Advances and Applications in Statistics
Volume 32, Number 1, 2013, Pages 15-31
Published Online: January 2013
Available online at http://pphmj.com/journals/adas.htm
Published by Pushpa Publishing House, Allahabad, INDIA

CONVERGENCE OF TEST SIZES FOR THE BLACKWELDER'S NON-INFERIORITY TEST

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

[1]Departamento de Ciencias Básicas
UPIITA del IPN, México, D.F., México

[2]Facultad de Ciencias Físico Matemáticas
BUAP, Pue., Puebla, México

Abstract

The performance of two often used statistical procedures - the classical asymptotic normal approximation and the same method with the Hauck-Anderson continuity correction - to test non-inferiority between two proportions was studied, recently [1]. That study evaluates the performance of these two methods calculating the test sizes by enumerating all possible cases rather than through simulation, the hypothesis tests approach was used and was done for sample sizes until 300; the main conclusion in that work is that for these sample sizes, behavior of test sizes is erratic and uncontrolled, and its value is nearly always far above the nominal significance level. In that order of ideas, we consider important to know the performance of test sizes for big sample sizes $(n \geq 300)$. That is, we extended this analysis in this new research, having as our main goal to study numerically the performance of test sizes for these two statistical procedures, but now considering big sample sizes. Due to the fact the involved test sizes are

2010 Mathematics Subject Classification: 62F05.

Keywords and phrases: non-inferiority test, Blackwelder test, significance level, binomial proportions, hypothesis test.

Received October 4, 2012

big, some new additional computational difficulties were presented in this work, these difficulties were solved by implementing some theoretical properties in a program we wrote, in C++, to compute test sizes, especially by incorporating the results obtained in [2].

Introduction

Non-inferiority tests are statistical procedures designed with the objective of determining if a new treatment is superior, equal or inferior, by a generally small margin to a treatment considered as standard. These tests are often used in clinical trials in which generally the new treatment must present at least an advantage as for example: cheaper, easier to apply or to have less contraindications. Among a great quantity of non-inferiority tests for two proportions [3-12], the classical asymptotic or Blackwelder test has an outstanding role because it is used very often in practice, probably due to its simplicity.

Asymptotic tests have the problem that not necessarily respect the nominal significance level for which they were constructed, for that reason is important to study the performance of test sizes.

On the other hand, test sizes calculation for non-inferiority test is a computationally intensive problem due to the presence of a nuisance parameter.

Li and Chuang-Stein [13] 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 that investigation but doing an exact calculation of type I errors and test sizes instead of estimation by simulation. As conclusions of work in [1] is that behavior of test sizes is erratic and uncontrolled; 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, it is natural to ask if for big sizes the behavior is more acceptable or not.

To answer this inquiry, in the present article we made a numerical study of the behavior of test sizes for these two noninferiority statistical procedures. This new analysis is based on sample sizes $n_1 = n_2$ 30, 40, 50, ..., 1000 nominal significance levels 0.025 and 0.05 and non-inferiority margins 0.05, 0.10, 0.15, 0.20; we also consider unbalanced designs with sample sizes $(n_1, n_2 = 1.5n_1)$ with $n_1 = 50, 100, 150, ..., 1500$; and $(n_1 = 1.5n_2, n_2)$ with $n_2 = 50, 100, 150, ..., 1500$.

In order to be able to carry out calculations of test sizes for big sample sizes in a reasonable time, it was necessary to solve several numerical difficulties, as will be seen below. These difficulties were solved by using recommendations in [2].

Classical Asymptotic Non-inferiority Test

The notation we will use in this paper is the same as in Almendra-Arao [1].

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 and new drug, respectively. And consider the hypothesis testing problem

$$H_0 : p_1 - p_2 \geq d_0 \text{ vs. } H_a : p_1 - p_2 < d_0, \tag{1}$$

where d_0 is a positive known constant. d_0 is referred to as noninferiority margin. The corresponding sample space in this situations is $\chi = \{0, ..., n_1\} \times \{0, ..., n_2\}$ and the parameter space can be conveniently represented as $\Theta = \{(p_1, p_2) \in [0, 1]^2\}$.

