Journal of Modern Applied Statistical Methods

Volume 1 | Issue 1 Article 18

5-1-2002

Two-Sided Equivalence Testing Of The Difference Between Two Means

R. Clifford Blair
University of South Florida

Stephen R. Cole
Bloomberg School of Public Health, The Johns Hopkins University

Journal Of Modern Applied Statistical Methods
Winter 2002, Vol. 1, No. 1, 139-142

1538 - 9472/02/$30.00

BRIEF REPORTS

Two-Sided Equivalence Testing Of The Difference Between Two Means

R. Clifford Blair
Department of Epidemiology and Biostatistics
College of Public Health, &
Jaeb Center For Health Research
University of South Florida

Stephen R.Cole
Department of Epidemiology
Bloomberg School of Public Health
The Johns Hopkins University

Studies designed to examine the equivalence of treatments are increasingly common in social and biomedical research. Herein, we outline the rationale and some nuances underlying equivalence testing of the difference between two means. Specifically, we note the odd relation between tests of hypothesis and confidence intervals in the equivalence setting.

Keywords: Equivalence, Statistical inference, Hypothesis testing, Confidence intervals

Introduction

Studies designed explicitly to examine the equivalence of two (or more)treatments are increasingly common in social and biomedical research. In such studies the null hypothesis maintains that the difference between treatments is at least of some specified magnitude, while the alternative specifies a lesser difference. Some consequences of stating hypotheses in this fashion are not obvious. For example, intention-to-treat analyses do not carry the same robust interpretation when there is noncompliance (Robins,1988),random measurement error may bias toward rejecting the null (Jones, et al.,1996), and significantly larger sample sizes may be required (Makuch & Johnson, 1986). In order to understand these and other consequences of equivalence testing one must first have an understanding of the basic tenets underlying the methodology.

The purpose of this report is to briefly outline the rationale and some of the nuances underlying equivalence testing of the difference between two means. For simplicity the context involves the difference between means but the explanations afforded apply with equal force to tests of the difference between two adjusted means as might be obtained from a two group ANCOVA analysis. Topics to be covered include, establishing equivalence by means of hypothesis tests, establishing equivalence by means of one and two-sided confidence intervals, power considerations, and efficacy tests.

Establishing 2-Sided Equivalence by a Pair of Nested Hypothesis Tests

Two-sided equivalence tests (ET) of two means begin with the establishment of an equivalence interval defined by the constants c and -c. Equivalence is declared if it can be established that the value $\mu_1 - \mu_2$ lies within the interval (-c,c).In order to make this determination, two one-sided tests of significance must be carried out as depicted in Figure 1.In this figure -c and c are set to -1 and 1 units, respectively. The curves depict the sampling distribution of $\overline{X}_1 - \overline{X}_2$ under each of the null hypotheses to be tested. In order to establish equivalence the null hypothesis $\mu_1 - \mu_2 \geq 1$ must be rejected in favor of the alternative $\mu_1 - \mu_2 < 1$ in order to establish the fact that $\mu_1 - \mu_2$ is below the upper bound of the equivalence interval. Similarly, a second test is necessary to show that $\mu_1 - \mu_2$ is greater than the lower bound of the equivalence interval.

Notice that both tests must attain significance in order to declare equivalence. Notice also that both null hypotheses cannot be true. Therefore, the Type I error rate will be determined by the critical region of only one of the two curves. If $\mu_1 - \mu_2 = 1$,the probability of a type I error (α) is the shaded critical region of the right hand curve.

