Journal of the Korean
Statistical Society
Vol. 23, No. 1, 1994

Equivalence Testing as an Alternative to Significance Testing †

Myung-Hoe Huh [1]

ABSTRACT

Sometimes a researcher with a view of conventional significance testing rejects his/her hypothesis, even though it could have not been rejected with a smaller sample. This can be a logical dilemma for a researcher who wants to "prove" a hypothesis rather than to show discrepancy from a null hypothesis. In this study, a new testing paradigm called *equivalence testing via confidence interval* will be developed so that it is suitable for the purpose of statistical proof.

KEYWORDS: Equivalence testing, Significance testing, Several means.

1. INTRODUCTION

Suppose that a reseacher is planning an experiment to show that A is not practically different from B with the postulation that sample A consisting of n observations is from $N(\mu_A, \sigma^2)$ and sample B of the same size from $N(\mu_B, \sigma^2)$. In such a circumstance, a significance testing formulation is to set the null and

[1]Department of Statistics, Korea University, Seoul 136-701, Korea.
†This research is supported by NON-DIRECTED RESEARCH FUND, Korea Research Foundation, 1993.

alternative hypothesis as $H_0 : \mu_A = \mu_B$ and $H_1 : \mu_A \neq \mu_B$, respectively. After collecting data, he/she rejects or accepts H_0 as directed by statistics textbooks. If H_0 is not rejected(i.e. accepted), he/she may insist that A is the same as B statistically.

Even though it is usual in practice, it is not logical by the following reasons. First, significance testing is a procedure for statistical disproof (not for statistical proof). Fisher (1960, p.16; 1st edition 1935) wrote that "the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation". Thus one can insist only whether the discrepancy between the data and the null hypothesis is statistically significant or not. Second, with significance tesing as described above, smaller sample size (or poorer research design) is more advantageous to the researcher. It is because p-value from a smaller sample is less significant for the same effect size. Thus it is a logical contradiction against the spirit of good science.

These kinds of problems are well known and studied in the area of drug research and approval with the following slight modifications (Weiner, 1981; Metzler and Huang, 1983; Westlake, 1988). Suppose that A is the original drug with a patent and B is a generic drug developed later. The maker of drug B wants to show that his/her generic drug is bioequivalent to the original drug A in order to get a government approval. Bioequivalence is usually measured and evaluated by the area under the curve(AUC) of some clinical variable from blood or urine samples with the margin of $\pm$ 20%. Due to intersubject variability, the bioequivalence reseach is typically performed by two-period crossover designs (A followed by B, or B followed by A). See Fisher and Wallenstein (1981) for crossover designs in medical research. Most of the bioequivalence testing procedures developed so far can be classified into one of the four rules (Metzler, 1991); Rule 1: t-based confidence intervals, Rule 2: Westlake's symmetric intervals, Rule 3: Bayesian analysis, Rule 4: Anderson-Hauck hypothesis test. Among the above rules, Rule 1 and 2 are based on confidence intervals. Recently, FDA(Food and Drug Administration) of the United States decided to officially support the confidence interval approach.

In this study, statistical results obtained in the bioequivalence area will be extended to more general problem of equivalence testing. Specifically, a testing procedure for equivalence of several means will be developed in Section 2 using confidence intervals. A numerical example is given in Section 3. Concluding remarks are given in Section 4.

2. EQUIVALENCE TESTING OF SEVERAL MEANS

2.1 Is the mean equivalent to a constant?

Suppose that $x_1, x_2, \cdots, x_n$ follow $N(\mu,\ \sigma^2)$ independently. And, we want to show that the mean μ is equivalent to μ_0. The following approval procedure is an obvious extension of Westlake(1972).

Step 1: Define the equivalence interval $\mathcal{E}$ for the mean μ as

$$\mathcal{E} = (\mu_0 - \delta, \mu_0 + \delta).$$

Step 2: Obtain the confidence interval $\mathcal{C}$ with confidence coefficient $1-\alpha$ for μ. Assuming that σ^2 is unknown,

$$\mathcal{C} = (\bar{x} - t_{\alpha/2}s/n^{1/2},\ \bar{x} + t_{\alpha/2}s/n^{1/2}),$$

where s^2 is the sample variance and $t_{\alpha/2}$ is the upper $\alpha/2$ quantile of t-distribution with $n-1$ degrees of freedom.

Step 3: If $\mathcal{C}$ is contained in $\mathcal{E}$, then the claim of equivalence can be accepted. Otherwise, the claim of equivalence should not be accepted.

Equivalence testing, as proposed here, requires pre-specification of the δ parameter denoting practical significance or equivalence limit. So one may think that the proposed procedure is not objective. But it is not quite true. Practical significance can be obtained by resorting to the common sense in each specific study area. For instance, in the drug research and approval area, it is generally agreed that δ is 20% of the general mean μ_0(Metzler, 1991, p.961). Another approach is possible using maximum effect-size, δ/σ , interpretation of meta-analysis (Song, 1992, p.17).

