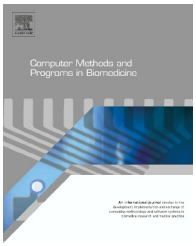
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 126 (2016) 110–117

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



A comparative appraisal of two equivalence tests for multiple standardized effects

*Gwowen Shieh**

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

ARTICLE INFO

Article history:
Received 1 May 2015
Received in revised form
11 December 2015
Accepted 11 December 2015

Keywords:
Effect size
Power
Sample size
Standardized mean difference

ABSTRACT

Equivalence testing is recommended as a better alternative to the traditional difference-based methods for demonstrating the comparability of two or more treatment effects. Although equivalent tests of two groups are widely discussed, the natural extensions for assessing equivalence between several groups have not been well examined. This article provides a detailed and schematic comparison of the ANOVA *F* and the studentized range tests for evaluating the comparability of several standardized effects. Power and sample size appraisals of the two grossly distinct approaches are conducted in terms of a constraint on the range of the standardized means when the standard deviation of the standardized means is fixed. Although neither method is uniformly more powerful, the studentized range test has a clear advantage in sample size requirements necessary to achieve a given power when the underlying effect configurations are close to the priori minimum difference for determining equivalence. For actual application of equivalence tests and advance planning of equivalence studies, both SAS and R computer codes are available as supplementary files to implement the calculations of critical values, *p*-values, power levels, and sample sizes.

1. Introduction

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. The methodology for establishing statistical equivalence has been typically developed for appraising the bioequivalence between two drug formulations in biopharmaceutical studies. A comprehensive review of the different types of equivalence tests can be found in Meyners [1]. Accordingly, the two one-sided tests procedure proposed by Schuirmann [2] and Westlake [3] is the most common method for equivalence assessment. It is essential to emphasize that power analyses and sample size calculations are often critical for investigators to credibly address designated hypotheses and support research questions. Considerable attention has been devoted to the power and sample size issues of the two one-sided tests procedure in the literature, such as Bristol [4], Chow, Shao, and Wang [5], Chow and Wang [6], Diletti, Hauschke, and Steinijans [7], Liu and Chow [8], Phillips [9], Schuirmann [10], and Wang and Chow [11].

Despite the equivalent tests of two groups are widely discussed, the natural extensions for assessing equivalence between several groups have received relatively little attention in the literature. Two different effect size measures have been proposed to represent the degree of disparity among several treatment groups (Cohen [12,13]). One index relies on the

* Tel.: +886 35712121.
E-mail address: gwshieh@mail.nctu.edu.tw

http://dx.doi.org/10.1016/j.cmpb.2015.12.004
0169-2607/

standard deviation of the standardized means and the second index is the range of the standardized means. Consequently, these two distinct measures of disparity among standardized means give rise to two different multiple-sample procedures: the ANOVA *F* test and the studentized range test. However, the research for the two methods has mainly focused on the test of the traditional null hypothesis of no difference in treatment means. In fact, the two approaches can be readily modified as viable equivalence tests. Specifically, Wellek [14] proposed an adjustment of the ANOVA *F* test based on the variance of the standardized means. On the other hand, the studentized range test was considered in Giani and Finner [15] for the range of the standardized means. Related discussions are presented in Cribbie, Arpin-Cribbie, and Gruman [16], Chen, Wen, and Wang [17], and the references therein.

The prescribed results in Wellek [14], Giani and Finner [15], Cribbie, Arpin-Cribbie, and Gruman [16], and Chen, Wen, and Wang [17] of the ANOVA *F* test and the studentized range test offer fundamental guidance for the equivalence problem of several treatments. Then a natural question of great interest is which of the two methods should be used because they are markedly different in theoretical principles and demand varying computational efforts. It was noted in Wellek [14] that power comparisons between the two approaches are futile because the associated distribution and hypothesis formulation are apparently distinct. Accordingly, no research to date has compared these two approaches regarding which method is more appropriate under what circumstances for determining whether treatment means are sufficiently near each other to be considered equivalent. Although the range of the standardized means is not a function of the standard deviation of the standardized means, there exists an intrinsic property for the lower and upper bounds of the range of the standardized means when the standard deviation of the standardized means is fixed (Pearson and Hartley [18]). Therefore, it provides a unified and meaningful viewpoint to evaluate the power behavior of the two grossly diverse techniques. More importantly, the explications presented later reveal that without a detailed appraisal, one may unknowingly employ an equivalence test procedure with lesser efficiency and inevitably confront the consequence of inadequate power performance and unsatisfactory research outcome.