Also of interest is the fact that the nominal level of the test establishes an upper bound for Type I errors rather than an explicit level. This derives from two factors: (1) If the null value exceeds c (e.g., 2 units) or is less than -c, the Type I error rate will necessarily be decreased. This is common to standard (i.e., efficacy) one-sided tests and will not be discussed here. (2) In the event that the standard error

R.Clifford Blair is Professor and Interim Chair, Department of Epidemiology and Biostatistics, College of Public Health MDC-56, University of South Florida,13201 Bruce B.Downs Blvd., Tampa FL, 33612-3805. E-mail: cblair@hsc.usf.edu. His areas of expertise are in computer-intensive statistical methods, multiple end point analysis, and control of family-wise error. Stephen R.Cole is Assistant Research Professor, Department of Epidemiology, The Johns Hopkins School of Hygiene and Public Health. He specializes in longitudinal data analysis, computer-intensive applications, and testing issues in medicine.

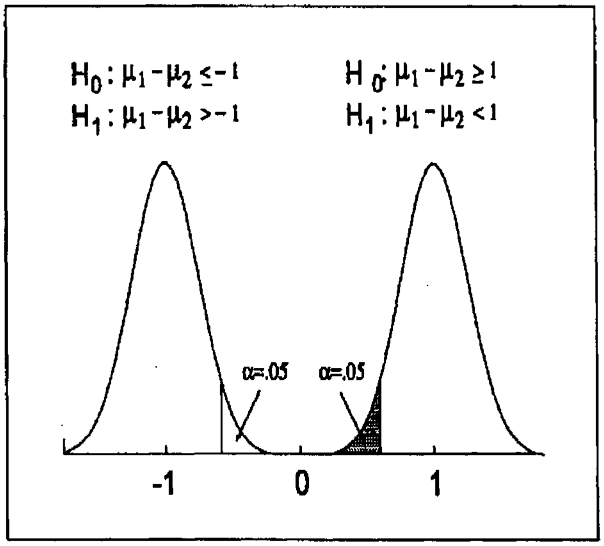


Figure 1: Hypothesis tests to establish quivalence of two means.

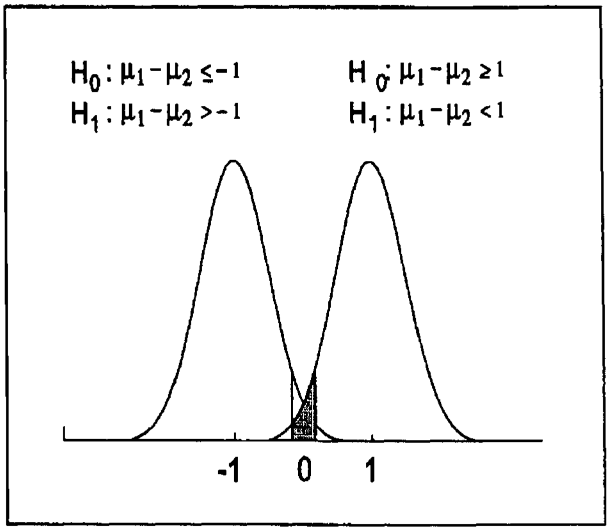


Figure 2: Conservative test of significance.

of the test statistic (SE) is too large and/or the length of the equivalence interval is too small, the two critical regions will overlap to a significant degree thereby producing a conservative test. This situation is depicted in Figure 2 where the Type I error rate is represented by the gray shaded area in the critical region of the right hand curve. In the extreme, the two critical regions may completely overlap so that the Type I error rate will be zero.

Establishing Equivalence by Means of One and Two-Sided Confidence Intervals

As with standard efficacy tests (Cox & Hinkley, 1974), there is a relationship between tests of hypotheses and confidence intervals used to establish equivalence. The relationship for equivalence is somewhat different from that for efficacy, however.

As depicted in Figure 3, the distance between the hypothesized null value and the beginning of the critical region is (approximately)1.65 standard errors (SE). Because the upper end of a one-sided 95 percent confidence interval is given by U = TS +1 .65SE where TS is the test statistic ($\bar{X}_1 - \bar{X}_2$ in the present case), it follows that any TS in the critical region of the right hand curve will produce a value of U that is less than c (or 1 in this sample).This situation is depicted in panel A of the figure. On the other hand, a value of U that is greater than c implies that TS is not in the critical region as shown in panel B. Thus, a value of U less than c implies rejection of $H:_0$, while a value greater than c implies a failure to reject. The same logic applies to the lower end of a one-sided 95 percent confidence interval and a test of hypothesis carried out on the lower curve. Thus, noting that neither of two one-sided 95 percent confidence intervals overlap c or -c is equivalent to noting that both hypothesis tests are significant.

