RESEARCH ARTICLE

Exact Power and Sample Size Calculations for the Two One-Sided Tests of Equivalence

Gwowen Shieh*

Department of Management Science, National Chiao Tung University, Hsinchu, 30010, Taiwan

* gwshieh@mail.nctu.edu.tw

Abstract

Equivalent testing has been strongly recommended for demonstrating the comparability of treatment effects in a wide variety of research fields including medical studies. Although the essential properties of the favorable two one-sided tests of equivalence have been addressed in the literature, the associated power and sample size calculations were illustrated mainly for selecting the most appropriate approximate method. Moreover, conventional power analysis does not consider the allocation restrictions and cost issues of different sample size choices. To extend the practical usefulness of the two one-sided tests procedure, this article describes exact approaches to sample size determinations under various allocation and cost considerations. Because the presented features are not generally available in common software packages, both R and SAS computer codes are presented to implement the suggested power and sample size computations for planning equivalence studies. The exact power function of the TOST procedure is employed to compute optimal sample sizes under four design schemes allowing for different allocation and cost concerns. The proposed power and sample size methodology should be useful for medical sciences to plan equivalence studies.

Citation: Shieh G (2016) Exact Power and Sample Size Calculations for the Two One-Sided Tests of Equivalence. PLoS ONE 11(9): e0162093. doi:10.1371/journal.pone.0162093

Editor: Jake Olivier, University of New South Wales, AUSTRALIA

Received: March 22, 2016

Accepted: August 17, 2016

Published: September 6, 2016

Data Availability Statement: Data are from the Rogers JL, Howard KI, Vessey JT (1993) Using significance tests to evaluate equivalence between two experimental groups. Psychological Bulletin 113: 553–565.

Funding: The author received no specific funding for this work.

Competing Interests: The author has declared that no competing interests exist.

Introduction

Equivalence tests have been widely adopted for demonstrating the bioequivalence between two drug formulations in biopharmaceutical studies. The notion of equivalence between treatment effects is equally relevant and potentially useful in other of research fields such as medical sciences. Although it is not the uniformly most powerful test and more powerful tests exist, the two one-sided tests (TOST) procedure proposed by Schuirmann [1] and Westlake [2] is the most common method for equivalence assessment under a two-group parallel design. A comprehensive review of the different types of equivalence tests was presented in Meyners [3]. Further details on the design and analysis of equivalence studies can be found in Chow and Liu [4], Chow, Shao, and Wang [5], Hauschke, Steinijans, and Pigeot [6], and Wellek [7].

It should be noted that the logic of the traditional difference-based tests and the formal equivalence-based tests are fundamentally distinct. Rogers et al. [8] emphasized that the traditional test and the equivalence test are not mutually exclusive. If both test procedures are performed, it is possible that both will be rejected, that neither will be rejected, or that one will be rejected and the other will not be rejected. Therefore, failing to reject a no-difference hypothesis test does not necessarily support the conclusion of equivalence as was stressed in Blackwelder [9]. Also, Cribbie, Gruman, and Arpin-Cribbie [10], Parkhurst [11], and Schuirmann [12] conducted comprehensive comparisons about the intrinsic appropriateness and theoretical properties between the TOST procedure and the two-sample *t* test for assessing the equivalence of two treatment means. More importantly, Allan and Cribbie [13] emphasized that the traditional tests are often inappropriately applied to establish equivalence in the psychological literature.

To improve the underutilized situation, the TOST procedure is strongly recommended, instead of the two-sample *t* test, when the research objective is to determine whether two treatment means are sufficiently near each other to be considered equivalent. The theoretical justification and computational ease are vital features of the TOST procedure for making statistical inferences. However, an empirical study requires adequate statistical power and sufficient sample size to detect designated hypotheses and examine research questions. The corresponding power calculations and sample size determinations must also be considered for a viable procedure to extend the applicability in planning research designs. Accordingly, considerable attention has been devoted to the power and sample size issues of the TOST procedure in the literature. Because the power function of the TOST is complicated in form, various expressions, approximations and computing algorithms have been proposed and discussed from different perspectives. The key findings are documented in Bristol [14], Chow, Shao, and Wang [15], Chow and Wang [16], Diletti, Hauschke, and Steinijans [17], Liu and Chow [18], Muller-Cohrs [19], Phillips [20], Schuirmann [12], Siqueira, Whitehead, Todd, and Lucini [21], and Wang and Chow [22], among others. It is essential to note that the inferential procedure and theoretical property of the TOST under a two-group parallel design immediately extend to the two-sequence and two-period crossover designs and the replicated crossover designs as explicated in Chow, Shao, and Wang [15], Chow and Wang [16], Siqueira et al. [21], and Wang and Chow [22].