The Blakwelder's or classical statistic to test (1) is

$$T_0(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\hat{\sigma}}, \tag{2}$$

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 in (2) 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 a 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 the Hauck-Anderson continuity correction (cc), which will be used in what follows.

Thus, we have the another test, which consider this cc, its statistic is given by

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

We will study in this work the tests T_1, and T_0. For the test T_1, we will use a similar notation as used for T_0.

Test Sizes Calculation

Let T used to denote either tests, T_0 or T_1.

Thus, we have that 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

$$\beta_T(p_1,\ p_2) = \sum_{(x_1,\, x_2)\in R_T(\alpha)} L(p_1,\ p_2;\ x_1,\ x_2).$$

Therefore, test size is given by $\sup_{\theta\in\Theta_0} \beta_T(p_1,\ p_2)$, where $\Theta_0 = \{(p_1,\ p_2) \in \Theta : p_1 - p_2 \geq d_0\}$ is the null space.

Since that 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 are 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 was shown in [1], the above redefinition of $\hat{\sigma}$ is essential for the critical regions of the statistical tests to conform the following two definitions that are necessary to reduce computation of test sizes.

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

Calculation of test sizes for non-inferiority is not a trivial problem, considered from the computational point of view. Hence in order to do these calculations for big sample sizes it is necessary to have on hand theoretical results that allows us for reduction of this time.

Thus, several theoretical results have been established aimed at reduce computational time to calculating test sizes for non-inferiority [2, 11, 12, 14-16], these results imposed by convenience that critical regions be Barnard

convex sets. However, that critical regions for non-inferiority tests be Barnard convex sets it is not only a convenient condition, it is actually necessary in order that non-inferiority test be coherent.

Another condition that it is useful fulfils critical regions for balanced designs it is defined in what 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 if $(x_1, x_2) \in R_T \Rightarrow (n - x_2, n - x_1) \in R_T$.

The calculation of test sizes was carried out following the "procedure used to calculate test sizes" presented in Almendra-Arao [2]. We made these calculations by using a program we wrote in C++ programming language. This program can be obtained from the second author on request.

Results

Balanced designs

In Figures 1 to 4, for $\alpha = 0.025$ are presented test sizes for different configurations for balanced designs.

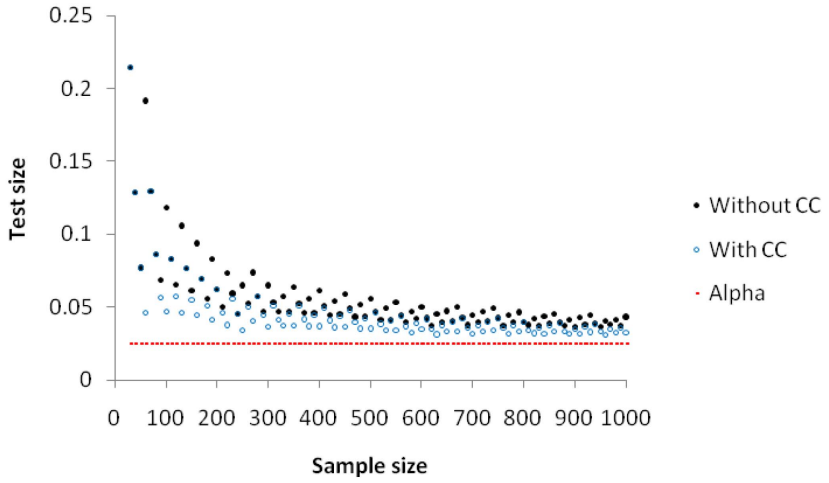

Figure 1. Test sizes for $\alpha = 0.25$, $d_0 = 0.05$ and balanced designs $(n_1 = n_2 = n)$ for $n = 30, 40, 50, ..., 1000$.

