■ 연구논문

Probability of Rejection Curve for Equivalence Testing Procedure†

Nae Kyung Sung

Dept. of Statistics, Ewha Womans University

Abstract

We investigate the small-sample behavior of the probability of rejection curves and its perfomance for a equivalence testing procedure based on confidence intervals which was developed with a motivation from bioequivalence studies. This type of equivalence studies are conducted frequently in pharmaceutical industries to compare the relative bioavailabilty of two formulations of a drug and can be applied various fields where assurance of quality equivalence is needed. From the Monte-Carlo simulation results we suggest proper sample sizes for the equivalence testing procedure.

1. Introduction

In the conventional statistical hypotheses testing, a researcher's own purpose is to reject the null hypothesis in most cases. Thus after establishing appropriate null and alternative hypotheses, one may conclude, based on the sample collected, that he reject or fail to reject the null hypothesis. Unfortunately, however, failing to reject a null hypothesis is not *proof* that the null hypothesis is true. It denotes only that there is not sufficient evidence to conclude that the null hypothesis is false. This testing principle imposes a serious logical problem on researchers whose purposes are to show that the null hypothesis is true.

This kind of arguments on the logical problem present in statistical tests of hypotheses is not new. Note, for instance, that the same argument holds for the Pearson's goodness-of-fit test. Refer to Inman (1994) for more details on the

† 이 연구는 1994년도 이화여자대학교 교내연구비의 지원을 받았음.

debates exchanged between K. Pearson and R. A. Fisher. In this context some statisticians call the goodness-of-fit test as the badness-of-fit test.

Consider a simple situation where we wish to show equivalence of two population means. Assume as usual that we draw a sample of size n_1 from $N(\mu_1, \sigma^2)$ and another of size n_2 from $N(\mu_2, \sigma^2)$, and that both samples are independent. In this case we set H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$. As noted above failing to reject the null hypothesis does not warrant statistically that two population means are equal. In addition, since increasing the sample size tends to yield more power of the tests, a small sample size or a less powerful research design is more advantageous, as long as the power of the test is concerned, for researchers who wish to show the equivalence of two population means.

Difficulty of statistical assessment of equivalence of two population means has been well-known, especially, in pharmaceutical industries and regulartory agencies of government such as the U. S. Food and Drug Administration (FDA) dealing with approval of a newly-developed pharmaceutical product. In this arena one must make decisions as to bioequivalence of pharmaceutical products manufactured by different pharmaceutical firms.

Motivated by bioequivalence studies prevalent in pharmaceutical industries given above, Huh (1994) proposed a testing procedure, as an alternative to the usual significance testing procedure, for general two-sample situations, which he called "equivalence testing via confidence interval" based on the t-based confidence interval method, useful for showing that two population means are equivalent. Though Huh's proposal is a direct extension of the t-based confidence interval method for bioequivalence, he fortified the testing procedure with sound large-sample properties.

It is apparent that this type of equivalence testing procedure can be applied to various fields where assurance of quality equivalence is needed.

In this article we observe and assess the performance of Huh's equivalence testing procedure when the sample size is small, which we believe is the most practical situation. The criterion we choose is the behavior of PR curves produced by the Monte-carlo simulation study. These curves are used for choosing the appropriate sample size for a equivalence study.

We focus only on the one-sample case in view of relative availability introduced in Section 2. Note that in view of statistical testing context the two-sample problem reduces to the one-sample problem.

2. Bioequivalence

Bioequivalence evaluations involve *relative bioavailability*, or *bioavailability* in short, which means a comparison of two or more dosage forms in terms of their relative rate and extent of absorption, according to Meyer (1988). Two dosage forms do not differ significantly in their rate and extent of absorption are termed *bioequivalent*. Hence, two bioequivalent formulations must make the active ingredient available in the circulating blood and should not differ in their therapeutic efficacy. The area under the concentration-time curve (AUC) is a favorite measure of bioavailability as Metzler (1974) pointed out.

Although it is not easy to tell how much difference in relative bioavailability is a limit in showing bioequivalence, a rule of thumb adopted in practice has been that a new test formulation is bioequivalent to an established reference formulation if the difference is less than 20% of the mean bioavailability of the reference formulation. This interval of acceptable relative bioavailability is called the acceptance interval.