Although the desirable properties of the two one-sided tests of equivalence, including the exact power function, have been well documented in the literature, the associated power and sample size calculations were illustrated mainly for selecting the most appropriate approximate method. At first sight, the approximate power functions are comparatively easy to use and seem to give practically useful results. But it does not retain all of the critical characteristics of the model configurations, and thus, there is no guarantee that the resulting sample size techniques will always give reliable performance. It was noted in Siqueira et al. [21] that simple approximations are satisfactory under certain conditions and the difficulty of calculating the sample size is not necessarily reduced by using the approximate formulas. On the other hand, with the advance of computer technology and the general availability of statistical software, computational simplicity is no longer a primary focus. Most importantly, the superiority of exact techniques in terms of accuracy is irreplaceable. Therefore, the exact power and sample size calculations should be considered instead. It is prudent to note that Bristol [14] and Schuirmann [12] described a particularly attractive and convenient expression for the exact power function of the TOST that can be readily implemented with the embedded normal and chi-square distribution functions in standard software systems.

Among others, Jan and Shieh [23] noted that conventional power analysis and sample size determination do not address matters of allocation restrictions and cost issues. However, researchers have been exploring design strategies that accommodate different constraints of the allocation structure and project funding while maintaining adequate power. Specifically, the allocation ratio of group sizes was fixed in the calculation of sample size for examining independent proportions in Fleiss, Tytun and Ury [24], while Heilbrun and McGee [25] considered sample size problem for the comparison of normal means when one sample size is specified in advance. Moreover, in an actual experiment, the available resources are generally limited and the cost for treating a subject often varies with treatment groups. For example, Nam [26] presented optimal sample sizes to maximize power for the comparison of the treatment and control under budget constraints. In contrast, Allison et al. [27] advocated designing statistically powerful studies while minimizing costs.

In view of the insufficient consideration on the exact sample size methodology in the literature, the present article aims to contribute to the development of optimal sample size determinations for the design of equivalence studies in two ways. First, the exact power function of the TOST procedure is employed to compute optimal sample sizes under four design schemes allowing for different allocation and cost concerns. The allocation schemes include (a) the ratio of group sizes is given, and (b) one sample size is specified. Moreover, the cost implications suggest optimally assigning subjects (a) to attain maximum power performance for a fixed cost, and (b) to meet a designated power level for the least cost. Second, because existing software packages do not accommodate power and sample size considerations with the same degree of generality as described in this article, computer algorithms are developed to facilitate the implementation of the suggested procedures. The proposed power and sample size methodology should be useful for medical sciences to plan equivalence studies.

Methods

Two one-sided tests procedure

Consider independent random samples from two normal populations with the following formulations:

$$X_{ij} \sim N(\mu_i, \sigma^2), \tag{1}$$

where μ_i, σ^2 are unknown parameters, $j = 1, \ldots, N_i$, and $i = 1$ and 2. For detecting the group effect $\mu_d = \mu_1 - \mu_2$ in terms of the hypothesis H_0: $\mu_d = 0$ versus H_1: $\mu_d \neq 0$, the common two-sample t statistic has the form

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S^*}, \tag{2}$$

where $\overline{X}_1 = \sum_{j=1}^{N_1} X_{1j}/N_1,\ \overline{X}_2 = \sum_{j=1}^{N_2} X_{2j}/N_2,\ S^{*2} = S^2(1/N_1 + 1/N_2),\ \ S^2 = \{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2\}/\nu,\ S_1^2 = \sum_{j=1}^{N_1} (X_{1j} - \overline{X}_1)^2/(N_1 - 1),\ \ S_2^2 = \sum_{j=1}^{N_2} (X_{2j} - \overline{X}_2)^2/(N_2 - 1)$, and $\nu = N_1 + N_2 - 2$.