It is obvious that a more direct assessment can be carried out by inspection of a two-sided 90 percent confidence interval where the lower and upper limits are constructed by subtracting and adding 1.65 SE to TS. In this case, significance is established by noting that the 90 percent confidence interval is completely contained in the interval (-c, c). In general, one would use a two-sided 100(1 -2 α) percent confidence interval to establish significance at the α level. Although convenient, reporting use of this methodology can lead to confusion on the part of readers not familiar with equivalence methods.

A statement of the form "A two-sided significance test was carried out at the .05 level by means of a 90 percent confidence interval " is almost certain to cause confusion. An equivalent statement concerning two one sided 95 percent confidence intervals seems slightly more palatable.

Power Considerations

For the present situation, power is defined as the probability of attaining significance when $\mu_1 - \mu_2$ is contained in the interval (-c, c). Equivalently, power may be defined as the probability that a properly constructed confidence interval will be completely contained in the interval (-c, c) when $\mu_1 - \mu_2$ is in the interval (-c, c). Power calculations are usually carried out under the assumption that $\mu_1 - \mu_2$ = 0, although other values may be chosen when the research situation warrants.

The shaded area of the middle distribution in Figure 4 depicts power of .95 for a two-sided equivalence test. Notice that the probability of failing to obtain a significant result in this situation is the unshaded portion in the tails

Figure 3: Relationship of a confidence interval to a test of hypothesis.

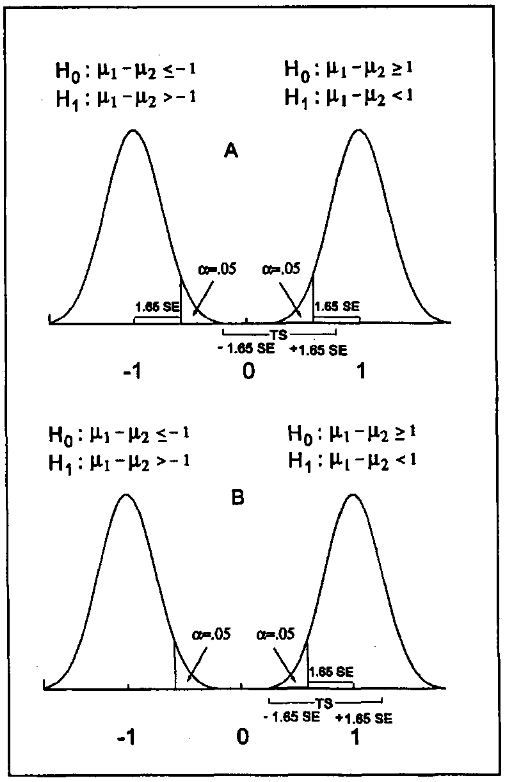
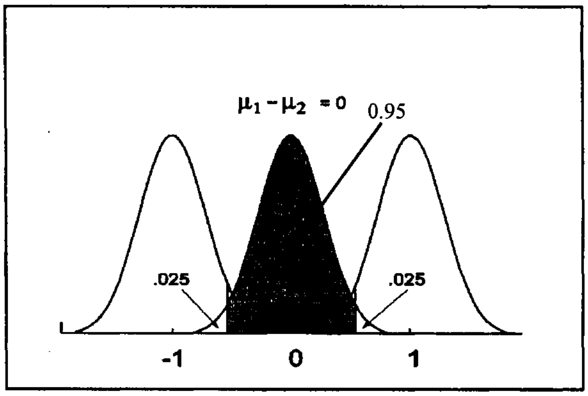


Figure 4: Power of a two-sided equivalence test.