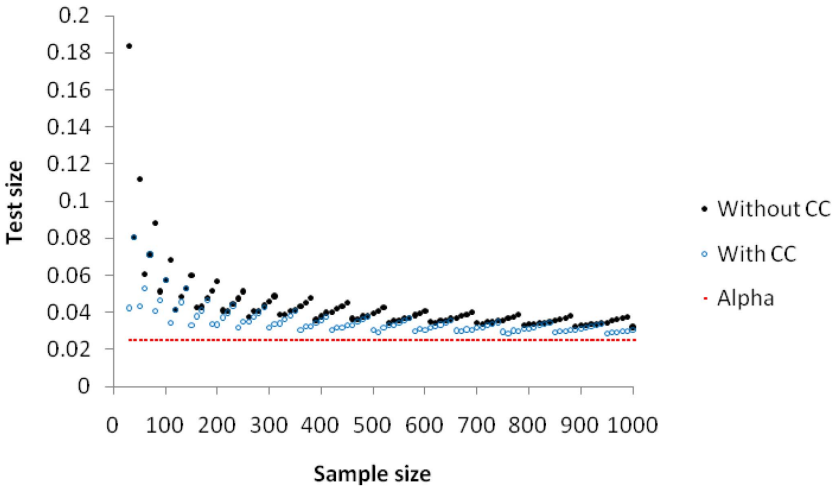

Figure 2. Test sizes for $\alpha = 0.25$, $d_0 = 0.10$ and balanced designs $(n_1 = n_2 = n)$ for $n = 30, 40, 50, ..., 1000$.

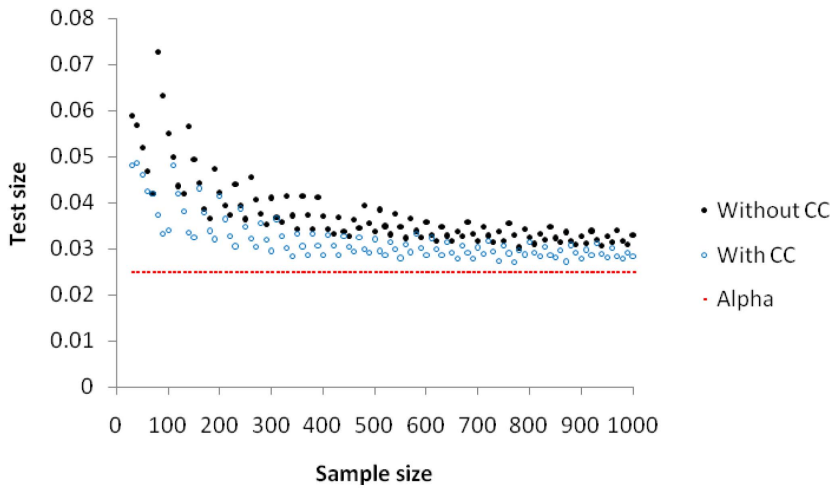

Figure 3. Test sizes for $\alpha = 0.25$, $d_0 = 0.15$ and balanced designs $(n_1 = n_2 = n)$ for $n = 30, 40, 50, ..., 1000$.

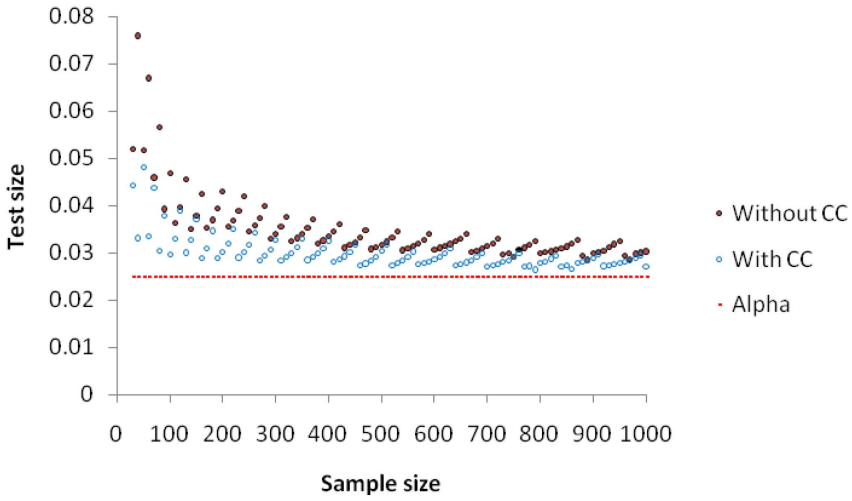

Figure 4. Test sizes for $\alpha = 0.25$, $d_0 = 0.20$ and balanced designs $(n_1 = n_2 = n)$ for $n = 30, 40, 50, ..., 1000$.

