Journal of Pharmacokinetics and Biopharmaceutics, Vol. 18, No. 2, 1990

PHARMACOMETRICS

Power of the Two One-Sided Tests Procedure in Bioequivalence

Kem F. Phillips[1]

Received May 16, 1989—Final November 21, 1989

The power of the two one-sided tests procedure for testing bioequivalence is derived from the bivariate noncentral t *distribution. Power curves are shown and their use in planning bioequivalence experiments discussed. Sample sizes computed in the usual manner from an analysis of variance are shown to be too small to assure a declaration of bioequivalence except under favorable conditions.*

KEY WORDS: bioequivalence; hypothesis testing; power; sample size.

INTRODUCTION

The two one-sided tests procedure has become the standard test of bioequivalence (1). Schuirmann (2), who proposed the test in its present form, has demonstrated that the test's logical and operational characteristics are superior to those of other tests of average bioavilability such as the "power approach" and Hauck and Amderson's procedure (3).

In the two one-sided tests procedure, bioequivalence is taken to be the *alternative* hypothesis: let μ_T be the mean bioavailability of the test product, μ_R the mean for the reference product, and θ_L and θ_U the lower and upper bounds defining bioequivalence. Then the null hypothesis, nonequivalence, is

$$H_0: \mu_T - \mu_R < \theta_L \quad \text{or} \quad \mu_T - \mu_R > \theta_U \tag{1}$$

and the alternative hypothesis, bioequivalence, is

$$H_1: \theta_L \leq \mu_T - \mu_R \leq \theta_U \tag{2}$$

[1]Department of Clinical and Scientific Affairs, Pfizer Pharmaceuticals, 235 East 42nd Street, New York, New York 10017.

0090-466X/90/0400-0137$06.00/0

The test is conducted by estimating $\mu_T - \mu_R$, the mean difference between the responses to the two treatments, and the standard deviation of this estimate. In a balanced study $\mu_T - \mu_R$ is estimated by the observed difference, $\bar{X}_T - \bar{X}_R$, which has a normal distribution, $N(\mu_T - \mu_R, 2\sigma^2/n)$, where n is the sample size and σ^2 is the error variance of the observations. Let $\hat{\sigma}^2$ be the mean-square error from the appropriate analysis of variance. Two t statistics are defined as

$$T_L = \frac{\bar{X}_T - \bar{X}_R - \theta_L}{\hat{\sigma}\sqrt{2/n}} \quad \text{and} \quad T_U = \frac{\bar{X}_T - \bar{X}_R - \theta_U}{\hat{\sigma}\sqrt{2/n}} \tag{3}$$

H_0 is rejected in favor of bioequivalence if T_L and $-T_U$ equal or exceed $t_{1-\alpha,\nu}$, where $\hat{\sigma}^2$ has ν degrees of freedom. The nominal level of the test is α, say 0.05, and $t_{1-\alpha,\nu}$ is the value at which the t distribution with ν degrees of freedom reaches $1-\alpha$. Referring to p. 660 of Schuirmann's article, T_L corresponds to t_1, T_U to $-t_2$, θ_L to θ_1 and θ_U to θ_2.

POWER OF THE TWO ONE-SIDED TESTS PROCEDURE

The power of a statistical test is the probability that the null hypothesis, H_0, will be rejected when the alternative hypothesis, H_1, is true. Since in the two one-sided tests procedure bioequivalence is the alternative hypothesis, the power of the test is the probability of accepting bioequivalence when the products are in fact bioequivalent, that is, when the true difference in their means lies in the interval $[\theta_L, \theta_U]$. Bioequivalence is accepted if T_L and $-T_U$ are at least $t_{1-\alpha,\nu}$, so that if $p(\cdot)$ denotes the power for given alternatives,

$$p(\mu_T - \mu_R, \sigma, \nu) = P\{T_L \geq t_{1-\alpha,\nu} \quad \text{and} \quad T_U \leq -t_{1-\alpha,\nu} | \mu_T - \mu_R, \sigma, \nu\} \tag{4}$$

Power can be calculated for each value in the interval, producing a power curve. The distributions of T_L and T_U also depend on the sample size, n, or degrees of freedom ν, and σ, giving rise to entire families of power curves.