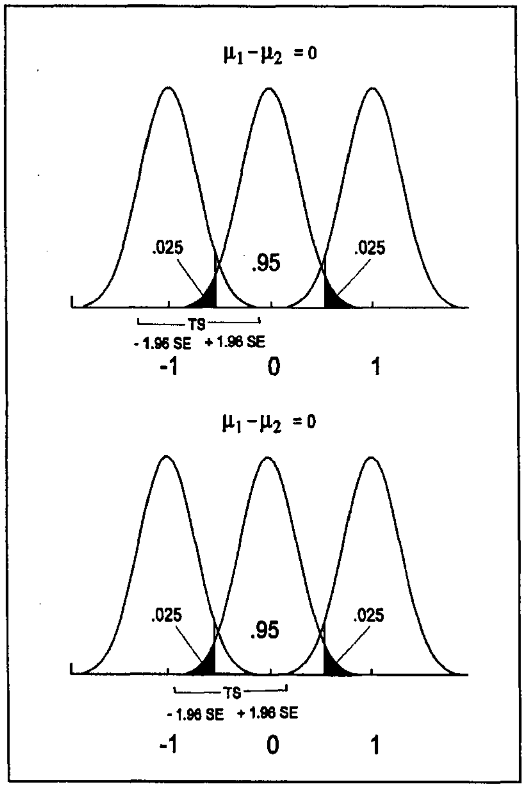


of the middle distribution. Because power is .95, the unshaded area is .05 with .025 allocated to each tail. Power for equivalence tests is often lower than that for efficacy tests of the same sample size due to the typically smaller effect sizes associated with equivalence tests relative to efficacy tests.

Efficacy Tests

Inherent in every equivalence test is an efficacy test of the null hypothesis $\mu_1 - \mu_2 = 0$. Look again at the middle distribution in Figure 4. This is the sampling distribution of the TS when $\mu_1 - \mu_2 = 0$ which in turn is the usual null distribution when testing the hypothesis $\mu_1 - \mu_2 = 0$. The probability of rejecting the efficacy test when there is no difference is one minus the power of the equivalence test. Thus, when establishing the power of the equivalence test one is also establishing the Type I error rate of the efficacy test. This leads to the following result: If significance is attained with the equivalence test, one can conclude that the treatments are equivalent with probability of error being α (usually .05). When non-significance is attained with the equivalence test one can state that there is a significant difference with probability of error being one minus the power of the equivalence test. By setting power at .95 this error rate becomes the traditional .05. Note that non-significance of the equivalence does not mean non-equivalence ($\mu_1 - \mu_2$ may well be in the equivalence interval), but does mean that there is *some,* albeit possibly practically irrelevant, difference in the treatments.

Figure 5: Using 95 percent CI to establish significance.



One last point should be noted. If the significanc level for the equivalence test is established at .025 and power at .95 the following results. (1) Significance can be determined by means of a two-sided 95 percent confidence interval. (2) If this interval is completely contained in (-c,c), equivalence is established with probability of error being.025. (3) If this interval is not completely contained in (-c,c), then a significant difference between μ_1 and μ_2 is established with probability of error being .05. (4) These determinations can also be made by noting whether or not zero is in the 95 percent confidence interval. (See Figure 5.) (5) The 95 percent confidence interval can be used to estimate $\mu_1 - \mu_2$ in the usual manner.

References

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman and Hall.

Jones, B., Jarvis, P., Lewis, J.A., & Ebbutt, A. F. (1996). Trials to assess equivalence: The importance of rigorous methods. *British Medical Journal, 313*, 36-39.

Makuch, R. W., & Johnson, R. F. (1986). Some issues in the design and interprettion of 'negative'clinical studies. *Archives of Internal Medicine, 146*, 989.

Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine, 17*, 269-302.