Since a bioequivalence decision is based on sample results, it is natural that the bioequivalence decision rule should have solid statistical backgrounds. Among many decision rules appeared until recently, a method of using the usual t-based confidence intervals for evaluating bioequivalence suggested by Metzler (1988, 1991) and Westlake (1972, 1976) is typical, which was approved officially by the FDA.

Other bioequivalence decision rules include a Bayesian approach given by Rodda and Davis (1980) and Mandallaz and Mau (1981), and the Anderson-Hauck hypothesis test formulated by Anderson and Hauck (1983).

For the t-based confidence interval method, one computes the conventional t-based confidence interval for the relative bioavailabilaty of the two formulations. If this confidence interval is contained in the preassigned acceptable interval, we decide bioequivalence.

One way to characterize these rules is to consider the probability that they will reject bioequivalence for a given value of the true relative bioavailability. If these probabilities of rejection are computed across a range of relative bioavailabilities, a probability of rejection curve (PR curve) can be drawn to show the characteristics of the rules. The PR curves depend on the variability of the AUCs, the sample size, the level of significance, and the level of protection against incorrectly deciding bioequivalence at the end points of the acceptance interval. A reasonable choice of the protection level is known to be 95%.

3. Simulations of PR Curves

Suppose that $X_1, X_2, \cdots, X_n$ are a random sample of size n from a normal distribution $N(\mu, \sigma^2)$. We wish to infer that the population mean μ is equivalent to μ_0. In this case the equivalence testing procedure is given as follows: we first obtain the usual t-based confidence interval $C=(\bar{x}-t_{n-1,\alpha/2}\ s/\sqrt{n}\ ,\ \bar{x}+t_{n-1,\alpha/2}\ s/\sqrt{n})$ for μ with confidence coefficient $1-\alpha$, where s^2 is the sample variance and $t_{n-1,\alpha/2}$ is the upper $\alpha/2$ quantile of the t distribution with $n-1$ degrees of freedom. If C is contained in a predefined acceptance interval $E=(\mu_0-\delta,\ \mu_0+\delta)$, $\delta>0$, then we accept the claim of equivalence. Otherwise, the claim of equivalence should not be accepted.

In order to investigate the small sample behavior of the equivalence testing procedure via confidence interval by the Monte-Carlo simulation study, we let μ be 1. In terms of bioequivalence study μ can be considered as *relative availability*. In this case the hypotheses become H_0: $\mu=1$ versus H_1: $\mu\neq1$.

A single set of n random observations from a normal distribution with mean 1 and standard deviation σ is generated. We used RANNOR, a normal random number generator in SAS 6.04 software. We selected 0.1, 0.2, and 0.3 as the values of σ. These values correspond to 10%, 20%, and 30% of coefficients of variation (CV), respectively. The sample sizes considered here are from 5 to 30 with an increment of 5.

For each sample generated, the usual t-based confidence interval is computed for a given value of significance level. The values of significance level we considered are 0.05 and 0.1. This confidence interval is compared to the predetermined acceptance interval $(\mu_0-\delta,\ \mu_0+\delta)$. As the values of δ we choose 0.1, 0.15, 0.20, and 0.25. For each combination of simulation design parameters, we repeat prescribed simulation steps 1,000 times and relative frequencies of rejecting the equivalence claim are calculated. Finally we applied the spline smoothing routine to values of empirical probabilities to produce the PR curves.

4. Behavior of PR Curves

Typical forms of PR curves for some combinations of parameters are given in 〈Figure 1〉 to 〈Figure 6〉.

A desirable PR curve, in general, should approach to zero probability of rejecting equivalence rapidly when the relative availability is around 1 and should approach to probability one when the relative availability is outside the

preassigned acceptance interval.

From the simulation results we observe as expected that the PR curves move downward (i) as δ increases, (ii) to 0 around $\mu=1$ as n becomes large, and (iii) as α increases. Since Huh developed the equivalence testing procedure according to the FDA guidelines that upto 20% difference of mean relative availability is allowed with a protection level of 95%, Figures show common phenomena that all of the PR curves pass through $1-\alpha/2$ point of probability of rejecting equivalence when the true mean μ is at the boundary $\mu_0 \pm \delta$. This is, however, rather surprising because the equivalence testing procedure via confidence interval Huh proposed is based on large-sample theory.