In Figures 5 to 8, for balanced designs, test sizes are presented for $\alpha = 0.05$ and for several configurations.

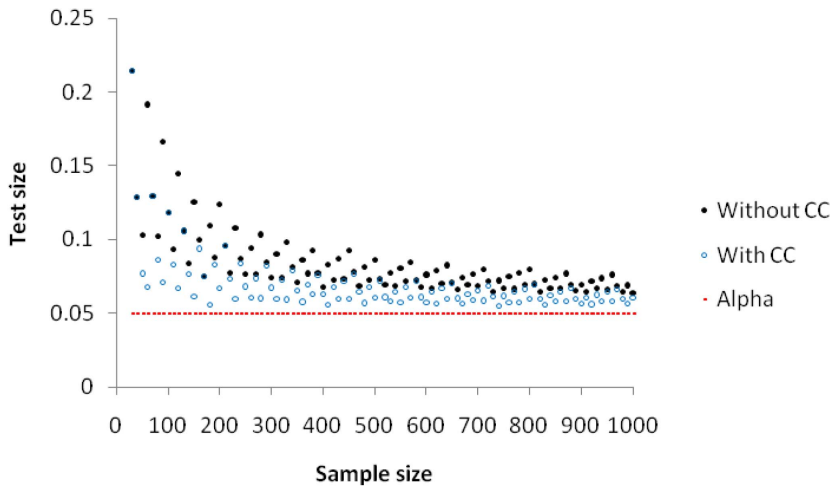

Figure 5. Test sizes for $\alpha = 0.05$, $d_0 = 0.05$ and balanced designs $(n_1 = n_2 = n)$ for $n = 30, 40, 50, ..., 1000$.

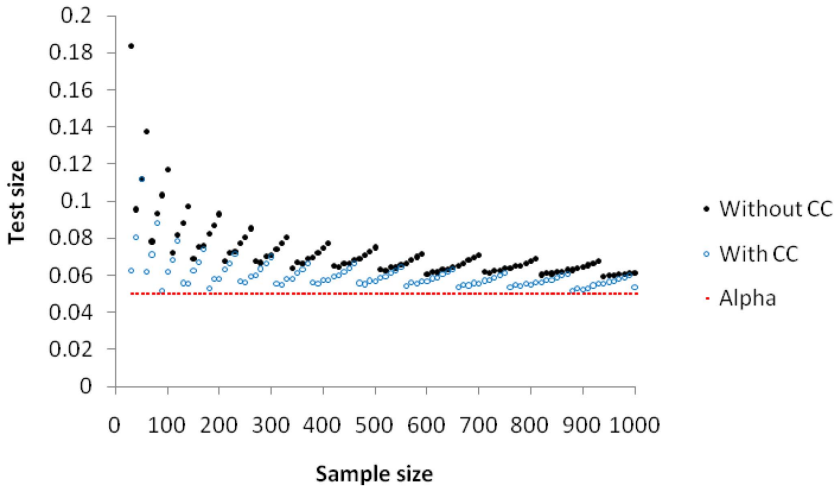

Figure 6. Test sizes for $\alpha = 0.05$, $d_0 = 0.10$ and balanced designs $(n_1 = n_2 = n)$ for $n = 30, 40, 50, ..., 1000$.