The primary focus of this article is the test of equivalence and without loss of generality, the null and alternative hypotheses are expressed as

$$\mathrm{H}_0 : \mu_d \leq -\Delta \text{ or } \mu_d \geq \Delta \text{ versus } \mathrm{H}_1 : -\Delta < \mu_d < \Delta, \tag{3}$$

where Δ (> 0) is a priori constant that represents the minimal difference for declaring equivalent means. It follows from the TOST procedure proposed by Schuirmann [1] and Westlake [2] that the null hypothesis is rejected at the significance level α if

$$T_1 = \frac{\overline{X}_1 - \overline{X}_2 + \Delta}{S^*} > t_{\nu,\alpha} \text{ and } T_2 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta}{S^*} < -t_{\nu,\alpha}, \tag{4}$$

where $t_{\nu,\alpha}$ is the upper 100·α-th percentile of the t distribution with degrees of freedom ν.

The exact power function Ψ_E of the TOST procedure is presented in Equation A2 of S1 File. The numerical computation of exact power requires the evaluation of the cumulative distribution function of a standard normal variable and the one-dimensional integration with respect to a chi-square probability distribution function. Since all related functions are presented in major statistical packages, the exact computations can be conducted with current computing systems. For advance planning of equivalence studies, the presented power function Ψ_E can be employed to calculate the sample sizes $\{N_{1E}, N_{2E}\}$ needed to attain the specified power $(1 - \beta)$ for the chosen significance level α, the model configurations $\{\mu_d, \sigma^2\}$, and the equivalence threshold Δ. In order to enhance the applicability of the TOST procedure, optimal sample size algorithms are described in the subsequent section for four design schemes under different allocation and cost considerations. The R [28] and SAS/IML [29] programs employed to perform the corresponding sample size calculations are available in S3 File and S4 File, respectively. While the optimal power and sample size considerations are primarily illustrated for the equivalence evaluations of two-group parallel designs, they may be directly extended to equivalence problems of two-sequence and two-period crossover designs as explicated in S2 File.

For illustration, simulation study was conducted to demonstrate the suggested exact approach for power and sample size calculations. The empirical assessment examines the two mean difference patterns and six standard deviation values presented in Table V of Siqueira et al. [21]. Specifically, the two sets of mean differences and standard deviations are $\mu_d = \{0.0, 0.1\}$ and $\sigma = \{0.10, 0.12, 0.14, 0.16, 0.18, 0.20\}$, respectively. Also, the chosen sample sizes were determined to achieve the power level 0.80 with the gold standard method reported in Siqueira et al. [21]. They only considered balanced design with the sample sizes $N_1 = N_2 = N$ and the gold standard method is an approximation with the power function Ψ_A given in Equation A5 of S1 File. Accordingly, the exact power function Ψ_E presented in Equation A3 is employed to compute the attained power for the twelve model configurations with $\alpha = 0.05$ and $\Delta = 0.2231$.

Moreover, estimates of the true power associated with given sample size and parameter configuration were computed via Monte Carlo simulation of 100,000 independent data sets. For each replicate, (N_1, N_2) normal outcomes are generated with the two-sample parallel design. Then, the test statistics T_1 and T_2 are computed and the simulated power is the proportion of the 100,000 replicates with the test statistics $T_1 > t_{\nu,\, 0.05}$ and $T_2 < -t_{\nu,\, 0.05}$. The adequacy for power and sample size calculation is determined by the difference between the simulated power and computed power. The computed power, simulated power, and the corresponding difference are summarized in Table 1 for the examined model settings. An inspection of the summarized results reveals that the suggested exact method based on the power function Ψ_E produces almost identical results with the simulation for all twelve cases. Specifically, the resulting absolute differences are all less than 0.003 and the largest discrepancy 0.0025 is incurred by the situation with $\mu_d = 0.1$, $\sigma = 0.12$, and $N = 13$. Hence, the presented exact approach and computer algorithm have the distinct advantage in computational accuracy.

