

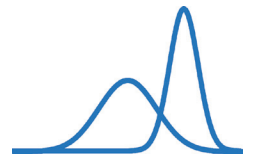
Developing and applying advanced research tools for human perceptual measurement

IFPress® Research Papers

Number 904

# Equivalence Hypothesis Testing

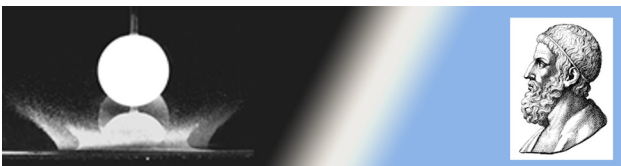
Daniel M. Ennis and John M. Ennis



**The Institute for Perception**

7629 Hull Street Rd.  
Richmond, VA 23235

Available at [www.ifpress.com](http://www.ifpress.com)



## Equivalence Hypothesis Testing

Daniel M. Ennis and John M. Ennis

*The Institute for Perception, 7629 Hull Street Road, VA 23235, USA*

---

### Abstract

In statistical applications, such as a comparison of two items, it is useful to know whether one item is equivalent to another. Similarly it is often desirable to know whether one item can act as a substitute for another. Applications of the concept of equivalence include blend and flavor modifications of products, substitution of generic drugs for brand-name drugs, modifications of products in response to government regulations, or component substitutions with more healthful or lower cost components. In addition, some companies develop products that are direct substitutes for those of their competitors and make advertising claims concerning their equivalence.

In a recent paper, Ennis and Ennis (2009) used an open interval to define equivalence and provided exact and approximate methods for testing a null hypothesis of nonequivalence. In this paper, a discussion of this newly developed theory of equivalence testing is presented along with a comparison to existing methods such as the “two one-sided tests” (TOST) method. We provide numerical examples to illustrate this new theory and we demonstrate that although the TOST is a convenient approximation it is fundamentally inconsistent with the specification of the null hypothesis.

**Keywords:** Ratios of normal random variables; confidence intervals; estimation; multiplicative comparisons; ratio statements; multiplicative statements

---

### 1. Introduction

Various terms have been used to describe what we will call “equivalence” in this paper. *Equivalence*, *bioequivalence*, *parity*, *equality*, and *similarity* are terms used to describe a condition in which two items differ within some predetermined bounds. *Bioequivalence* is a term used in the context of drug effect differences and the term *parity* is often used in the context of consumer product differences. It seems preferable to avoid using the term “similarity” in this context as it has a specific meaning in the psychology literature (Ashby and Ennis, 2007; Ennis, 1988, 1992; Ennis and Johnson, 1993; Ennis *et al.* (1988); Ennis and Ashby, 1993; Tversky, 1977). In addition, the recently reissued American Society for Testing and Materials Claims Guide (ASTM E 1958, 2006) distinguishes two types of parity advertising claims, which are termed *equality* and *unsurpassed*. An equality claim can be made when two products are essentially equivalent within some application specific bounds on an attribute of interest. The term “equality” is somewhat misleading as it implies no bounds in the definition of equality, even though these bounds are included. Two drugs are *bioequivalent* if their

effects are equivalent within defined bounds on some clinically relevant scale. In this paper “equivalent” will be used in place of application-specific terms such as “bioequivalent” and “equality.”

The domain of applications for equivalence hypothesis testing is quite broad. In false advertising cases brought under the provisions of the Lanham Act (Title 15, §1125 of the United States Code) an advertiser may be sued for an alleged false claim that its product is equivalent on some performance measure to a competitor’s product. Examples include claims that dropped call rates are equivalent among cell phone providers, that an artificial sweetener is equivalent to a natural sweetener, or that one tooth whitening product is as effective as another. Another type of application occurs when, for instance, a consumer products company makes a change in an ingredient or process to reduce cost, improve healthfulness or comply with regulations. A contemporary example includes changes made to food products to address the public health issue of childhood obesity. In this last case, the objective may be to produce a more healthful product that has a similar taste or odor profile to an existing product.

Among the several available tests for equivalence, the “two one-sided tests” (TOST) is perhaps the most well known (Berger and Hsu, 1996a; Westlake, 1981; Schuirmann, 1981). Berger and Hsu (1996a) have also discussed three other tests of interest, a test due to Anderson and Hauck (1983), an unbiased test due to Brown *et al.* (1995) and a slightly biased test that they themselves proposed.

