The Korean Communications
in Statistics Vol. 5, No. 2, 1998
pp. 491-501

A General Procedure for Testing Equivalence [1]

Sung Nae Kyung [2]

Abstract

Motivated by bioequivalence studies which involve comparisons of pharmaceutically equivalent dosage forms, we propose a more general decision rule for showing equivalence simultaneously between multiple means and a control mean. Namely, this testing procedure is concerned with the situation in that one must make decisions as to the bioequivalence of an original drug product and several generic formulations of that drug. This general test is developed by considering a spherical confidence region, which is a direct extension of the usual t-based confidence interval rule formally approved by the U.S. Food and Drug Administration. We characterize the test by the probability of rejection curves and assess its performance via Monte-Carlo simulation. Since the munufacturer's main concern is the proper choice of sample sizes, we provide optimal sample sizes from the Monte-Carlo simulation results. We also consider an application of the generalized equivalence test to a repeated measures design.

1. Introduction

As has been pointed out in Sung (1994), 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 debates exchanged between K. Pearson and R. A. Fisher.

---

1) Research was supported by 1997 Ewha Faculty Research Fund.
2) Associate Professor, Department of Statistics, Ewha Womans University, Seoul 120-750, Korea.

Accordingly it is quite apparent that we need a rather different decision rule if we wish to show that the null hypothesis is true.

This type of problem is prevalent, especially, in pharmaceutical industries and regulatory agencies such as the U.S. Food and Drug Administration (FDA) dealing with approval of newly-developed generic pharmaceutical products. In this arena one must make decisions as to bioequivalence of pharmaceutical products manufactured by different pharmaceutical firms.

The FDA officially defined that *bioavailability* refers to the rate and extent to which the active drug ingredient or therapeutic moiety is absorbed from a drug product and becomes available at the site of drug action. Also, as Meyer (1988) pointed out, *relative bioavailability* refers to a comparison of two or more dosage forms in terms of their relative rate and extent of absorption.

Accordingly, two dosage forms that 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) discussed.

Much efforts have been undertaken to develop a decision rule based on statistical principles deciding if a test formulation is bioequivalent to a reference formulation.

Among the several decision rules proposed so far by Metzler (1988, 1991), Westlake (1972, 1976), Rodda and Davis (1980), Mandallaz and Mau (1981), and Anderson and Hauck (1983), the simplest and most intuitive test for bioequivalence is to use the usual t-based confidence intervals:

Let $X_1, X_2, \cdots, X_n$ be a random sample from $N(\mu, \sigma^2)$. We wish to infer that μ is equivalent to a known μ_0. Note again that in this situation the purpose of inference is to show that the null hypothesis is true. In this case we obtain the t-based $100(1-\alpha)\%$ confidence interval C with endpoints $\bar{x} \pm t_{n-1,\alpha/2}(s/\sqrt{n})$, where $\bar{x}$ is the sample mean and s^2 is the sample variance. If C is contained in a predefined acceptance interval $E=(\mu_0-\delta, \mu_0+\delta)$, $\delta>0$, then we accept the claim of bioequivalence. Otherwise the claim of bioequivalence should not be accepted.

Extending the concept of decision rules for showing bioequivalence, Huh (1994) proposed a new testing procedure of showing equivalence among k independent samples in a more general context with H_0: $\mu_1=\mu_2=\cdots=\mu_k$, where μ_i, $i=1,\cdots, k$, is the ith population mean, and Sung (1994) suggested proper sample sizes for the independent two-sample case via Monte-Carlo simulation.

Though Huh's decision rule is directly applicable to arbitrary number of independent samples, it is unfortunately not easy to explore its behavior exactly, nor to provide an appropriate statistical assessment of its performance even with the aid of computer simulation.

In this paper we propose a generalized equivalence testing procedure to deal with a situation

where several test formulations are compared simultaneously to a reference formulation. This test is a direct extension of the usual t-based confidence interval rule and can also be viewed as a modified version of Huh's rule.

We characterize the test by the *probability of rejection* (PR) curves and assess its performance via Monte-Carlo simulation and determine optimal sample sizes for equivalence test. We also consider an application of the test to a simple repeated measures design.

