n_2 = 990$ the test sizes are 0.61344 and 0.059717 for T_0 and T_1, respectively, these values correspond to an inflation of 22.69% and 19.43%, respectively.

Test sizes were adjusted accordingly, particularly for sample sizes greater than or equal to 300. The adjustments required for the nominal level were less in the case of T_1 than in T_0 due that better performance of test sizes for T_1 and the adjusted test sizes for T_1 were better than those for T_0.

Unbalanced designs

Now, to show the behavior of the test sizes, the adjusted test sizes and the adjusted nominal significance level, for unbalanced designs we present plots including the values of test sizes, ts_{adj} and α_{adj} in the horizontal axis the sample size is plotted (n_1) with $n_2 = 1.5n_1$. Figures 3-4 correspond to T_0 and T_1, respectively, both for $d_0 = 0.10$, $\alpha = 0.05$, n_1, $n_2 = 1.5n_1$.

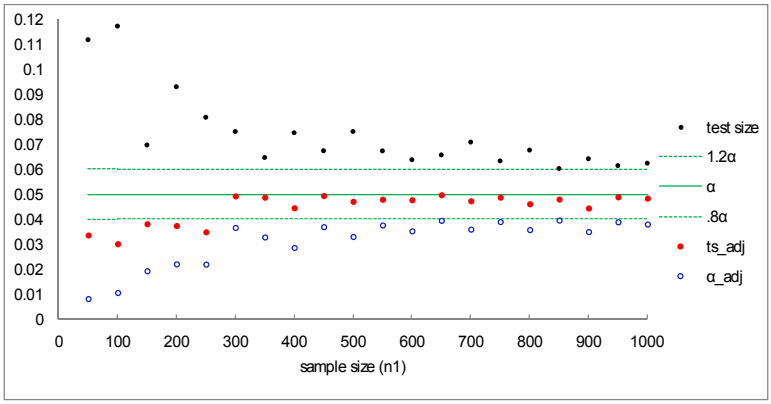

Figure 3. Test sizes, ts_{adj} and α_{adj}, for T_0, $n_2 = 1.5n_1$, $\alpha = 0.05$, $d_0 = 0.10$.

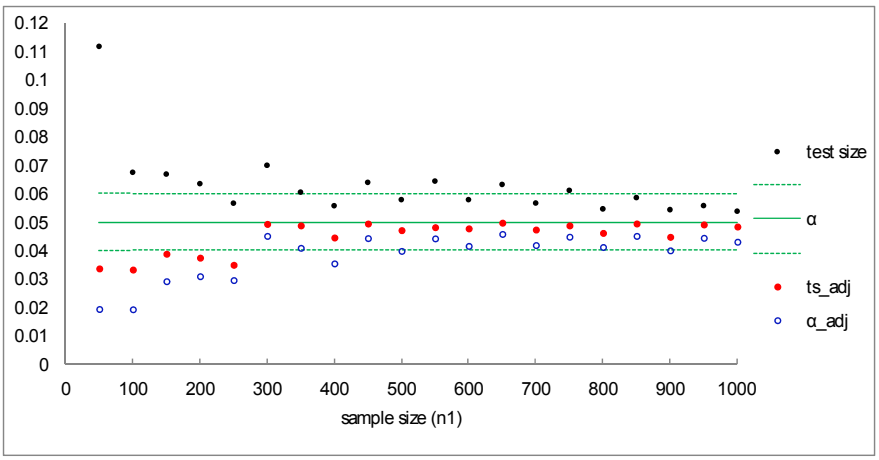

Figure 4. Test sizes, ts_{adj} and α_{adj}, for T_1, $n_2 = 1.5n_1$, $\alpha = 0.05$, $d_0 = 0.10$.

Figures 3-4 show that the original test size is always greater than the nominal test size, in general terms test sizes decrease when sample sizes increase and test sizes are quite inflated inclusive for large sample sizes. In this situation, for $n_1 = 550$, $n_2 = 825$ the test sizes are 0.067322 and 0.064421 for T_0 and T_1, respectively, corresponding inflation was of 34.64% and 28.84%, respectively.

As for the balanced case, for the unbalanced one, test sizes could be adjusted in a convenient form, especially for sample sizes greater than or equal to 300.

Also it is noted that the adjusted required for the nominal level were less in the case of T_1 than for T_0 due to better performance of test sizes for T_1 and the adjusted test sizes for T_0 were better than those for T_0.

Example

For illustration on how to use the proposed procedure, the following example is presented, in which we explore the data corresponding to a published non-inferiority trial, originally presented and analyzed by Rodary

et al. [19]. For the analysis, in each example we will apply both tests, T_0 and T_1.