In this paper we review a new theory for equivalence testing laid out in Ennis and Ennis (2009). The goals of this paper are to explain this new theory and to further contrast this theory with the existing theory for equivalence testing. For technical details the reader will be referred to Ennis and Ennis (2009). We will start by considering binomially distributed random variables before considering normally distributed random variables in the cases of either known or unknown variance. In each instance we will provide examples. We will then contrast our new methods with existing methods for equivalence testing, including the TOST, before concluding with a summary.

## 2. Binomially Distributed Random Variables

Suppose that  $X$  is a binomially distributed random variable with parameters  $N$  and  $\mu$ , and that  $x$  is a particular realization of  $X$ . In a two-alternative forced choice (2-AFC) or a preference task,  $x$  would represent the choice count of one of the two products. If there is no difference between the treatments then  $\mu$  equals 0.5. We define equivalence to mean that  $\mu$  falls within a value  $\theta$  of 0.5 where  $\theta$  is a predetermined positive constant specific to the application. For instance,  $\theta$  would be 0.05 if the bounds defining equivalence were 45% and 55%. We have the following null ( $H_0$ ) and alternative ( $H_a$ ) hypotheses,

$$H_0: \mu \leq 0.5 - \theta \text{ or } \mu \geq 0.5 + \theta \tag{1}$$

$$H_a: 0.5 - \theta < \mu < 0.5 + \theta. \tag{2}$$

Ennis and Ennis (2009) proposed an exact test of the equivalence hypothesis by considering

$$p = \sum_{k=0}^{N-m} \binom{N}{k} (0.5-\theta)^k (0.5+\theta)^{N-k} - \sum_{k=0}^{m-1} \binom{N}{k} (0.5-\theta)^k (0.5+\theta)^{N-k} \tag{3}$$

where  $m = \min(x, N-x)$ . Note that this equation is the basis for tables that have been recently published (Ennis, 2008).

### 2.1 Binomially Distributed Data - Example 1

In a difference test of two products with different sweeteners among 600 consumers, 295 of them chose the first product as sweeter. If the bounds used to define equivalence are 0.45 and 0.55 then  $\theta = 0.05$ . Assuming the null hypothesis of nonequivalence and using (3),

$$p = \sum_{k=0}^{305} \binom{600}{k} (0.5-0.05)^k (0.5+0.05)^{600-k} - \sum_{k=0}^{294} \binom{600}{k} (0.5-0.05)^k (0.5+0.05)^{600-k} = 0.0205.$$

In this case, the null hypothesis of nonequivalence is rejected in favor of the alternative hypothesis of equivalence at an  $\alpha$  level of 0.05. The two products are considered to be equivalently sweet.

### 2.2 Binomially Distributed Data - Example 2

In a preference test of two cola products among 1000 consumers, 420 of them preferred the first product. If the bounds used to define equivalence are 0.4 and 0.6 then  $\theta = 0.1$ . Using (3),

$$p = \sum_{k=0}^{580} \binom{1000}{k} (0.5-0.1)^k (0.5+0.1)^{1000-k} - \sum_{k=0}^{419} \binom{1000}{k} (0.5-0.1)^k (0.5+0.1)^{1000-k} = 0.1043.$$

In the second case we cannot reject the null hypothesis of nonequivalence at an  $\alpha$  level of 0.05 and therefore cannot rule out the possibility with reasonable confidence that one of the products is preferred.

## 3. Normally Distributed Random Variables

Assume that  $X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ . The null and alternative hypotheses can be stated as

$$H_0: \mu \leq -\theta \text{ or } \mu \geq +\theta \tag{4}$$

$$H_a: -\theta < \mu < +\theta. \tag{5}$$

or identically,

$$H_0: \mu^2 \geq \theta^2 \tag{6}$$

$$H_a: \mu^2 < \theta^2. \tag{7}$$

### 3.1 Normally Distributed Data - Known Variance

Ennis and Ennis (2009) showed that an exact test of the null hypothesis against the alternative is to find

$$p = \Pr[\chi_1'^2(\lambda) < x^2/\sigma^2], \tag{8}$$

where  $\lambda = [\theta^2/\sigma^2]$ .  $\chi_1'^2(\lambda)$  is a noncentral chi-square random variable with one degree of freedom and noncentrality parameter  $\lambda$ .

There are numerous software tools available to evaluate (8). Ennis and Ennis (2009) also noted that the noncentral chi-square in (8) can be expressed in terms of differences of standard normal distribution functions,

$$p = \Phi\left(\frac{|x|-\theta}{\sigma}\right) - \Phi\left(\frac{-|x|+\theta}{\sigma}\right) \tag{9}$$

Standard normal tables can be used to evaluate (9).