Also it can be noted that, from a view point of decision-rule, the equivalence testing is different from the significance testing; the former is based on the covering relationship between the confidence interval and the equivalence interval, while the latter is based on the covering relationship between a null hypothesis parameter point and the confidence interval.

This procedure for equivalence testing has the following properties.

Property 1a: When the true mean μ is at the boundary $\mu_0 \pm \delta$ of equivalence interval $\mathcal{E}$, probability of rejecting the equivalence claim (PREC) is approximately equal to $1 - \alpha/2$, for sufficiently large n.

Property 1b: When μ is at $\mu_1 \equiv \mu_0 \pm (\delta - t_{\alpha/2}\sigma/\sqrt{n})$, PREC is approximately equal to 0.5, for sufficiently large n.

Property 1c: When μ is at $\mu_2 \equiv \mu_0 \pm (\delta - 2t_{\alpha/2}\sigma/\sqrt{n})$, PREC is approximately equal to $\alpha/2$, for sufficiently large n.

2.2 Are two means equivalent to each other?

Suppose that $x_1, x_2, \cdots, x_m$ are a random sample from $N(\mu_1, \sigma^2)$, and that $y_1, y_2, \cdots, y_n$ are a random sample from $N(\mu_2, \sigma^2)$. Furthermore, two samples are assumed to be independent. The following is a proposed procedure for testing the hypothesis that two means μ_1 and μ_2 are equivalent to each other. This procedure has similar properties to Properties 1a-1c.

Step 1: Define the equivalence region $\mathcal{E}$ for two means μ_1 and μ_2 as

$$\mathcal{E} = \{(\mu_1, \mu_2) : -\delta \leq \mu_1 - \mu_2 \leq \delta\}.$$

Step 2: Obtain the confidence interval $\mathcal{C}$ with confidence coefficient $1-\alpha$ for the difference $\mu_1 - \mu_2$. Assuming that σ^2 is unknown,

$$\mathcal{C} = (\bar{x} - \bar{y} - t_{\alpha/2}s(1/m + 1/n)^{1/2},\ \bar{x} - \bar{y} + t_{\alpha/2}s(1/m + 1/n)^{1/2}),$$

where s^2 is the pooled sample variance and $t_{\alpha/2}$ is the upper $\alpha/2$ quantile of t-distribution with $m + n - 2$ degrees of freedom.

Step 3: If $\mathcal{C}$ is contained in $\mathcal{E}$, then the claim of equivalence can be accepted. Otherwise, the claim of equivalence should not be accepted.

2.3 Are several means equivalent among them?

Suppose y_{ij}'s are independent random observations from $N(\mu_i, \sigma^2)$, $i = 1, \cdots, I$, and $j = 1, \cdots, n_i$. and we are interested in the equivalence of I population means $\mu_1, \mu_2, \cdots, \mu_I$. For this, a reasonably extended version of the above procedures could be written as in the followings.

Step 1: Define the equivalence region $\mathcal{E}$ for I population means $\mu_1, \mu_2, \cdots, \mu_I$ by the requirement that the average squared pairwise differences is less than some specified value: Set

$$\mathcal{E} = \{(\mu_1, \mu_2, \cdots, \mu_I) : \sum_{i=1}^{I} \sum_{j=i+1}^{I} (\mu_i - \mu_j)^2 / \{I(I-1)/2\} \leq \delta^2\}.$$

Note that, for $\mu = \sum_i \mu_i / I$.

$$\sum_{i=1}^{I} \sum_{j=i+1}^{I} (\mu_i - \mu_j)^2 = \sum_i \sum_j \{(\mu_i - \mu) - (\mu_j - \mu)\}^2 / 2 = I \sum_i (\mu_i - \mu)^2,$$

where $\mu = \sum_i \mu_i / I$. Therefore the equivalence region can be written also as

$$\sum_i (\mu_i - \mu)^2 / (I-1) \leq \delta^2 / 2.$$

Step 2: Obtain the simultaneous confidence region $\mathcal{C}$ with confidence coefficient $1-\alpha$ for $(\mu_1, \mu_2, \cdots, \mu_I)$. Assuming that σ^2 is unknown,

$$\mathcal{C} = \{(\mu_1, \mu_2, \cdots, \mu_I) : \sum_i n_i \{(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) - (\mu_i - \mu)\}^2 / (I-1) \leq s^2 F_\alpha\},$$

where s^2 is the pooled sample variance and F_α is the upper α quantile of F-distribution with $I-1$ and $(n_1-1)+\cdots+(n_I-1)$ degrees of freedom.

Step 3: If $\mathcal{C}$ is contained in $\mathcal{E}$, then the claim of equivalence can be accepted. Otherwise, the claim of equivalence should not be accepted.

Especially when $n_1 = \cdots = n_I \ (= n)$, the equivalence is declared if and only if

$$[\sum_i \{(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 / (I-1)]^{1/2} + [(s^2/n) F_\alpha]^{1/2} \leq [\delta^2/2]^{1/2} \tag{1}$$

In this case, it is difficult to directly evaluate the PREC values because the computation involves noncentral-F random quantities. However, it can be shown that this procedure has some consistency properties due to the monotonically shrinking property of confidence regions with growing sample size.