To show that chemotherapy (a new treatment) is not inferior to radiotherapy (control or standard treatment), authors in Rodary et al. [19] presented a randomized clinical trial of 164 children in which the statistical analysis was based on a non-inferiority margin $d_0 = 0.10$ and a nominal significance level $\alpha = 0.05$. The chemotherapy and radiation group success response rates were 83/88 and 69/76, respectively.

Chan's asymptotic test (see Chan [9]), has the same form as T_0, the only (and very important) difference is that to estimate standard deviation in the denominator Chan's test uses as proportion estimators the maximum likelihood estimators restricted under the null hypothesis.

With Chan's asymptotic test, Chan [9] analyzed these results, for this asymptotic test, Chan [10] computed the test size obtaining the value 0.0578. That is to say, the "real" test size was 0.0578 which is 15.6% greater than the nominal value (0.05), in other words, the test is liberal as the test size is greater than the nominal value in this situation.

Below we discuss the same example using the Blackwelder test.

By running the program, we wrote for this work, we obtain directly the values presented in Table 1. As shown in Table 1, the test sizes are 0.11209 and 0.05636 for the test with and without continuity correction, respectively.

Again, we conclude that the test is liberal, that is, not conservative.

As $T_0(69, 83) = -3.2733$ and $T_1(69, 83) = -3.1132$ and as both values are less than $-z_\alpha = -1.64$, then both tests reject the null hypothesis concluding non-inferiority.

However, the above conclusion was obtained at the significance levels of 0.112087 and 0.05636 for T_0 and T_1, respectively.

Table 1. Values of test sizes for $\alpha = 0.05$, and adjusted nominal levels (α_{adj}) to obtain an adjusted test of size $ts_{adj} \leq \alpha$, for both T_0 and T_1

	α_adj	ts_adj	test size
T_0	0.01250	0.04697	0.11209
T_1	0.02500	0.04697	0.05636

Although the adjusted nominal level α_{adj} for the test without continuity correction (0.01250) is smaller than that for the test with continuity correction (0.02500), in this case both tests obtained the same adjusted test size (ts_{adj}), that is 0.04697.

In this situation, $T_0(69, 83) = -3.2733 < -z_{\alpha_{adj}} = -z_{0.0125} = -2.2414$, and $T_1(69, 83) = -3.1132 < -z_{\alpha_{adj}} = -z_{0.025} = -1.96$, therefore for both tests the conclusion is to reject the null hypothesis confirming the non-inferiority of the chemotherapy over the radiotherapy to the significance level 0.046973.

Now consider the hypothetical case in which we would obtain two less successes for the new treatment (chemotherapy) and two more successes for the reference treatment (radiotherapy), that is, that the chemotherapy and radiation group success response rates were 81/88 and 71/76, respectively.

Thus, we have that $T_0(71, 81) = -2.1291 < -z_{\alpha} = -z_{0.05} = -1.64$. Therefore, we must reject the null hypothesis with a true test size of $0.112087 = 2.24\alpha$. However, using the test T_0 with the adjusted test size 0.046973, one obtains $T_0(71, 81) = -2.1291 > -z_{\alpha_{adj}} = -z_{0.0125} = -2.2414$, and therefore in this case the null hypothesis is not rejected.

Thus, if one is not willing to accept a type I error larger than the nominal level (in fact too large), and instead accept a type I error less than the nominal level, then one should not reject the null hypothesis for this hypothetic case as was shown.

Discussion and Conclusions

Asymptotic tests usually behave in a liberal form, that is, test sizes are greater than the nominal values. In this work, we have analyzed the performance of the classical asymptotic non-inferiority test for non-inferiority for two independent proportions. The idea was to force the test to have a type I error less than or equal to the nominal level, which can be guaranteed if the test size is less than or equal to the nominal level.

In clinical investigations unbalanced designs are prevalent. Thus, because of the enormous quantity of possibilities, it is not practical to prepare tables for many different configurations of significance levels, sample sizes and non-inferiority margins.

Therefore, for practical purposes, we advise the use of the program we have written for that goal, that is, for adjusting both the test size and the nominal level. This program can be obtained directly by request to the authors.

For balanced design, we have analyzed the configurations $n_1 = n_2 = 30$, 40, ..., 1000, $\alpha = 0.025, 0.05$, $d_0 = 0.05, 0.10, 0.15$, for T_0 and T_1, whereas that for the unbalanced design we studied the configurations n_1, $n_2 = 1.5n_1$, for $n_1 = 50, 100, ..., 1000$ and n_2, $n_1 = 1.5n_2$, for $n_2 = 50, 100, ..., 1000$ for $\alpha = 0.025, 0.05$, $d_0 = 0.05, 0.10, 0.15$, or T_0 and T_1.