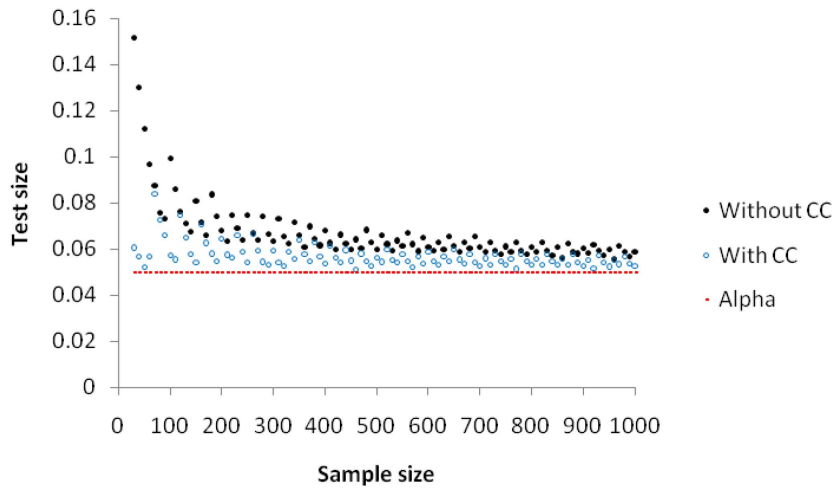

Figure 7. Test sizes for $\alpha = 0.05$, $d_0 = 0.15$ and balanced designs $(n_1 = n_2 = n)$ for $n = 30, 40, 50, ..., 1000$.

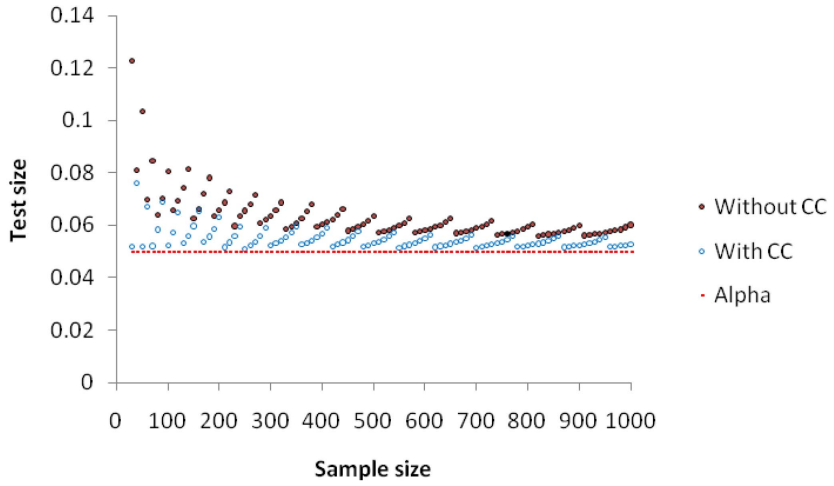

Figure 8. Test sizes for $\alpha = 0.05$, $d_0 = 0.20$ and balanced designs $(n_1 = n_2 = n)$ for $n = 30, 40, 50, ..., 1000$.

In Figures 1 to 8, it is noted that in all cases that test sizes are above the nominal significance level and uncontrolled.

In Table 1 are presented the percentages of test sizes that belong to the interval indicated. For example, the value 87.76 in the fourth column corresponds to the percentage of test sizes that belong to the interval $[1.2\alpha, 1]$ for $\alpha = 0.025$ and $d = 0.10$. Also, the highest values for each case are highlighted in bold.

Table 1. Percentage of test sizes belonging to specified intervals

α		0.025		0.05	
d_0	Interval	Test Size		Test Size	
		T_0	T_1	T_0	T_1
0.05	[0, α]	0.00	0.00	0.00	0.00
	(α, 1.2α)	0.00	0.00	0.00	30.61
	[1.2α, 1]	**100.00**	**100.00**	**100.00**	**69.39**
0.1	[0, α]	0.00	0.00	0.00	0.00
	(α, 1.2α)	0.00	12.24	2.04	**67.35**
	[1.2α, 1]	**100.00**	**87.76**	**97.96**	32.65
0.15	[0, α]	0.00	0.00	0.00	0.00
	(α, 1.2α)	0.00	43.88	28.57	**85.71**
	[1.2α, 1]	**100.00**	**56.12**	**71.43**	14.29
0.2	[0, α]	0.00	0.00	0.00	0.00
	(α, 1.2α)	10.20	**67.35**	47.96	**93.88**
	[1.2α, 1]	**89.80**	32.65	**52.04**	6.12

Unbalanced designs

In Figures 9 to 11, for unbalanced designs with $n_1 = n_2 = 1.5n_1$ for $n_1 = 50, 100, 150, ..., 1500,$ test sizes are presented for $\alpha = 0.025$ and for different configurations.