### 3.1 Normally Distributed Data - Examples 1 and 2

In many applications the variance is unknown. However, in choice experiments the variance can be determined and the above method can be used via the normal approxima-

tion to the binomial. In the example from Section 2.1, the bounds defining equivalence occur at 0.45 and 0.55. Of 600 consumers, 295 preferred the first alternative. After this value is continuity corrected to 294.5, the sample estimate of the population mean is 0.490833, from which we derive

$$x = 0.490833 - 0.5 = -0.009167.$$

Under the assumption of the null hypothesis, the variance is known at the equivalence bounds,

$$\sigma^2 = \left( \frac{\sqrt{0.45 \times 0.55}}{1000} \right)^2 = 0.02031^2.$$

Using (9) we have

$$p \approx \Phi\left(\frac{0.009167-0.05}{0.02031}\right) - \Phi\left(\frac{-(0.009167+0.05)}{0.02031}\right) = 0.0204.$$

In the example from Section 2.2, 420 consumers from a sample of 1000 preferred one product to the other. In this case  $p \approx 0.1041$ . These two probabilities compare quite well with the exact values of 0.0205 and 0.1043.

It should be noted that (9) provides an exact test of the equivalence hypothesis if the variance is known and the random variable is normally distributed. The reason that the above examples are approximate is that the normal is used to approximate the binomial.

### 3.3 Normally Distributed Data - Unknown Variance

As in Section 3.1, suppose that when two products are identical that  $\mu = 0$ . When the variance is unknown, the noncentrality parameter in (8) and (9) is unknown. One could consider the sample variance as a substitute for the unknown variance but this approach induces liberality in the test of the null hypothesis. Instead we adjust the noncentrality parameter using a positive constant to ensure that the Type I error remains at or below the nominal level. Ennis and Ennis (2009) referred to this constant as  $c$ , and this constant may be determined for the cases where either the population variance is completely unknown or the population variance is known not to exceed some upper bound.

Ennis and Ennis (2009) provided the technical justification behind the claims of the previous paragraph as well as a table of  $c$  values to be used in bioequivalence trials. More generally, tables of  $c$  values can be generated using the theory in Ennis and Ennis (2009) for any set of experimental specifications. Once an appropriate value of  $c$  is determined, it is straightforward to let  $\lambda'' = \frac{c\theta^2}{s^2}$  and to conduct an equivalence test based on

$$p \approx \Pr\left[\chi_1^2(\lambda'') < \frac{x^2}{s^2}\right].$$

Using normal distribution functions, this becomes

$$p \approx \Phi\left(\frac{|x| - \sqrt{c}\theta}{s}\right) - \Phi\left(\frac{-(|x| + \sqrt{c}\theta)}{s}\right).$$

If  $p$  is less than a pre-specified  $\alpha$  level then the null hypothesis of nonequivalence is rejected in favor of the alter-

native hypothesis of equivalence. For the sake of reference, this method of equivalence testing is known as the Adjusted Noncentral Chi-Square (ANC) method.

### 3.4 Normally Distributed Data - Example 3

For this example we assume that when two drugs are equivalent on some clinical measure of efficacy, such as  $\log(\text{area under a concentration versus time curve})$ , that the difference between their means falls within  $\pm \log(1.25)$ . Here the multiple of 1.25 indicates that one drug is 125% more efficacious than the other. This means that  $\theta = \log(1.25)$  in this case. Suppose that for two drugs the difference in  $\log(\text{efficacy})$  between treatments is 0.04 and the sample standard error was 0.103, which is incidentally the error reported in a clinical trial by Westlake (1976) following a conversion to natural logs. Using the tabulated value 0.8204 for  $c$  we reject the null hypothesis at an  $\alpha$  level of 0.05 because

$$\begin{aligned} \Phi\left(\frac{0.04}{0.103} - \frac{\sqrt{0.8204}\log(1.25)}{0.103}\right) - \Phi\left(-\left(\frac{0.04}{0.103} + \frac{\sqrt{0.8204}\log(1.25)}{0.103}\right)\right) \\ = 0.048. \end{aligned}$$