Design schemes

With the exact power function of the TOST procedure, this study examines research designs with the allocating and budgetary constraints. First, the ratio $r = N_2/N_1$ between the two group sizes may be fixed in advance, so the task is to decide the minimum sample size N_1 $(N_2 = rN_1)$ required to achieve the specified power level. Second, one of the two sample sizes, say N_2, may be pre-assigned, and so the smallest size N_1 required to satisfy the designated power should be found. Third, what is the least cost for a research study to maintain its desired power level? Fourth, how can the maximum power be attained in a scientific investigation with a limited budget?

Table 1. The computed power and simulated power of the two one-sided test for α = 0.05, Δ = 0.2231, and equal sample sizes $N_1 = N_2 = N$.

μ_d	σ	*N*	Computed power	Simulated power	Difference
0.00	0.10	5	0.8823	0.8806	0.0017
	0.12	6	0.8220	0.8218	0.0002
	0.14	8	0.8333	0.8331	0.0002
	0.16	10	0.8238	0.8242	–0.0004
	0.18	12	0.8049	0.8035	0.0014
	0.20	15	0.8181	0.8192	–0.0011
0.10	0.10	9	0.8033	0.8050	–0.0017
	0.12	13	0.8148	0.8123	0.0025
	0.14	17	0.8062	0.8065	–0.0003
	0.16	22	0.8066	0.8068	–0.0002
	0.18	28	0.8110	0.8106	0.0004
	0.20	34	0.8070	0.8083	–0.0013

doi:10.1371/journal.pone.0162093.t001

Design I: sample size ratio is fixed. Consider the scenario that the sample size ratio $r = N_2/N_1$ is pre-assigned, and for ease of illustration, the ratio is assumed as $r \geq 1$. The common balanced design with equal sample sizes is the special case with $r = 1$. An incremental process can be conducted to determine the minimum sample size N_1 needed to attain the specified power 1 –β for the chosen significance level α and parameter values $\{\mu_d, \sigma^2, \Delta\}$. For comparative purpose and computational ease, the approximate normal distribution $T \sim N(\lambda, 1)$ provides a convenient solution where $0\lambda = \mu_d/\sigma^*$ and $\sigma^{*2} = \sigma^2(1/N_1 + 1/N_2)$. To simplify the computation, the starting sample size N_{1Z} computed by the normal approximation would be the smallest integer that satisfies the inequality $N_{1Z} \geq (1 + /r)\sigma^2(z_\alpha + z_\beta)^2/(\Delta–|\mu_d|)^2$ where z_α and z_β are the upper 100·αth and 100·βth percentiles of the standard normal distribution, respectively.

Design II: one sample size is fixed. Without loss of generality, the sample size N_2 of the second group is held constant. Just as in the previous case, the minimum sample size N_1 needed to ensure the specified power 1 –β can be found by an iterative search for the chosen significance level α and parameter values $\{\mu_d, \sigma^2, \Delta\}$. The starting sample size N_{1Z}, based on the normal approximation as is described in the previous situation, is chosen as the smallest integer that satisfies the inequality $N_{1Z} \geq 1/\{(\Delta–|\mu_d|)^2/[\sigma^2(z_\alpha + z_\beta)^2]–/N_2\}$.

Design III: total cost is fixed and the actual power needs to be maximized. Suppose C_F is the overhead cost of the study, and C_1 and C_2 are the costs per subject in the first and second groups, respectively; then the total cost of the study is $C = C_F + C_1N_1 + C_2N_2$. Accordingly, the traditional consideration of the total number of subjects can be viewed as a special case of the cost function C, with $C_F = 0$ and $C_1 = C_2 = 1$. Within the normality context, Pentico [30] showed that the optimal allocation with different unit sampling costs is when the ratio of the sample sizes assumes the equality $N_1/N_2 = C_2^{1/2}/C_1^{1/2}$. For a fixed value of total cost C and a specified fixed cost C_F, the maximum power is obtained with the sample size combination

$$N_{1Z} = \frac{C_2^{1/2}(C - C_F)}{C_1C_2^{1/2} + C_2C_1^{1/2}} \text{ and } N_{2Z} = \frac{C_1^{1/2}(C - C_F)}{C_1C_2^{1/2} + C_2C_1^{1/2}}.$$