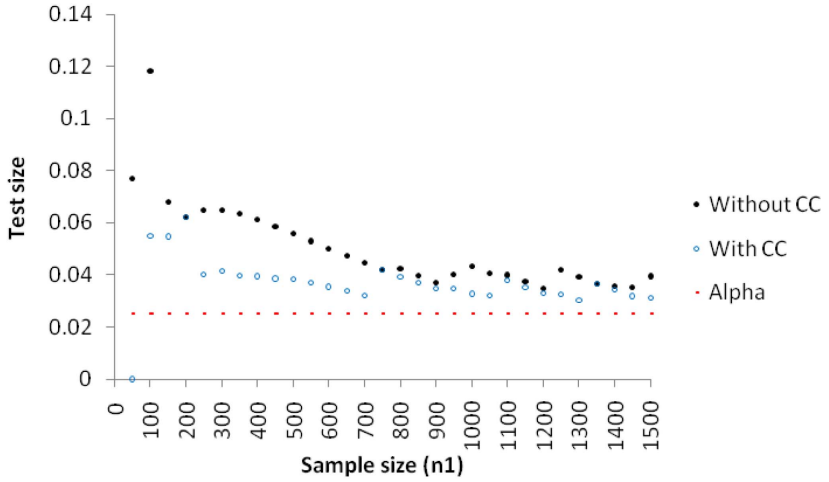

Figure 9. Test sizes for $\alpha = 0.025$, $d_0 = 0.05$ and unbalanced designs $(\text{for } n_1,\ n_2 = 1.5n_1)$ for $n_1 = 50, 100, 150, ..., 1500$.

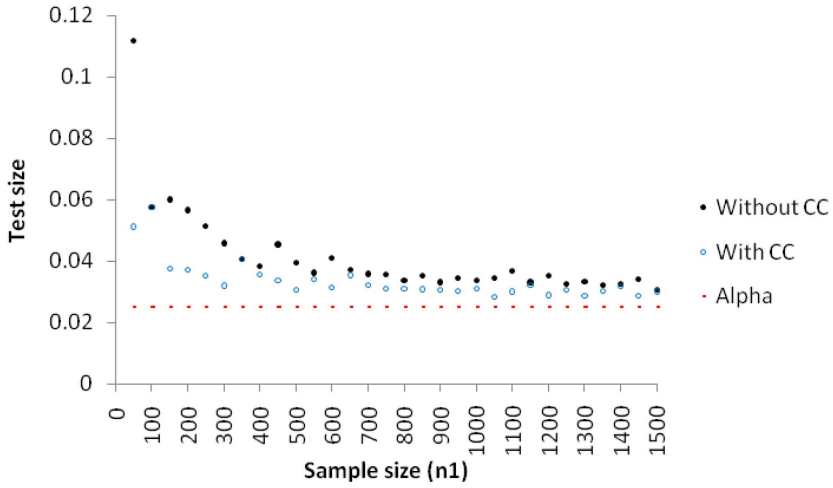

Figure 10. Test sizes for $\alpha = 0.025$, $d_0 = 0.10$ and unbalanced designs $(\text{for } n_1,\ n_2 = 1.5n_1)$ for $n_1 = 50, 100, 150, ..., 1500$.

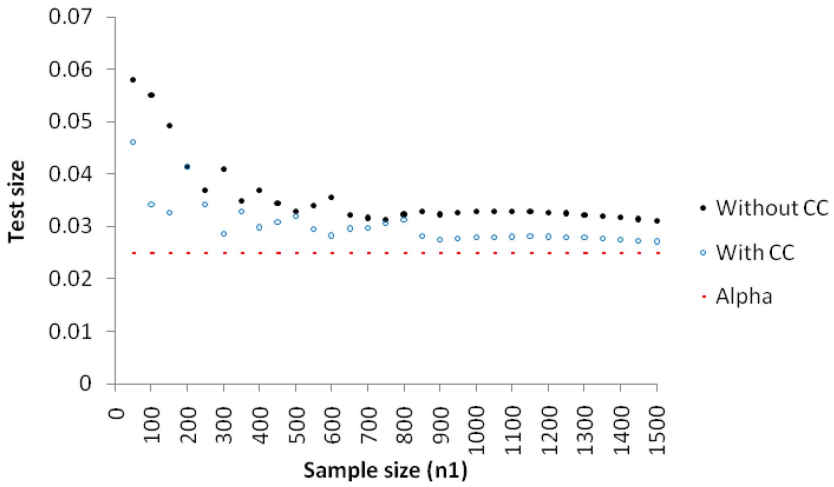

Figure 11. Test sizes for $\alpha = 0.025$, $d_0 = 0.15$ and unbalanced designs $(\text{for } n_1,\ n_2 = 1.5n_1)$ for $n_1 = 50, 100, 150, ..., 1500$.