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

2. Simultaneous Test for Equivalence

Suppose that we draw a sample of size n from each of k independent normally distributed populations, respectively; *i.e.*, we assume $Y_{ij} \sim N(\mu_i, \sigma^2)$, $i=1,\cdots,k$; $j=1,\cdots,n$. Denote the value of the reference formulation as μ_c.

We wish to provide a decision rule to test equivalence: H_0: $\mu_1=\mu_2=\cdots=\mu_k=\mu_c$.

We predefine the acceptance region as follows:

$$E=\left\{(\mu_1,\cdots,\mu_k): \sum_{i=1}^{k}(\mu_i-\mu_c)^2/k \le \delta^2\right\}, \quad \delta \ge 0$$

That is, E consists of k-dimensional vectors $(\mu_1,\cdots,\mu_k)$ with distance less than or equal to $\delta\sqrt{k}$ from $(\mu_c,\cdots,\mu_c)$.

From the fact that $\sum_{i=1}^{k}(\overline{Y_i}-\mu_i)^2/k \sim (s^2/n)F_{k,k(n-1)}$, where s^2 is the pooled sample variance, the $100(1-\alpha)$ confidence region for $(\mu_1,\cdots,\mu_k)$ can be constructed as follows:

$$C=\left\{(\mu_1,\cdots,\mu_k): \sum_{i=1}^{k}(\overline{Y_i}-\mu_i)^2/k \le (s^2/n)F_{k,k(n-1),\alpha}\right\}$$

Hence if $(\mu_1,\cdots,\mu_k)$ satisfying C also satisfies E, then we accept the null hypothesis; that is, we accept the claim of equivalence.

Since $(\mu_1,\cdots,\mu_k)$ satisfying E are interior points of a sphere with center $(\mu_c,\cdots,\mu_c)$ and radius $\sqrt{k(s^2/n)F_{k,k(n-1),\alpha}}$, one can obtain the following simple rule:

$$\textit{Accept equivalence if } \sqrt{\frac{\sum(\overline{Y_i}-\mu_c)^2}{k}} + \sqrt{\frac{s^2}{n}F_{k,k(n-1),\alpha}} \le \delta \tag{2.1}$$

The decision rule using the usual t-based confidence intervals for evaluating bioequivalence

suggested by Metzler (1974) and Westlake (1972) is a special case of (2.1) for which $k=1$.

3. Simulation: PR Curves and Sample Sizes

In order to assess the performance of the bioequivalence testing rule (2.1), we utilize a Monte-Carlo simulation study.

For convenience we let μ_c to be 1 and $k \geq 2$ sets of n random variates with mean d and standard deviation σ are generated using a normal random number generator RANNOR in SAS 6.11 software.

For $k=1(1)5$, we selected 0.1(0.1)0.3 as σ, $-0.3(0.01)0.3$ as d, 2(2)30, 35, 40 as the sample size n, 0.1(0.05)0.25 as δ, and 0.1 and 0.05 as the significance level α. Note, in particular, that the values of σ correspond to 10% to 30% of the coefficients of variation (CV), respectively.

For each combination of simulation parameters 10,000 runs are repeated and relative frequencies of rejecting the equivalence claim are calculated to produce the PR curves.

Specification of $(\mu_1, \cdots, \mu_k)$ needs a special attention.

In order to generate μ_i's, $i=1,\cdots,k$, we used the transformation of Mustard (1964) to spherical coordinates, namely,

$$\begin{aligned}
\mu_1 &= 1 + d\sin\theta_1 \sin\theta_2 \cdots \sin\theta_{k-1} \\
\mu_2 &= 1 + d\sin\theta_1 \sin\theta_2 \cdots \sin\theta_{k-2}\cos\theta_{k-1} \\
&\vdots \\
\mu_{k-1} &= 1 + d\sin\theta_1 \cos\theta_2 \\
\mu_k &= 1 + d\cos\theta_1
\end{aligned}$$

where d is a candidate value of $\delta\sqrt{k}$.

Typical forms of PR curves for some combinations of parameters are given in Figure 1 to Figure 8.