Notably, the optimal property is valid only when the statistic T has a normal distribution. However, this is not the case here and a two-step procedure is conducted to find the exact optimum. First, a detailed power assessment is performed for the sample size combinations $\{N_1, N_2\}$ with N_1 from N_{1min} to N_{1max} and $N_2 = Floor[(C–C_F–C_1N_1)/C_2]$, where $N_{1min} = Floor(N_{1Z})–3$, $N_{1max} = Floor[\{C–C_F–C_2(Floor(N_{2Z})– 3)\}/C_1]$, and the function $Floor(a)$ returns the largest integer that is less than or equal to a. Second, the optimal sample size allocation is the one giving the largest power.

Design IV: target power is fixed and the total cost needs to be minimized. In addition to the preceding scenario with limited budget, a distinct approach to take into account both power and cost issues is to find the optimal sample size combination which minimizes the total cost and attains the pre-chosen target power. In view of the discrete character of sample size, the exact procedure is conducted in three steps.

First, in order to achieve the nominal power 1 –β while minimizing total cost $C = C_F + C_1N_{1Z} + C_2N_{2Z}$, a nearly optimal sample size combination under the normal distribution for T is

$$N_{1Z} = \frac{(1 + C_2^{1/2}/C_1^{1/2})\sigma^2(z_\alpha + z_{\beta*})^2}{(\Delta - |\mu_d|)^2} \text{ and } N_{2Z^*} = \frac{(1 + C_1^{1/2}/C_2^{1/2})\sigma^2(z_\alpha + z_{\beta*})^2}{(\Delta - |\mu_d|)^2}$$

where $z_{\beta^*} = z_{\beta/2}$ if $\mu_d = 0$, and $z_{\beta^*} = z_\beta$ if $\mu_d \neq 0$. It can be seen that $N_{1Z^*}/N_{2Z^*} = C_2^{1/2}/C_1^{1/2}$ and $\sigma^2(1/N_{1Z^*} + 1/N_{2Z^*}) = (\sigma–|\mu_d|)^2/(z_\alpha + z_{\beta^*})^2$. Then, the power computation and cost evaluation are conducted for sample size combinations with N_1 from N_{1min} to N_{1max} and a proper value of $N_2 \geq Floor[1/\{(\Delta–|\mu_d|)^2/[\sigma^2(z_\alpha + z_{\beta^*})^2] – 1/N_1\}]$ satisfying the required power, where $N_{1min} = max\{5, Ceil(N_{1Z^*})– 2\}$, $N_{1max} = Ceil(N_{1Z^*}) + 10$, the *max* function selects the largest value of the elements, and the function *Ceil*(a) returns the smallest integer that is greater than or equal to a. Second, the optimal sample size allocation is the one giving the smallest cost while maintaining the specified power level. Third, there may be more than one combination giving the same amount of least cost. A further screening and selection process is conducted to find the one $\{N_{1E}, N_{2E}\}$ producing the largest power.

Results

To illustrate the computational aspects of the suggested procedures for design planning, the example of Minnesota Multiphasic Personality Inventory (MMPI) similarities between alcohol and drug-dependent subjects presented in Rogers et al. [8] is extended here to sample size determinations for equivalence testing under various design schemes. Detailed discussions and related results of the MMPI differences between alcoholics and drug abusers can be found in Cannon, Bell, and Fowler [31].

Due to the prospective nature of advance research planning, the general guidelines suggest that typical sources like published finding or expert opinion can offer plausible and reasonable planning values for the vital characteristics of mean effects, variance components, and equivalence threshold. As an illustration of sample size determination for planning equivalence study, the reported summary statistics of the Masculinity-Femininity scale for the drug and alcohol-dependent groups are modified as the population means and variance. Specifically, $\mu_d = 61.4–59.2 = 2.2$, $\sigma = 9.78$. With these specifications, significance level $\alpha = 0.05$, and equivalence bound $\Delta = 5.92$ (10% of the MMPI scores of the alcoholic subjects), the numerical computation showed that the resulting power for the TOST is $\Psi_E = 0.7711$ for the reported sample sizes $\{N_1, N_2\} = \{49, 207\}$ of the MMPI study. The attained power is slightly less than the fairly common and somehow minimal level of 0.80.