3. A NUMERICAL EXAMPLE

Snedecor and Cochran(1967, p.259) presents a cooking experiment in which four types of fats are compared as to the amount absorbed in doughnuts. The experiment is carried out by completely randomized design. The data and ANOVA table is reproduced in Table 1. Note that p-value is less than 1%, suggesting that the means are significantly different in a statistical sense. But, suppose that experimenter's aim is to show that fat absorption is not influenced by the type of fats in a practical sense, where no practical difference is defined by $\sum\sum_{i<j}(\mu_i - \mu_j)^2/6 \leq 17^2 (= \delta^2)$. Note that 17 is about 10% of the grand mean.

Then follow three steps outlined in Section 2.3: with $\alpha = .1$, check Eq.(1)

$$[\sum_i \{(\bar{y}_{i.} - \bar{y}_{..})^2/(I-1)]^{1/2} + [(s^2/n)F_{.1}]^{1/2} \leq [\delta^2/2]^{1/2}$$

$$\underset{9.54}{(545.5/6)^{1/2}} + \underset{6.33}{(100.9/6 \times 2.38)^{1/2}} \leq \underset{12.02}{(17^2/2)^{1/2}}. \quad \text{No!}$$

So, in this case, Eq. (1) does not hold. Therefore we should not accept researcher's hypothesis of equivalence. However, he/she can possibly claim the equivalence with a sample large enough

$$(s^2/n)F_{.1} \leq (12.02 - 9.54)^2$$

or

$$n \geq s^2/2.48^2 F_{.1} = 16.41 F_{.1},$$

which is roughly equal to $n \geq 35$. (Here, $F_{.1}$ is the upper 0.1 quantile of F-distribution with 3 and $4(n-1)$ degrees of freedom.)

4. CONCLUDING REMARKS

The underlying idea of this study can be applied to other testing problems such as the equivalence of correlations. Moore(1985, p.324) wrote that "with 1000 observations, an observed correlation of only $r = 0.08$ is significant

evidence at the $\alpha = 0.01$ level that the correlation in the populations is not zero but positive. $\cdots$ We might conclude that for practical purpose there is no association between these variables, even though we are confident (at the 1% level) that this is not literally true", suggesting the need to differentiate practical significance from statistical signficance.

From Properties 1b, 1c, and their extensions, we can see that a larger sample is more advantageous to researchers who want to claim the equivalence (if it is true), in line with the results of Metzler(1991) with crossover trials for the case of two means. This study, however, did not cover PREC(probability of rejecting the equivalence claim) or power study for the case of several means, leaving it to someone else or future study.

Table 1. Doughnut Data and its ANOVA

	Fat 1	Fat 2	Fat 3	Fat 4	
	164	178	175	155	
	172	191	193	166	
	168	197	178	149	
	177	182	171	164	
	156	185	163	170	
	195	177	176	168	
Mean	172	185	176	162	173.8
S.D.	13.3	7.8	9.9	8.2	

Source	D.F.	SS	MS	F	p-value
Fats	3	1636.5	545.5	5.41	0.0069
Error	20	2018	100.9		
Total	23	3654.5			

REFERENCES

(1) Fisher, A.C. and Wallenstein, S.(1981) Crossover designs in medical research, in *Statistics in the Pharmaceutical Industry* (edited by C.R. Buncher and J-Y. Tsay). Dekker, New York.

(2) Fisher, R.A.(1960; 1st edition 1935) *The Design of Experiments.* 7th edition. Hafner, New York.

(3) Metzler, C.M.(1991) Sample size for equivalence studies, *Statistics in Medicine*, 10, 961-970.

(4) Metzler, C.M. and Huang, D.C.(1983) Statistical methods for bioavailability, *Clinical Research Practices and Drug Regulatory Affairs*, 1, 109-132.

(5) Moore, D.S.(1985) *Statistics: Concepts and Controversies*, Second edition. W.H. Freeman and Company, New York.

(6) Snedecor, G.W. and Cochran, W.G.(1967) *Statistical Methods*, 6th edition. Iowa State University Press, Ames.

(7) Song, H.H.(1992) *Meta-Analysis* (written in Korean). Freedom Academy, Seoul.

(8) Weiner, D.L.(1981) Design and analysis of bioavailability studies, in *Statistics in the Pharmaceutical Industry* (edited by C.R. Buncher and J-Y. Tsay). Dekker, New York.

(9) Westlake, W.J.(1972) Use of confidence intervals in analysis of comparative bioavailability trials, *Journal of Pharmaceutical Sciences*, 61, 1340-1341.

(10) Westlake, W.J.(1988) Bioavailability and bioequivalence of pharmaceutical formulations, in *Biopharmaceutical Statistics for Drug Development* (edited by K.E. Peace). Dekker, New York.