〈Figure 1〉 shows four PR curves with $\delta=0.1$, 0.15, 0.2, and 0.25 when the sample size $n=5$, the coefficient of variation CV=10%, and the significance level $\alpha=0.05$. Among these four PR curves, only the PR curve for $\delta=0.25$ has a desirable form. The PR curve for $\delta=0.2$ appears to be not bad, but barely acceptable. Remaining two PR curves are not recommendable.

The parameter values for 〈Figure 2〉 are the same as 〈Figure 1〉 except that α is set to 0.1. 〈Figure 2〉 shows a similar pattern compared to 〈Figure 1〉.

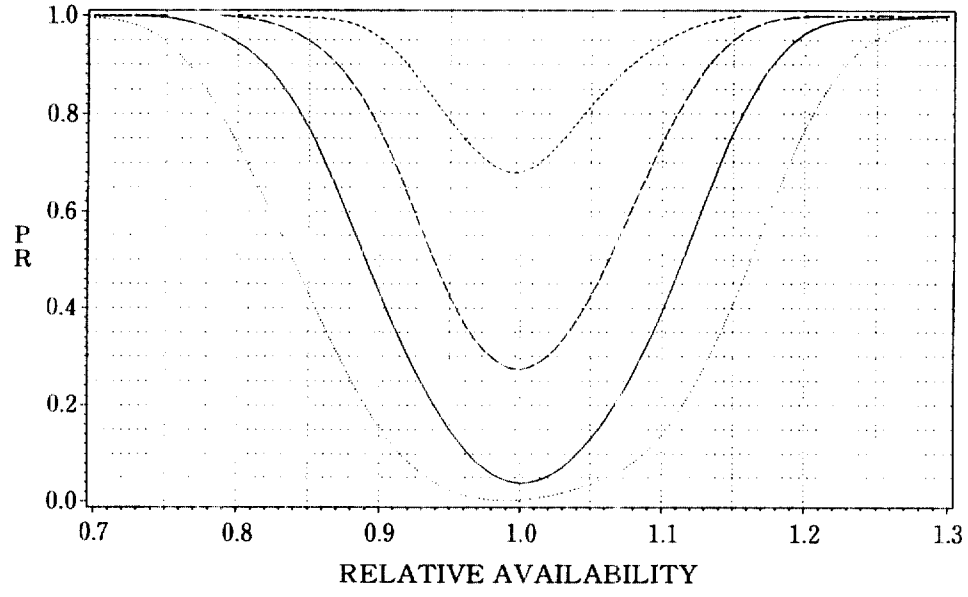
〈Figure 3〉 and 〈Figure 4〉 are the cases of $n=10$. Only α values are different. These Figures in common suggest that the PR curves for $\delta=0.15$, 0.2, and 0.25 behave very well.

〈Figure 5〉 and 〈Figure 6〉 are the case of $n=20$ and CV=20%. Namely observations are more dispersed than those in 〈Figure 1〉 to 〈Figure 4〉. In these cases the PR curves for $\delta=0.2$ and 0.25 performed satisfactorily.