In addition, the notion of least favorable configurations of treatment means emphasized in Giani and Finner [15] and Chen, Wen, and Wang [17] is vital to determine the critical values for conducting equivalence tests of unimportant differences. Tables of such critical values are generally not available to applied researchers and it is impossible to implement the test procedures without an efficient software package. Moreover, in order to enhance the usefulness of a test procedure, the corresponding power and sample size calculations must also be considered to extend its applicability in planning research studies. Due to the complexity of both the studentized range and the ANOVA *F* test procedures, Chen, Wen, and Wang [17] and Wellek [14] also addressed the corresponding computational issues under different computing systems. However, the presented algorithms do not provide all the desired features for data analysis and design planning. Arguably, the lack of efficient and convenient computer software impedes the practical use of equivalence tests and the theoretical development of equivalence research. Therefore, it is prudent to develop a full account of computer programs for implementing the necessary calculations in equivalence studies.

In view of the inadequate results in the literature, the present article aims to contribute to the analysis and design of equivalence studies in two ways. First, within the context of equivalence framework, the fundamental properties of the standard *F* test and the studentized range test are reviewed to document the importance of the problem and the characteristics of available methods. Moreover, extensive numerical assessments are performed to reveal the power performance and sample size requirement of the ANOVA *F* test and the studentized range test under a wide range of model configurations. The appraisals discern not only which method is most suitable under what circumstances but also the actual differences between the contending test procedures. Second, to facilitate the application of the examined approaches, the corresponding SAS and R computer codes are developed to compute the critical values, observed significance levels, attained power levels, and required sample sizes. Note that the implementation of both methods involves specialized programs not currently available in prevailing statistical packages. The constructed software programs provide a unified set of algorithms for design planning and data analysis of equivalence investigations.

2. Test procedures

Consider the one-way fixed-effects ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{1}$$

where Y_{ij} is the value of the response variable in the jth trial for the ith factor level, μ_i are treatment means, ε_{ij} are independent $N(0, \sigma^2)$ errors with $i = 1, \ldots, G$ (≥ 2) and $j = 1, \ldots, N$. To characterize the degree of departure from no treatment effect, two distinctive measures for the balanced design were proposed in Cohen [12,13]. The first index is the standard deviation of the standardized means

$$f = \frac{\sigma_\mu}{\sigma} \tag{2}$$

where $\sigma = (\sigma^2)^{1/2}$, $\sigma_\mu = (\sigma_\mu^2)^{1/2}$, $\sigma_\mu^2 = \sum_{i=1}^{G}(\mu_i - \bar{\mu})^2/G$ is the average dispersion between the treatment means, and $\bar{\mu} = \sum_{i=1}^{G}\mu_i/G$ is the mean of the treatment effects. The second index is based on the range of the standardized means

$$\delta_R = \frac{\mu_{max} - \mu_{min}}{\sigma}, \tag{3}$$

where μ_{max} and μ_{min} are the maximum and the minimum of the G treatment means, respectively. In general, the two effect sizes f and δ_R have no direct functional relationship except for $G = 2$ that $f = \delta_R/2 = |\delta|/2$, where $\delta = (\mu_1 - \mu_2)/\sigma$ is the well-known standardized mean difference. Notably, the corresponding inferential procedures are also substantially different. The general guidance of Cohen [12,13] suggests that the small, medium, and large effects in terms of f and δ could be defined