Figures 1 to 3 show the behavior of the PR curves by varying the sample size for $\delta=0.2$, $\alpha=0.1$, $CV=10\%$. Specifically, Figure 1 is for $k=1$ and so are Figure 2 and Figure 3 for $k=2$, 3, respectively.

Figures 4 to 6 show identical situations as in Figures 1 to 3 except that the CV is set to 20%; *i.e.*, we are dealing with the case of larger data variation.

Figure 7 illustrates the situation of using a significance level of 0.05.

Figure 8 shows the behavior of the PR curves by varying the δ values for a fixed sample size of 10 when $k=3$, $\alpha=0.1$, $CV=20\%$.

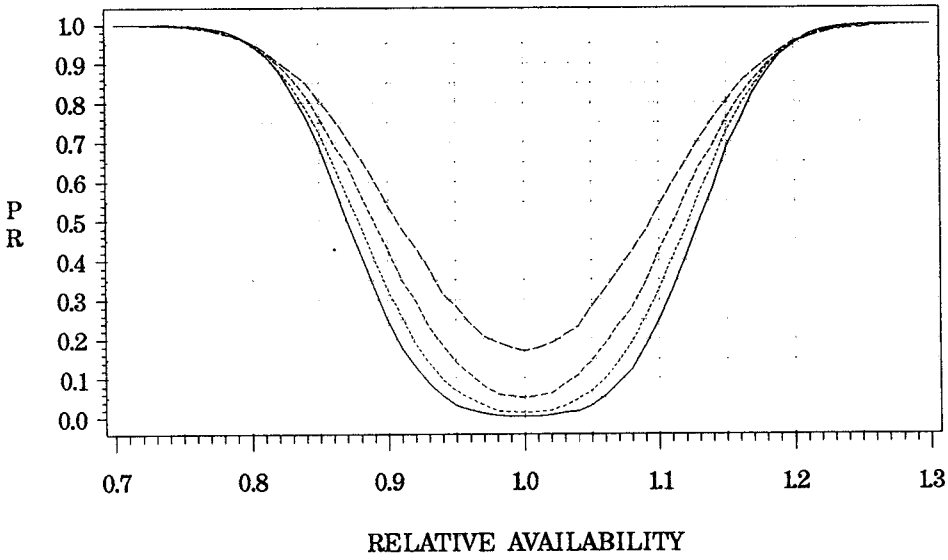

Figure 1. $k=1$, $\delta=0.2$, $\alpha=0.1$, $CV=10\%$; $n=4, 5, 6, 7$

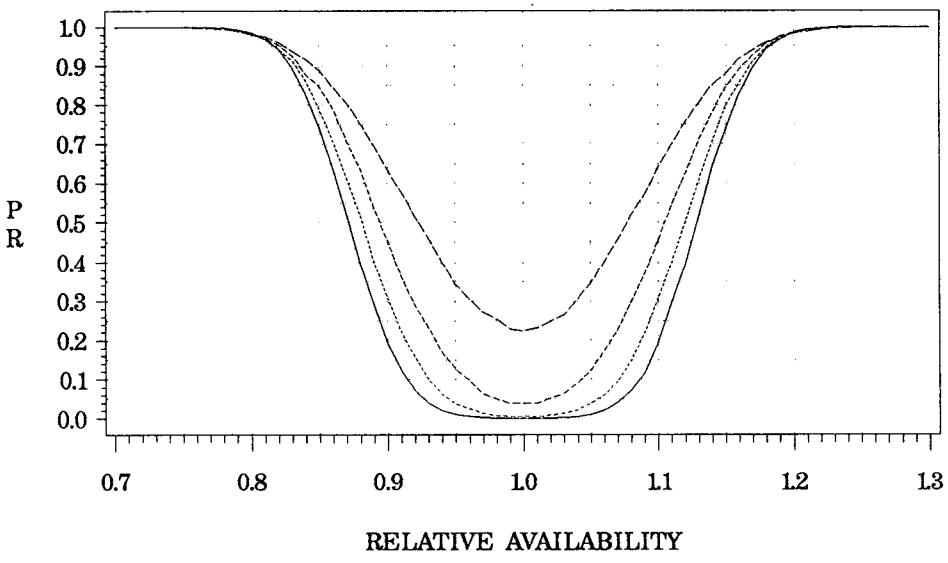

Figure 2. $k=2$, $\delta=0.2$, $\alpha=0.1$, $CV=10\%$; $n=3, 4, 5, 6$

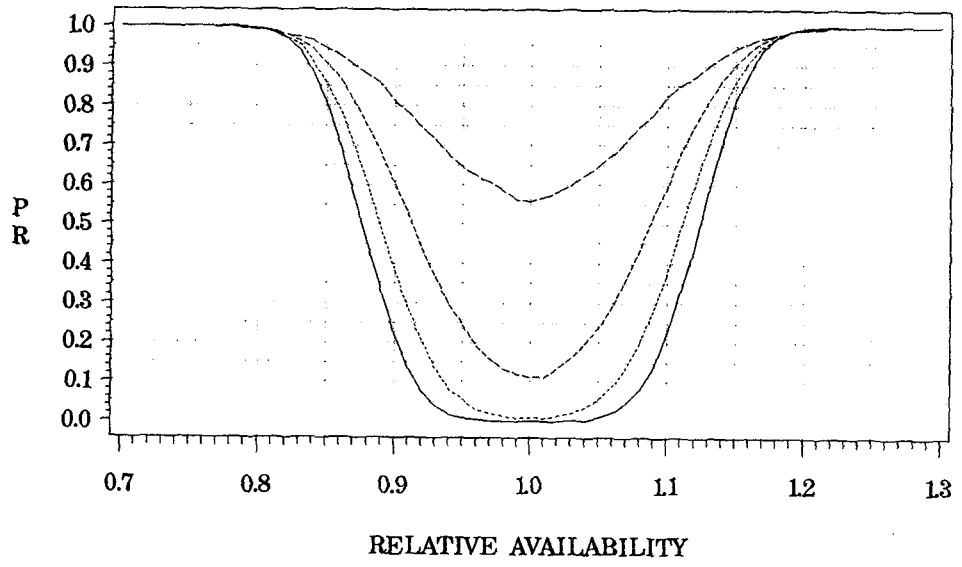

Figure 3. $k=3$, $\delta=0.2$, $\alpha=0.1$, $CV=10\%$; $n=2, 3, 4, 5$

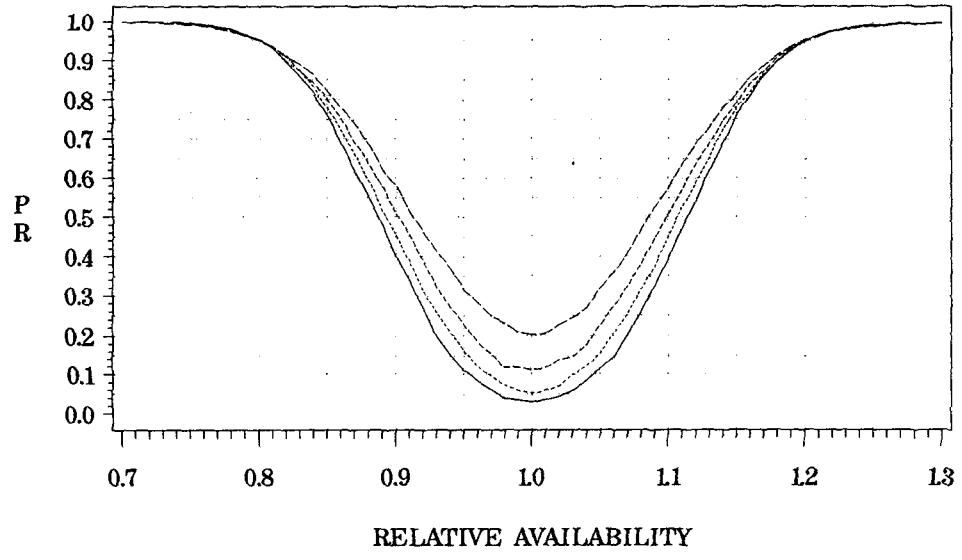

Figure 4. $k=1$, $\delta=0.2$, $\alpha=0.1$, $CV=20\%$; $n=10, 12, 14, 16$