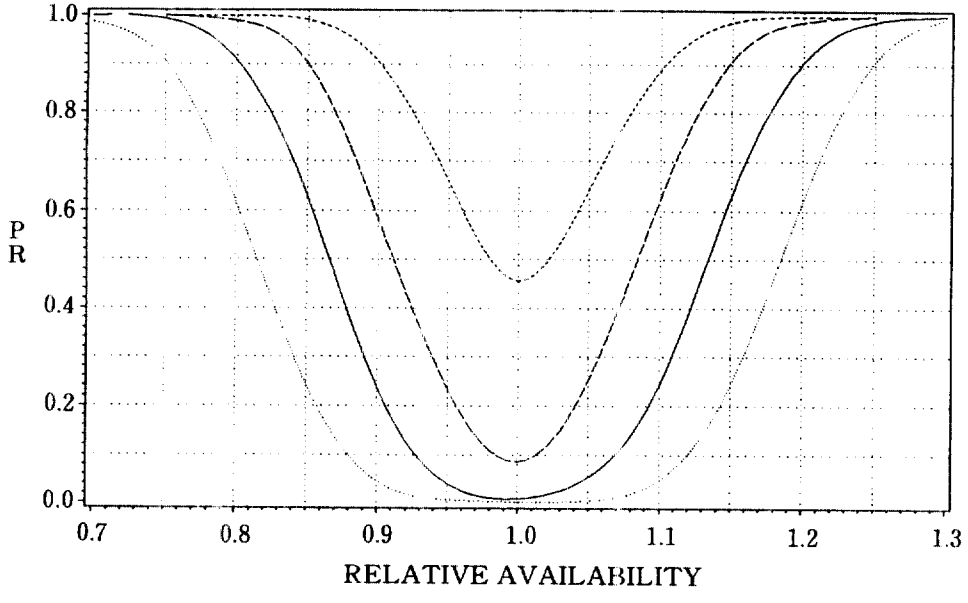
〈 Table 1 〉 Recommended sample size for equivalence testing

$\alpha=0.05$	δ			
CV	0.10	0.15	0.20	0.25
10%	≥25	≥10	≥10	≥ 5
20%	–	≥30	≥20	≥15
30%	–	–	–	≥25

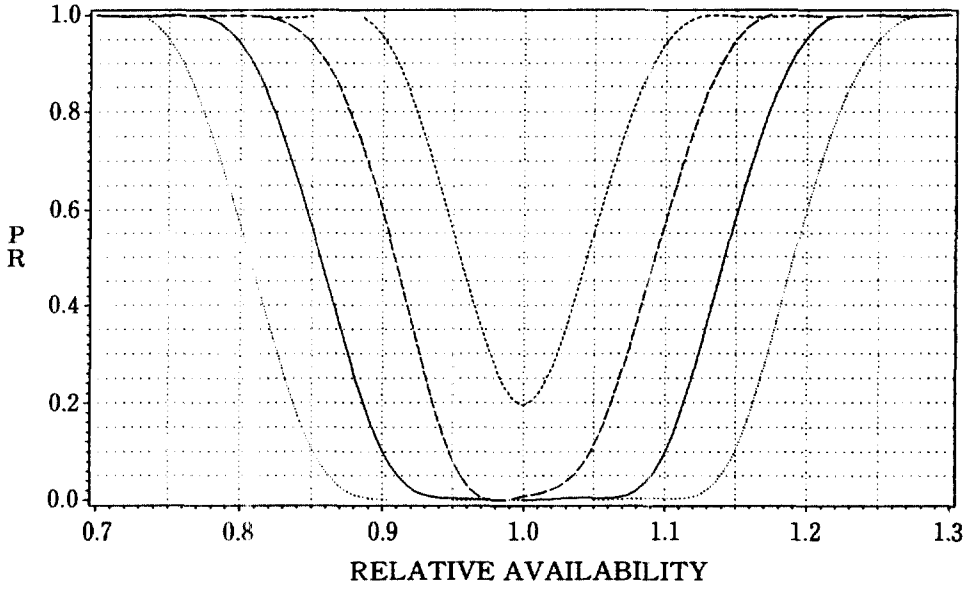
$\alpha=0.10$	δ			
CV	0.10	0.15	0.20	0.25
10%	≥15	≥10	≥ 5	≥ 5
20%	–	≥20	≥15	≥10
30%	–	–	≥30	≥20