In order to warrant a decent chance of assessing the equivalence property with the pre-assigned the sample size ratio $r = N_2/N_1 = 4$, the optimal sample sizes {54, 216} are required to attain the designated power 0.80. Alternatively, when the sample size $N_2 = 210$ is fixed beforehand, it requires a group size of $N_1 = 55$ in order to achieve the selected power 0.80. However, it is important to take budget issues into account. For illustration, assume the unit sampling costs for the two treatment groups are $C_1 = 4$ and $C_2 = 1$. Under cost consideration with the overhead cost $C_F = 0$ and the total budget $C = 400$ units, the optimal sample size solution is {67, 132} which has an actual power of 0.8111. On the other hand, the optimal sample sizes {65, 128} are required to attain the designated power 0.80 with the least total cost. The detailed computation showed that the attained power and total cost are 0.8005 and 388, respectively. The prescribed parameter configurations are incorporated in the user specifications of the supplementary R and SAS/IML programs. Researchers can easily identify the statements containing the exemplifying values in the computer code and then modify the programs to accommodate their own model specifications. Nonetheless, the suggested procedures will yield accurate power calculations and sample size determinations provided that all the required information is properly specified.

Conclusions

Many studies are designed explicitly to show that two treatments are functionally equivalent or that a new method is as effective as a well-established method under the same condition. Under such circumstance, the traditional tests are inappropriate to establish equivalence, because failing to reject a no-difference hypothesis test does not necessarily support the conclusion of equivalence. Notably, the TOST procedure for establishing statistical equivalence has been used effectively across a wide range of research disciplines. As a contrasting and concrete example, the TOST procedure were illustrated with the MMPI equivalence evaluations between alcohol and drug-dependent subjects in Rogers et al. [8], while the previous study of Cannon, Bell, and Fowler [31] focused on the research issues of MMPI differences between alcoholics and drug abusers. Consequently, it is advisable that investigators should determine when equivalency testing is useful and delineate meaningful equivalence bounds relative to the substantive issues in their expert areas of research.

To enhance the usefulness of TOST methodology, it is prudent to develop a full account of computer programs for implementing the necessary calculations in equivalence studies. Evidently, the lack of efficient and convenient computer software impedes the practical use of equivalence tests and the theoretical development of equivalence research. This article examines the power and sample size problem of equivalence testing for the means from two independent and normally distributed populations with an unknown variance. The exact power function of the TOST procedure is described and employed to compute the optimal sample sizes under various allocation and cost considerations. In view of the importance of power and sample size calculations in design planning and the limited features of available software packages, computer programs are developed to facilitate the usage of the proposed techniques. Overall, the illustrated power and sample size calculations and accompanying algorithms reinforce the theoretical and practical implications of TOST in equivalence studies.

Supporting Information

S1 File. Power function of the two one-sided tests.
(DOCX)

S2 File. The two one-sided tests for 2 × 2 crossover designs.
(DOCX)

S3 File. R programs.
(DOCX)

S4 File. SAS programs.
(DOCX)

Author Contributions

Conceptualization: GS.

Formal analysis: GS.

Investigation: GS.

Methodology: GS.

Resources: GS.

Software: GS.

Validation: GS.

Writing – original draft: GS.

Writing – review & editing: GS.