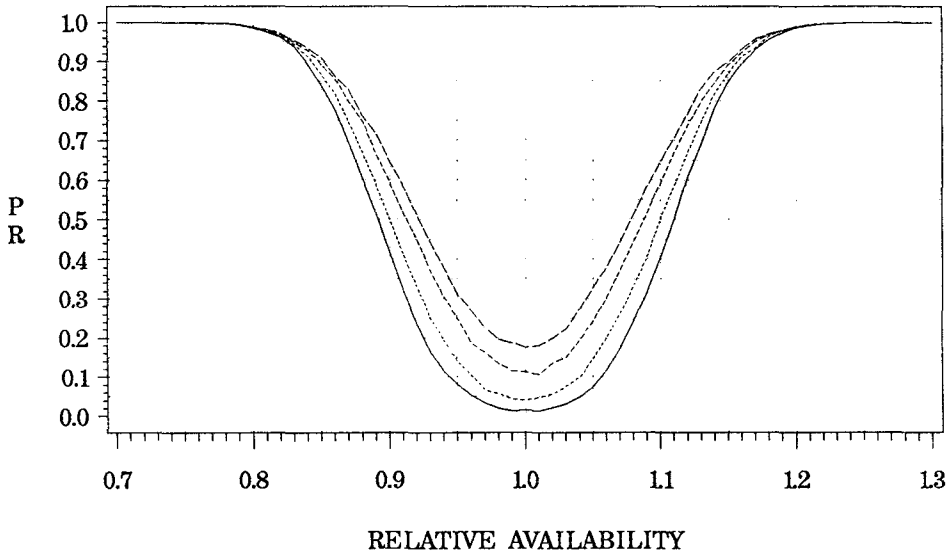

Figure 5. $k=2$, $\delta=0.2$, $\alpha=0.1$, $CV=20\%$; $n=9, 10, 12, 14$

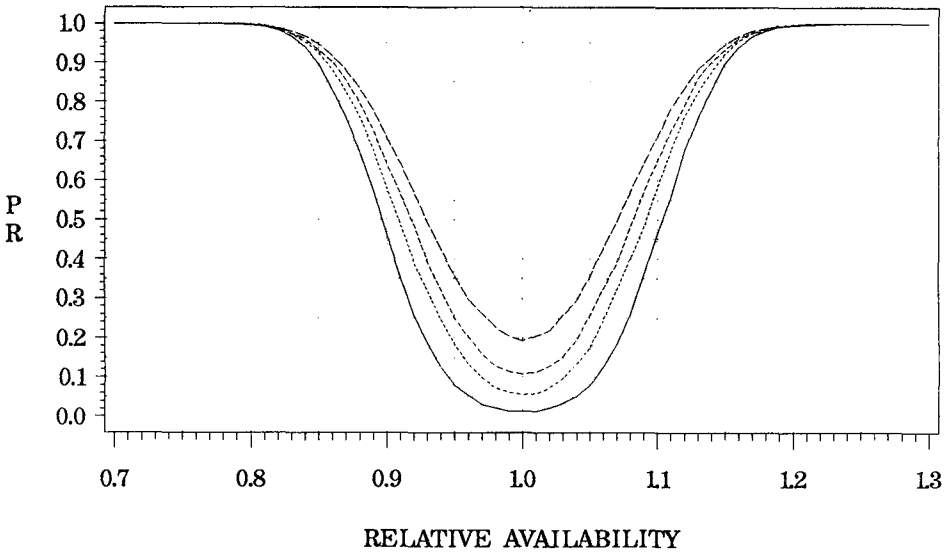

Figure 6. $k=3$, $\delta=0.2$, $\alpha=0.1$, $CV=20\%$; $n=8, 9, 10, 12$

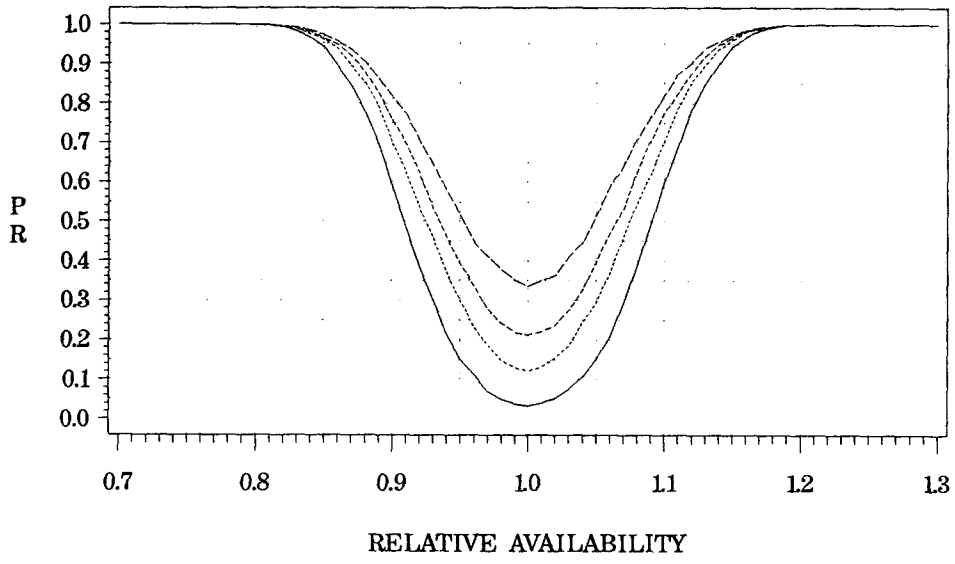

Figure 7. $k=1$, $\delta=0.2$, $\alpha=0.05$, $CV=20\%$; $n=8, 9, 10, 12$

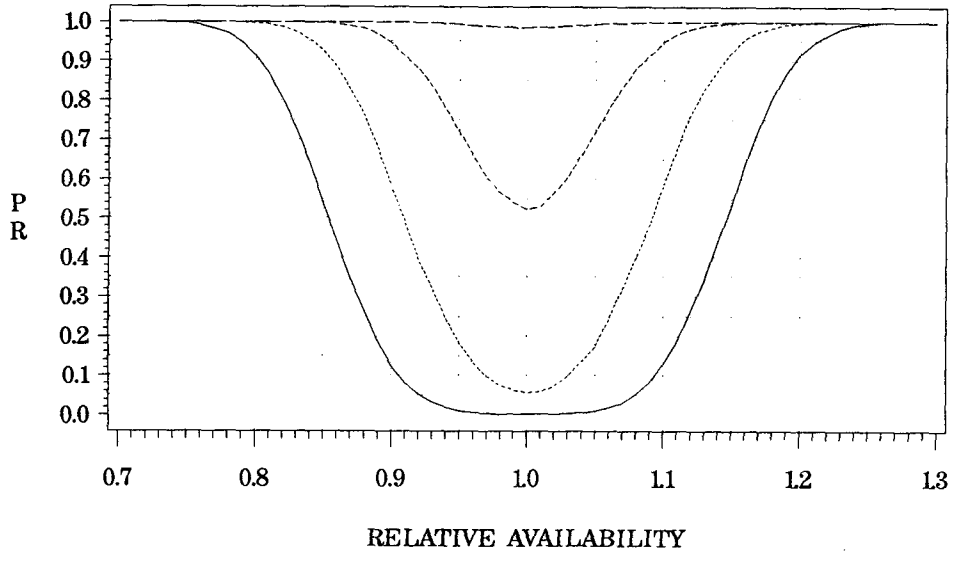

Figure 8. $k=3$, $n=10$, $\alpha=0.1$, $CV=20\%$; $\delta=0.1, 0.15, 0.2, 0.25$

A desirable PR curve, in general, should approach rapidly to zero probability of rejecting equivalence when the true relative availability is around 1, and should approach to probability one when the relative availability is at the boundary or outside the preassigned acceptance region.

From the PR curves one may observe some apparent facts that the PR curves move downward (*i*) as δ increases, (*ii*) to 0 around the relative availability of 1 as the sample size n becomes large, (*iii*) as the data variation represented by the CV becomes small, and (*iv*) as the significance level α increases.