## 4. Alternative Methods

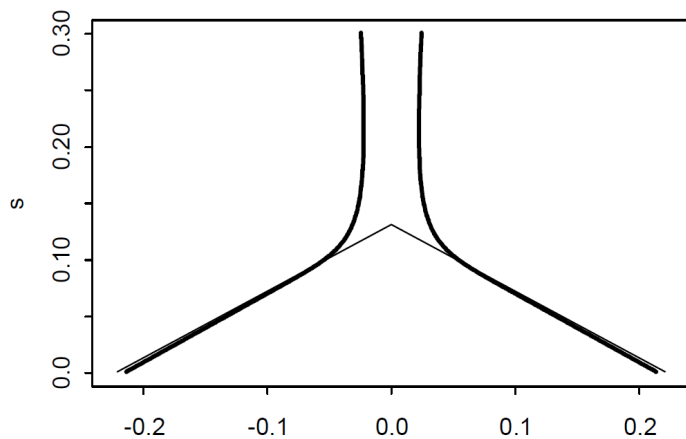
### 4.1 TOST

The TOST [Westlake (1981) and Schuirmann (1981)] involves conducting two one-sided tests and rejecting the null hypothesis if both tests reject at a particular  $\alpha$  level. An exact test of the nonequivalence null hypothesis was given earlier when the variance is known. The TOST does not provide the same results in this case, implying that it is not an exact equivalence test. In fact the TOST is fundamentally flawed because it does not consider the correlation between the two statistics it uses. To better understand the TOST it is of interest to consider the joint distribution of the two random variables involved,  $T_u$  and  $T_l$ , corresponding to tests at the upper and lower bounds, respectively.

$$T_u = \frac{X-\theta}{s} \text{ and } T_l = \frac{X+\theta}{s}.$$

$T_u$  follows a central  $t$  distribution with mean 0 and variance  $r/(r-2)$ .  $T_l$  follows a noncentral  $t$  distribution with noncentrality parameter  $2\theta/\sigma$ . Ennis and Ennis (2009) showed that the correlation coefficient between  $T_u$  and  $T_l$  depends on  $\sigma$ ,  $\theta$  and  $r$ . As the standard deviation of  $X$  increases, the power decreases in part due to the effect of this correlation.

A more detailed account of the performance of the TOST is evident from its rejection region, which is shown in Figure 1. This region summarizes the experimental results that would lead to a rejection of the null hypothesis with  $\sigma$ ,  $\theta$ , and  $r$  defined. Under the specifications in Figure 1, values for  $s$  above 0.131 are not contained within the TOST rejection region, so if a trial exhibits a standard error of that magnitude or greater, the null hypothesis will never



**Figure 1.** The rejection region for the Adjusted Noncentral Chi-square (ANC) method described by the sample variable,  $x$ , and the sample standard error,  $s$ , for trials with 30 degrees of freedom,  $\theta = \log(1.25)$  and  $\alpha = 0.05$ . The rejection region occurs below and between the two thick curved lines. The TOST rejection region in each figure falls below and between the two intersecting lines.

be rejected. This is a very strange property for a statistical test and what makes it even more peculiar is that this value changes depending on the specifications. It is sometimes argued that this flaw is a virtue because the TOST is unaffected by unbounded variance (Meredith and Heise, 1996; Shuirmann, 1996). However, this property did not originate with the null hypothesis. It has been stated that unbounded variance may lead to a situation where  $x$  exceeds  $\theta$  while  $x$  is still inside the rejection region, which may seem counterintuitive. There are two points to be made in response to this concern, one practical and the other theoretical. First, the likelihood of this happening in practice is essentially nil in equivalence testing involving consumer responses because of the large sample sizes relative to the rating variance. The variance for a 9-point rating scale, such as that employed to measure liking, is typically in the range of about 2.5 to 3.5, depending on the product category. With a sample size of 500, it would be highly improbable to obtain standard errors of the difference between means exceeding about 0.13, since more than 99% of them would be less than this value. Even with a sample size of 100, which is not recommended for equivalence testing, this upper point is about 0.31. With  $\theta$  set at 0.3, the difference between the means on a 9-point scale used to define equivalence, there is essentially no risk that a difference outside the equivalence bounds would occur that would provide support for an equivalence hypothesis. In drug testing, if large variances do occur, it would be worth examining the cause of variability before conducting equivalence hypothesis testing with any method including the TOST. The second point is that the fact that the rejection region flares out at very large variances simply means that there are multiple roots to the solution of the rejection region boundary equation provided by Ennis and Ennis

(2009). Whether this is counterintuitive depends on one's intuition. However, these roots are implied immediately when we write down the null and alternative equivalence hypotheses. If one wants to avoid the possibility of multiple roots, then the solution is to modify the hypothesis, not to invent an inconsistent test. For example, if we consider the statement  $x^2 - 4 = 0$  there are two roots. If one is only interested in a positive root, one should say  $x^2 - 4 = 0$ ,  $x > 0$ . As Berger and Hsu (1996b) explain, a more consistent approach is to put an upper bound on the variance as part of the hypothesis rather than using the test itself to control the effect of unbounded variance artificially. Unfortunately this latter choice is made when the TOST is used.