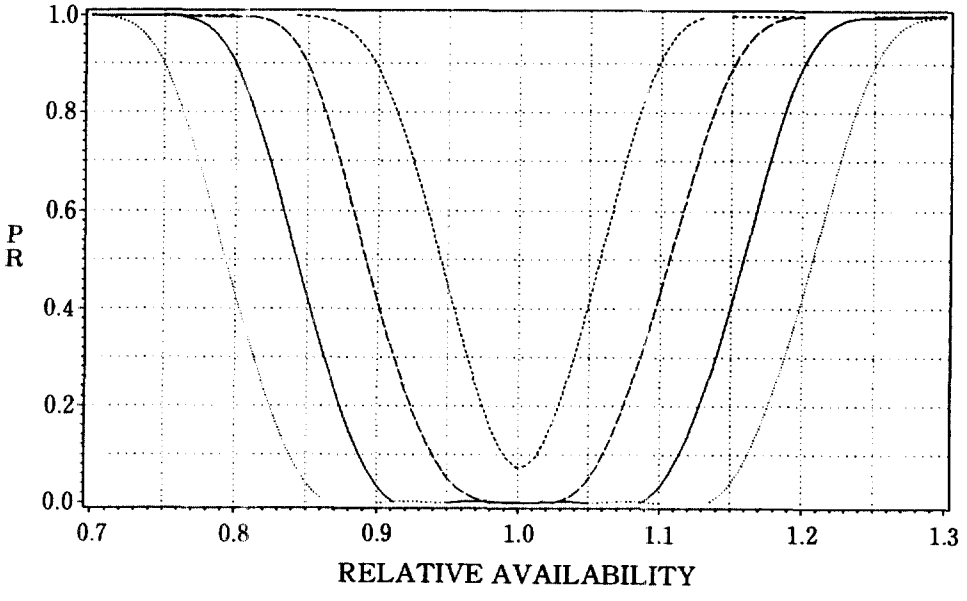
〈 Figure 1 〉 PR curves: n=5, alpha=0.05, CV=10%, delta=0.1, 0.15, 0.2, 0.25



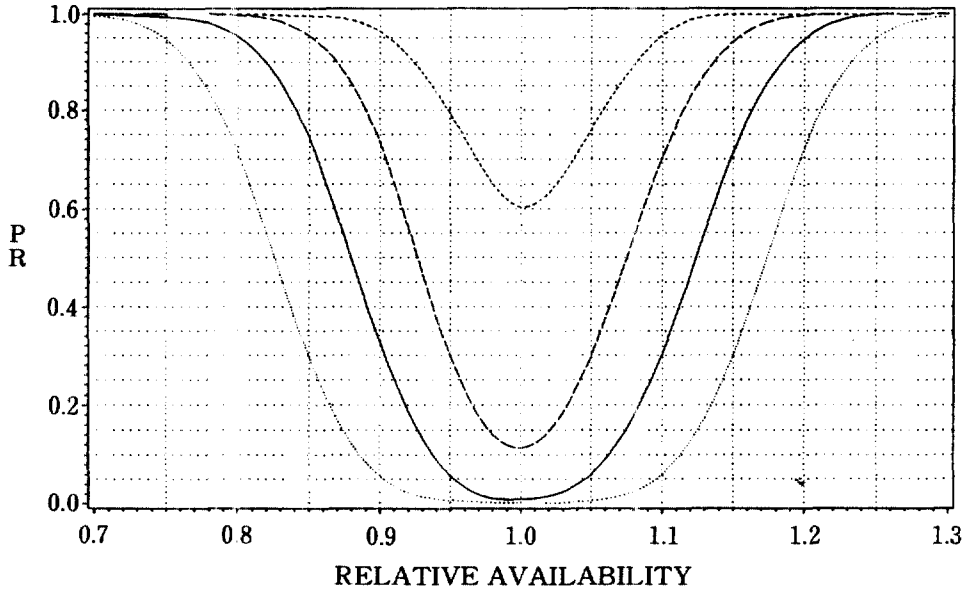
〈 Figure 2 〉 PR curves: n=5, alpha=0.10, CV=10%, delta=0.1, 0.15, 0.2, 0.25