The only FDA guideline concerned about the bioequivalence decision rules requests that upto 20% difference of mean relative availability is allowed with a protection level of 95%, when $k=1$. Figures 1 to 8 meet the requirement satisfactorily. Moreover, the PR curves pass through $(1-\alpha/2k)$ point of probability of rejecting equivalence when the true relative availability is at the boundary. On this phenomina, refer to Huh (1994).

The optimal sample sizes for the equivalence test when $\delta=0.2$ are given in Table 1. Note that $\delta=0.2$ is usually the most practical choice.

Table 1. Optimal sample sizes for equivalence test when $\delta=0.2$

α	k	n		
		CV=10%	CV=20%	CV=30%
0.1	1	5	14	30
	2	4	12	25
	3	3	10	22
	4	3	9	20
	5	3	9	19
0.05	1	6	17	37
	2	5	14	28
	3	4	12	24
	4	4	10	22
	5	3	10	20

4. An Application to Repeated Measures Design

The most common type of experimental design encountered in bioequivalence studies is the crossover design. Note that at here only the simplest crossover designs such as the three-period crossover design are considered.

The large-sample multivariate confidence region based on the usual Hotelling's T^2 statistic is given as follows:

$$n(\bar{y}-\mu)' S^{-1}(\bar{y}-\mu) \leq \chi^2_{k,\alpha} \tag{4.1}$$

Here, S is the sample variance-covariance matrix, $\bar{y}$ the sample mean vector, and μ the population mean vector.

The crossover design belongs to repeated measures designs which require a special covariance structure of sphericity. From the sphericity assumption the population covariance matrix is of the form: $\Sigma_{k\times k} = \sigma^2[(1-\rho)I_k + \rho 11']$, where 1 is the column vector of 1's and ρ is the common correlation coefficient of treatment pairs.

A maximum likelihood estimator of ρ is given as: $\hat{\rho}=(F-1)/[F+(k-1)]$, where F is the usual ratio of the between-subjects mean square to residual mean square (MSE). We take MSE as $\hat{\sigma}^2$, which is denoted by s^2.

The formula (4.1) can be written as

$$(n/s^2)[p\sum(\bar{Y}_i-\mu_i)^2 + q\sum_i\sum_{\neq j}(\bar{Y}_i-\mu_i)(\bar{Y}_j-\mu_j)] \leq \chi^2_{k,\alpha}, \tag{4.2}$$

where $p=(1-2\hat{\rho}+\hat{\rho}k)/[(1-\hat{\rho})(1-\hat{\rho}+\hat{\rho}k)]$, $q=-\hat{\rho}/[(1-\hat{\rho})(1-\hat{\rho}+\hat{\rho}k)]$. The left-hand side of (4.2) is a confidence ellipsoid with axis length $\sqrt{\lambda_i\chi^2_{k,\alpha}/n}$, where λ_i's are the eigenvalues of S.

Hence a conservative test rule for equivalence for repeated measures design is to accept the claim if

$$\sqrt{\sum(\bar{Y}_i-\mu_c)^2/k} + \sqrt{\max|\lambda_i|}\sqrt{\chi^2_{k,\alpha}/n} \leq \delta.$$

Acknowledgements

The author wishes to thank Professor Huh, Department of Statistics, Korea university, for kindly sending his early manuscript on equivalence test for review, which leads to this paper.

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, 2663-2692.

[2] Huh, M. H. (1994). Equivalence testing as an alternative to significance testing, *Journal of the Korean Statistical Society*, Vol. 23, 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, 2-11.

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

[5] Metzler, C. M. (1974). Bioavailability - a problem in equivalence, *Biometrics*, Vol. 30, 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. 267-285.

[7] Metzler, C. M. (1991). Sample sizes for bioequivalence studies, *Statistics in Medicine*, Vol. 10, 961-970.

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

[9] Mustard, D. (1964). Numerical integration over the n-dimensional spherical shell, *Mathematical Computation*, Vol 18, 578-589.

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

[11] Sung, N. K. (1994). Probability of rejection curve for equivalence testing procedure. *Journal of Korean Society for Quality Management*, Vol. 22, 102-110.

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

[13] Westlake, W. J. (1976). Symmetric confidence intervals for bioequivalence trials. *Biometrics*. Vol. 32. 741-744.