References

1. Schuirmann DL. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. Biometrics. 1981; 37: 617.
2. Westlake WJ. Response to T.B.L. Kirkwood: Bioequivalence testing–a need to rethink. Biometrics. 1981; 3: 589–594.
3. Meyners M. Equivalence tests–A review. Food Quality and Preference. 2012; 26: 231–245.
4. Chow SC, Liu JP. Design and analysis of bioavailability and bioequivalence studies (3rd ed.). New York, NY: Chapman & Hall/CRC; 2008.
5. Chow SC, Shao J, Wang H. Sample size calculation in clinical research. New York, NY: Marcel Dekker; 2003.
6. Hauschke D, Steinijans V, Pigeot I. Bioequivalence studies in drug development: Methods and applications. Chichester: John Wiley & Sons; 2007.
7. Wellek S. Testing statistical hypotheses of equivalence and noninferiority (2nd ed.). New York, NY: CRC Press; 2010.
8. Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimental groups. Psychological Bulletin. 1993; 113: 553–565. PMID: 8316613
9. Blackwelder WC. "Proving the null hypothesis" in clinical trails. Controlled Clinical Trials. 1982; 3: 345–353. PMID: 7160191
10. Cribbie RA, Gruman J, Arpin-Cribbie C. Recommendations for applying tests of equivalence. Journal of Clinical Psychology. 2004; 60: 1–10. PMID: 14692005
11. Parkhurst DF. Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. BioScience. 2001; 51: 1051–1057.
12. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics. 1987; 15: 657–680. PMID: 3450848
13. Allan TA, Cribbie RA. Evaluating the equivalence of, or difference between, psychological treatments: An exploration of recent intervention studies. Canadian Journal of Behavioral Science. 2013; 45: 320–328.
14. Bristol DR. Probabilities and sample sizes for the two one-sided tests procedure. Communications in Statistics-Theory & Methods. 1993; 22: 1953–1961.
15. Chow SC, Shao J, Wang H. A note on sample size calculation for mean comparisons based on noncentral t-statistics. Journal of Biopharmaceutical Statistics. 2002; 12: 441–456. PMID: 12477068
16. Chow SC, Wang H. On sample size calculation in bioequivalence trials. Journal of Pharmacokinetics and Pharmacodynamics. 2001; 28: 155–169. PMID: 11381568
17. Diletti E, Hauschke D, Steinijans VW. Sample size determination for bioequivalence assessment by means of confidence intervals. International Journal of Clinical Pharmacology, Therapy and Toxicology. 1991; 29: 1–8.

18. Liu JP, Chow SC. Sample size determination for the two one-sided tests procedure in bioequivalence. Journal of Pharmacokinetics and Biopharmaceutics. 1992; 20: 101–104. PMID: 1588502

19. Muller-Cohrs J. The power of the Anderson-Hauck's test and the double t-test. Biometrical Journal. 1990; 32: 259–266.

20. Phillips KF. Power of the two one-sided tests procedure in bioequivalence. Journal of Pharmacokinetics and Biopharmaceutics. 1990; 18: 137–144. PMID: 2348380

21. Siqueira AL, Whitehead A, Todd S, Lucini MM. Comparison of sample size formula for 2 × 2 cross-over designs applied to bioequivalence studies. Pharmaceutical Statistics. 2005; 4: 233–243.

22. Wang H, Chow SC. On statistical power for average bioequivalence testing under replicated crossover designs. Journal of Biopharmaceutical Statistics. 2002; 12: 295–309. PMID: 12448572

23. Jan SL, Shieh G. Optimal sample sizes for Welch's test under various allocation and cost considerations. Behavior Research Methods. 2011; 43: 1014–1022. doi: 10.3758/s13428-011-0095-7 PMID: 21512873

24. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics. 1980; 36: 343–346. PMID: 26625475

25. Heilbrun LK, McGee DL. Sample size determination for the comparison of normal means when one sample size is fixed. Computational Statistics & Data Analysis. 1985; 3: 99–102.

26. Nam JM. Optimum sample sizes for the comparison of the control and treatment. Biometrics. 1973; 29: 101–108. PMID: 4691048

27. Allison DB, Allison RL, Faith MS, Paultre F, Pi-Sunyer X. Power and money: Designing statistically powerful studies while minimizing financial costs. Psychological Methods. 1997; 2: 20–33.

28. R Development Core Team. R: A language and environment for statistical computing [Computer software and manual]; 2014. Retrieved from http://www.r-project.org.

29. SAS Institute. SAS/IML User's Guide, Version 9.3. Cary, NC: SAS Institute Inc; 2014.

30. Pentico DW. On the determination and use of optimal sample sizes for estimating the difference in means. The American Statistician. 1981; 35: 41–42.

31. Cannon DS, Bell WE, Fowler DR. MMPI differences between alcoholics and drug abusers: Effect of age and race. Psychological Assessment: A Journal of Consulting and Clinical Psychology. 1990; 2: 51–55.