In Figures 12 to 14, for unbalanced designs with $n_1,\ n_2 = 1.5n_1$ for $n_1 = 50, 100, 150, ..., 1500$ test sizes are presented for $\alpha = 0.05$ and for different configurations.

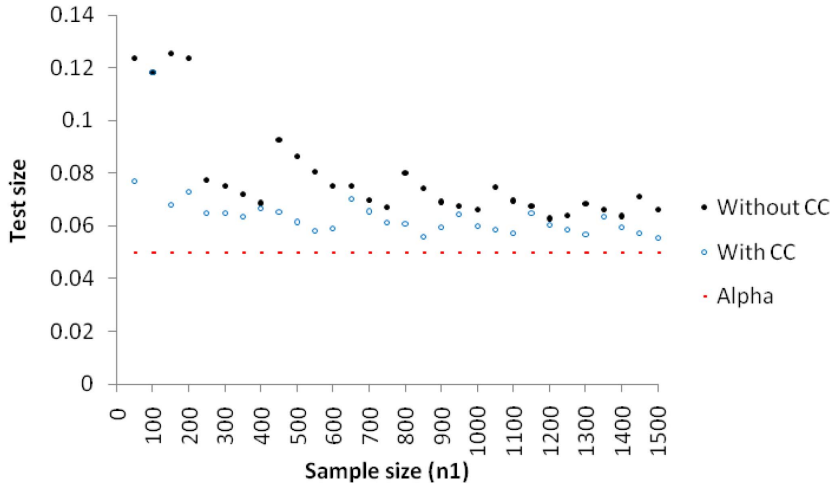

Figure 12. Test sizes for $\alpha = 0.05$, $d_0 = 0.05$ and unbalanced designs $(\text{for } n_1,\ n_2 = 1.5n_1)$ for $n_1 = 50, 100, 150, ..., 1500$.

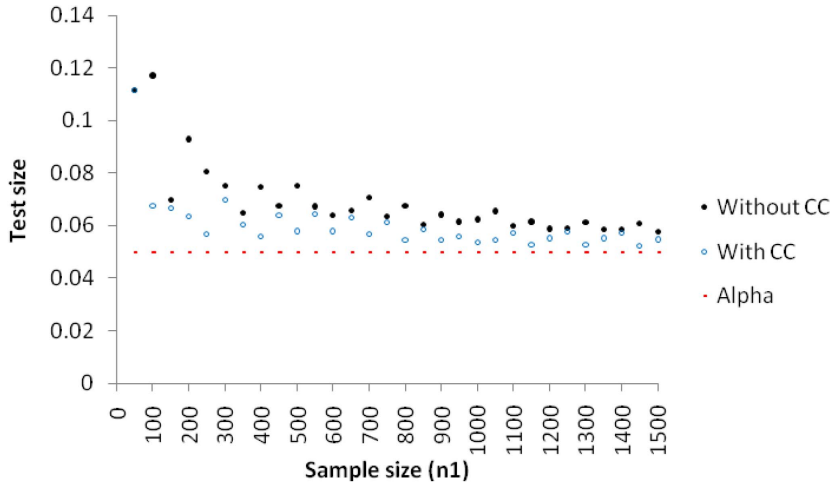

Figure 13. Test sizes for $\alpha = 0.05$, $d_0 = 0.10$ and unbalanced designs $(\text{for } n_1,\ n_2 = 1.5n_1)$ for $n_1 = 50, 100, 150, ..., 1500$.