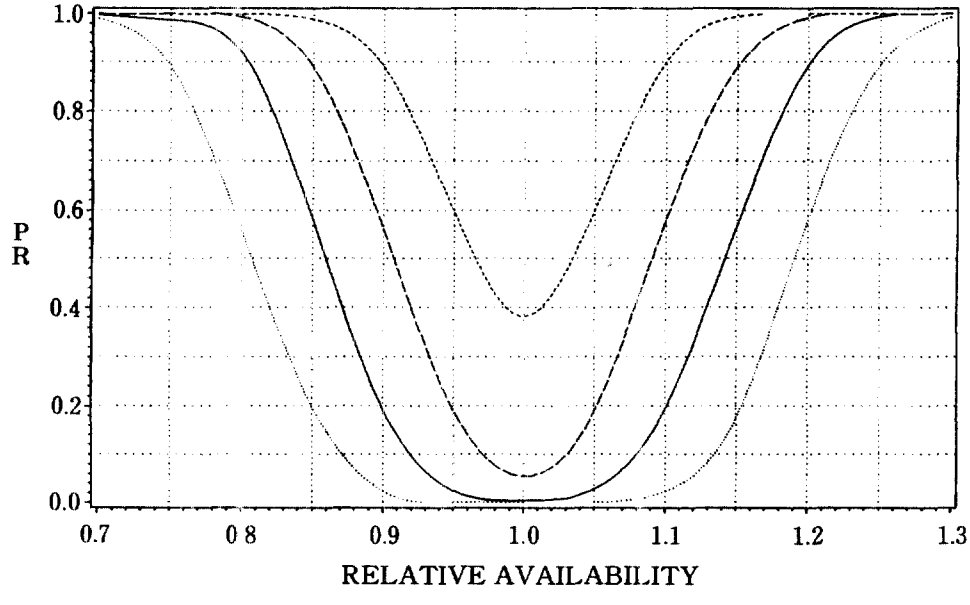
〈 Figure 3 〉 PR curves: n=10, alpha=0.05, CV=10%, delta=0.1, 0.15, 0.2, 0.25



〈 Figure 4 〉 PR curves: n=10, alpha=0.10, CV=10%, delta=0.1, 0.15, 0.2, 0.25



〈 Figure 5 〉 PR curves: n=20, alpha=0.05, CV=20%, delta=0.1, 0.15, 0.2, 0.25



〈 Figure 6 〉 PR curves: n=20, alpha=0.10, CV=20%, delta=0.1, 0.15, 0.2, 0.25

The behavior of PR curves is not always acceptable, especially when the sample size is very small and CV is large. It also depends on the significance level. 〈Table 1〉 shows a summary of recommended sample sizes for various cases. From 〈Table 1〉, one may see that the equivalence testing procedure requires rather large sample sizes in most cases.

Acknowledgements

The author wishes to thank the anonymous referee for a careful review.

References

[1] Anderson, S. and Hauck, W. W. (1983), "A new procedure for testing equivalence in comparative bioavailability and other clinical trials," *Communications in Statistics – Theory and Methods*, Vol. 12, pp. 2663–2692.

[2] Huh, M. H. (1994), "Equivalence testing as an alternative to significance testing," *Journal of the Korean Statistical Society*, Vol. 23, pp. 199–206.

[3] Inman, H. F. (1994), "Karl Pearson and R. A. Fisher on statistical tests: A 1935 exchange from Nature," *American Statistician*, Vol. 48, pp. 2–11.

[4] Mandallaz, D. and Mau, J. (1981), "Comparison of different methods for decision making in bioequivalence assessment," *Biometrics*, Vol. 37, pp. 213–222.

[5] Metzler, C. M. (1974), "Bioavailabilty–a problem in equivalence," *Biometrics*, Vol. 30, pp. 309–317.

[6] Metzler, C. M. (1988), "Statistical methods for deciding bioequivalence of formulations," in *Drug absorption from sustained release formulations* edited by Yacobi, A. and Halperin-Walega, E., Pergamon Press, pp. 217–238.

[7] Metzler, C. M. (1991), "Sample sizes for bioequivalence studies," *Statistics in medicine*, Vol. 10, pp. 961–970.

[8] Meyer, M. C. (1988), "Bioavailability of drugs and bioequivalence," *Encyclopedia of Pharmaceutical Technology*, Vol. 1, pp. 477–494.

[9] Rodda, B. E. and Davis, R. L. (1980), "Determining the probability of an important difference in bioavailability," *Clinical Pharmacology and Therapeutics*, Vol. 28, pp. 252–257.

[10] Westlake, W. J. (1972), "Use of confidence intervals in analysis of comparative bioavailability trials," *Journal of Pharmaceutical Sciences*, Vol. 61, pp. 1340–1341.

[11] Westlake, W. J. (1976), "Symmetric confidence intervals for bioequivalence trials," *Biometrics*, Vol. 32, pp. 741–744.