This fact about the TOST has real world consequences. There may be cases when a generic drug is equivalent to a brand name drug, but the TOST does not reject the null hypothesis of nonequivalence. Using the bioequivalence example discussed earlier, for which  $x$  was 0.04 and the standard error was 0.103, the test based on (11) with an adjusted noncentrality parameter was significant using an  $\alpha$  level of 0.05. However, the TOST would not reject the null hypothesis at  $\alpha = 0.05$ .

#### 4.2 Other Methods

Ennis and Ennis (2009) have discussed the use of the noncentral  $F$  distribution function as an alternative to the adjusted noncentral chi-square method discussed earlier. This approach is identical to a difference in noncentral  $t$  distribution functions but is not the same as a test due to Anderson and Hauck (1983). The test of Anderson and Hauck is an approximation for the case where the variance is unknown and is based on a difference of two  $t$  distribution functions. This test does not control for the Type I error at a specified level and is somewhat liberal. There are two factors contributing to this liberality. One is their specification for the unknown noncentrality parameter used in their test and the second is their use of the central  $t$  distribution function. Two other methods, due to Brown *et al.* (1995) and Berger and Hsu (1996a) involve making adjustments to the TOST's rejection region to improve power and prevent the Type I error from exceeding a stated  $\alpha$  level.

## 5. Summary

In equivalence hypothesis testing, open intervals are used to define the meaning of equivalence. The null hypothesis is that the items are nonequivalent as determined by the appropriate definition. The alternative hypothesis is that the items tested are equivalent within the defined boundaries and an exact test exists of the null hypothesis for binomial and normally distributed data with known variance. When the variance is unknown an adjustment to the noncentrality parameter removes a potential problem with liberality. This new method is preferable to the TOST as it is a direct test of the null hypothesis. In fact, the TOST is fundamen-

tally flawed because it does not account for the correlation between its component statistics and it shuts off the test power in a manner inconsistent with the null hypothesis.

As we conclude, we note that more research is needed on equivalence hypothesis testing and in particular we note that it would be useful to develop new approaches for multi-parameter cases.

---



---

## References

- Anderson, S., & Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Comm. Statist. Theory Methods*, 12, 2663-2692.
- Ashby, F.G., & Ennis, D.M. (2007). Similarity measures. *Scholarpedia*, 2(12):4116
- ASTM E 1958 (2006). *Standard Guide for Sensory Claim Substantiation*. ASTM International.
- Berger, R.L., & Hsu, J.C. (1996a). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4), 283-302.
- Berger, R.L., & Hsu, J.C. (1996b). Rejoinder. *Statistical Science*, 11(4), 315-319.
- Brown, L.D., Hwang, J.T.G., & Munk, A. (1995). An unbiased test for the bioequivalence problem. Technical Report, Cornell University.
- Ennis, D. M. (1988). Confusable and discriminable stimuli: Comments on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General*, 117(4), 408-411.
- Ennis, D. M. (1992). Modeling similarity and identification when there are momentary fluctuations in psychological magnitudes. In F. Gregory Ashby (Ed.), *Multidimensional Models of Perception and Cognition*, Mahwah, NJ: Lawrence Erlbaum Associates.
- Ennis, D.M. (2008). Tables for parity testing. *Journal of Sensory Studies* 23 80-91.
- Ennis, D.M., & Ashby, F.G. (1993). The relative sensitivities of same-different and identification judgment models to perceptual dependence. *Psychometrika*, 58(2), 257- 279.
- Ennis, D.M., & Ennis, J.M. (2009). Hypothesis testing for equivalence based on symmetric open intervals. *Communications in Statistics*, 38(11), 1792-1803.
- Ennis, D. M., & Johnson, N. L. (1993). Thurstone-Shepard similarity models as special cases of moment generating functions. *Journal of Mathematical Psychology*, 37(1), 104-110.
- Ennis, D. M., Palen, J., & Mullen, K. (1988). A multidimensional stochastic theory of similarity. *Journal of Mathematical Psychology*, 32(4), 449-465.
- Meredith, M.P., & Heise, M.A. (1996). Comment. *Statistical Science*, 11(4), 304-306.
- Schuurmann, D.J. (1996) Comment. *Statistical Science*, 11(4), 312-313.
- Schuurmann, D.J. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics*, 37, 617.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.
- Westlake, W.J. (1976). Symmetric confidence intervals for bioequivalence trials. *Biometrics*, 32, 741-744.
- Westlake, W.J. (1981). Response to T.B.L. Kirkwood: bioequivalence testing – a need to rethink. *Biometrics*, 37, 589-594.