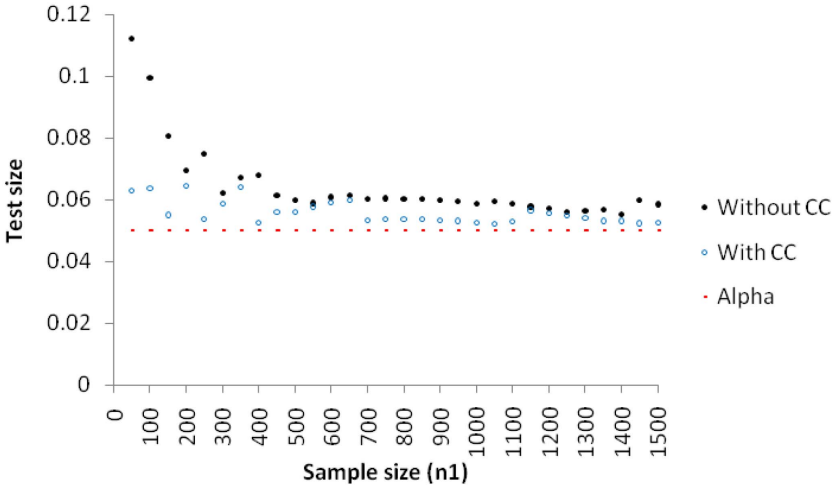

Figure 14. Test sizes for $\alpha = 0.05$, $d_0 = 0.15$ and unbalanced designs $(\text{for } n_1,\ n_2 = 1.5n_1)$ for $n_1 = 50, 100, 150, ..., 1500$.

In Figures 9 to 14, it is noted that in all cases that test sizes are uncontrolled and above the nominal significance level.

Conclusions

Analyzing the results presented in Table 1, which was established for balanced design, we note that in no case the test size belongs to the interval $[0, \alpha]$, this means that these tests do not preserve test size for configurations considered for balance designs. Also, Table 1 shows that the minimum percentage of significance levels belonging to the interval $(\alpha, 1.2\alpha)$ is 0%, and the maximum is 93.88%; this maximum value is obtained for T_1 when $\alpha = 0.05$, and $d_0 = 0.20$. While at the same time, the minimum percentage for the interval $[1.2\alpha, 1]$ is 6.12% and it is reached by the statistic T_1 when $\alpha = 0.05$ and $d_0 = 0.20$, and the maximum is 100%.

Thus, for balanced designs, from results of Table 1 and Figures 1 to 8, we obtain similar conclusions to those attained by Almendra-Arao [1], that is, that the behavior of test sizes is liberal, erratic and uncontrolled, and its value is nearly always far above from the nominal significance level.

For unbalanced case, from Figures 9 to 14, we have similar conclusions as those for balanced case, in other words, also for unbalanced designs tests behave in liberal, erratic and uncontrolled way.

Previous conclusions mean that although it is known theoretically that test sizes converges to nominal significance level, this convergence is very slow. In other words, convergence of test sizes to nominal significance level is not sufficiently rapid as to guarantee that nominal significance level is respected for sample sizes as big as 1000, for balanced designs, and for sample sizes as big as 1500, for unbalanced designs.

Thus, we can say that although the analyzed tests are very simple to use, they are not ready to be used in practice in the way that they are.

With the aim of making these tests adequate for their use in practice, we think that future investigation can be carried out to determine the nominal significance level to be specified to obtain a given objective significance level (α).

Acknowledgements

First author wants to thank to SNI-CONACyT, EDI-IPN, COFAA-IPN and Project SIP-IPN 20110192 for their partial support.

References

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

[2] F. Almendra-Arao, Efficient calculation of test sizes for noninferiority, Computational Statistics and Data Analysis (in press).

[3] 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.

[4] W. Blackwelder, Proving the null hypothesis in clinical trials, Controlled Clinical Trials 3 (1982), 345-353.

[5] O. Miettinen and M. Nurminen, Comparative analysis of two rates, Statist. Med. 4 (1985), 213-226.

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

[7] 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. Statist. Med. 9 (1990), 1447-1454.

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

[9] J. Chen, Y. Tsong and S. Kang, Tests for equivalence or noninferiority between two proportions, Drug Inf. J. 34 (2000), 569-578.

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

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

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

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

[14] 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 J. 2 (1999), 149-170.

[15] H. Frick, Undominated *p*-values and property *C* for unconditional one-sided two sample binomial tests, Biometrical J. 42 (2000), 715-728.

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