For all of these configurations the behavior of test sizes and adjusted values in the proposed form are similar.

In this work, we have focused on adjusting the nominal levels and test sizes. It would serve well to carry out a power study to assess the impact of these adjustments applied in the power of the test, however for reasons of anticipated length and space concerns, we postpone such analysis.

The proposed method is based on the computational binary search algorithm which is a deceptively simple algorithm, however as in each

iteration the algorithm eliminates half of the remaining possibilities, making binary searches very efficient. Specifically, for *k* iterations, we have an error less than or equal to $1/2^{k+1}$. In this work, we took $k = 8$ consequently the error in determining the adjusted nominal levels is less than or equal to 0.00195313.

Finally, note that the topic exposed in this article provides practical application to previous works and the ideas exposed in this paper can be applied straightforward to any asymptotic non-inferiority test for two independent proportions, for in doing so, it is enough to change the statistic under consideration and to adapt the domain of *p* in formula (3) to the respective statistic.

Acknowledgments

The first author would like to thank SNI-CONACyT, EDI-IPN, COFAA-IPN and project SIP-IPN 20160687 for their partial support. The authors are very grateful for all of the valuable suggestions and comments provided by the unknown reviewers.

References

[1] F. Almendra-Arao, A study of the classical noninferiority test for two binomial proportions, Drug Information Journal 43 (2009), 567-572.

[2] F. Almendra-Arao, Barnard convex sets, Communications in Statistics - Theory and Methods 40(14) (2011), 2574-2582.

[3] F. Almendra-Arao, Efficient calculation of test sizes for non-inferiority, Computational Statistics and Data Analysis 56 (2012), 4138-4145.

[4] F. Almendra-Arao, A new non-inferiority test for independent dichotomous variables based on a shrinkage proportion estimator, Journal of Biopharmaceutical Statistics 25(1) (2015), 157-169.

[5] F. Almendra-Arao, J. J. Castro-Alva and H. Reyes-Cervantes, Convergence of test sizes for the Blackwelder's non-inferiority test, Advances and Applications in Statistics 31(1) (2013), 15-31.

[6] F. Almendra-Arao and D. Sotres-Ramos, Some properties of non-inferiority tests for two independent probabilities, Communications in Statistics - Theory and Methods 41 (2012), 1636-1646.

[7] F. Almendra-Arao and Sotres-Ramos, On the requirement that critical regions for comparing two independent proportions must be Barnard convex sets, Therapeutic Innovation and Regulatory Science 48(2) (2014), 208-212.

[8] W. Blackwelder, Proving the null hypothesis" in clinical trials, Controlled Clinical Trials (1982), 345-353.

[9] I. S. F. Chan, Exact tests of equivalence and efficacy with a non zero lower bound for comparative studies, Stat. Med. 17 (1998), 1403-1413.

[10] I. S. F. Chan, Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods, Statistical Methods in Medical Research 12 (2003), 37-58.

[11] J. Chen, Y. Tsong and S. Kang, Tests for equivalence or non-inferiority between two proportions, Drug Information Journal 34 (2000), 569-578.

[12] C. W. Dunett and M. Gent, Significance testing to establish equivalence between drugs with special reference to data in the form 2×2 tables, Biometrics 33 (1977), 593-602.

[13] C. Farrington and G. Manning, 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 (1990), 1447-1454.

[14] W. Hauck and S. Anderson, A comparison of large-sample confidence interval methods for the difference of two binomial probabilities, The American Statistician 40 (1986), 318-322.

[15] Z. Li and C. Chuang-Stein, A note on comparing two binomial proportions in confirmatory non-inferiority trials, Drug Information Journal 40 (2006), 203-208.

[16] A. A. Martin and T. I. Herranz, Asymptotical test on the equivalence, substantial difference and non-inferiority problems with two proportions, Biometrical Journal 46 (2004a), 305-319.

[17] A. A. Martin and T. I. Herranz, Exact unconditional non-classics tests on the difference of two proportions, Comput. Statist. Data Anal. 45 (2004b), 373-388.

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

[19] C. Rodary, C. Com-Nougue and M. F. Tournade, How to establish equivalence between treatments: a one-sided clinical trial in pediatric oncology, Statistics in Medicine 8 (1989), 593-598.

[20] J. Röhmel and U. Mansmann, Unconditional non-asymptotic one sided tests for independent binomial proportions when the interest lies in showing non-inferiority and or superiority, Biometrical Journal 2 (1999), 149-170.

[21] D. Tu, A comparative study of some statistical procedures in establishing therapeutic equivalence of nonsystemic drugs with binary endpoints, Drug Information Journal 31 (1997), 1291-1300.