Schuirmann's article (2, pp. 670, 671) contains examples of power curves; probabilities for those curves were calculated using numerical integration. Owen (5) has shown that the vector (T_L, T_U) has the bivariate noncentral t distribution, with ν degrees of freedom, and noncentrality parameters δ_L and δ_U, with

$$\delta_L = \frac{\mu_T - \mu_R - \theta_L}{\sigma\sqrt{2/n}} \quad \text{and} \quad \delta_U = \frac{\mu_T - \mu_R - \theta_U}{\sigma\sqrt{2/n}}. \tag{5}$$

By Eq. (4), this distribution underlies the power function for the two one-sided tests procedure. In addition to providing explicit formulas for

computing power, the bivariate noncentral t distribution connects the two one-sided tests procedure with other statistical methodologies, such as sampling plans (4), which motivated Owen's work on the bivariate t distribution. Although Owens's formulas are complicated and involve a definite integral that is not available in standard computing packages, the formulas can be programmed relatively easily in a language such as SAS (6).

POWER CURVES

Power curves are presented in Figs. 1-4. It was assumed in the calculations that $\nu = n - 2$, which holds for a two-period crossover experiment, allowing for period and sequence effects. Each figure contains curves for sample sizes of 9, 12, 18, 24, 30, 40, and 60, and is based on a value of σ ranging from 10 to 30%, σ being expressed as a percentage of a reference mean, that is, as a coefficient of variation; values of the difference in means are expressed as percentages of the same reference mean. The limits defining bioequivalence, θ_L and θ_U, are ±20%, but only the right half of this range is shown since the curves are symmetric around zero.

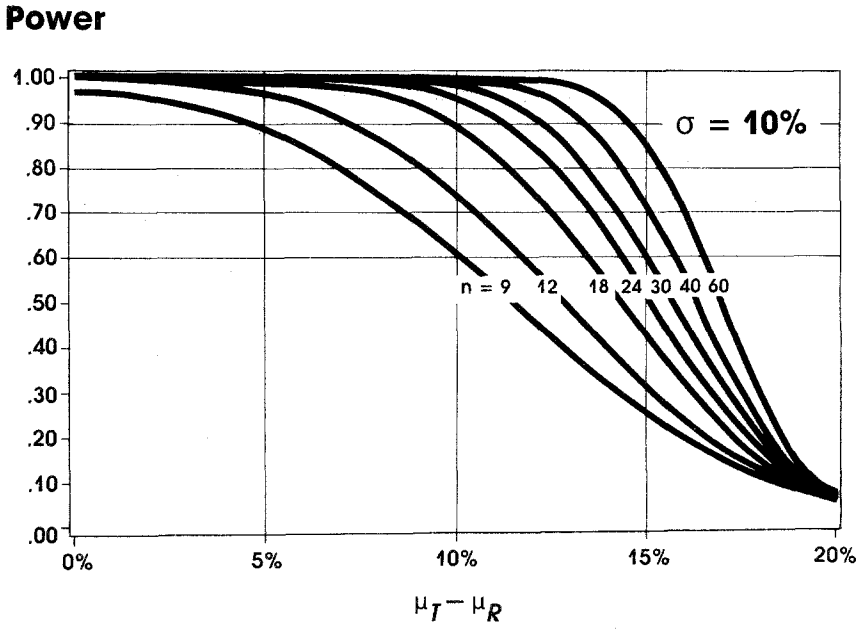

Fig. 1. Power curves for the two one-sided tests procedure with coefficient of variation 10% and samples of $n = 9$, 12, 18, 24, 30, 40, and 60.

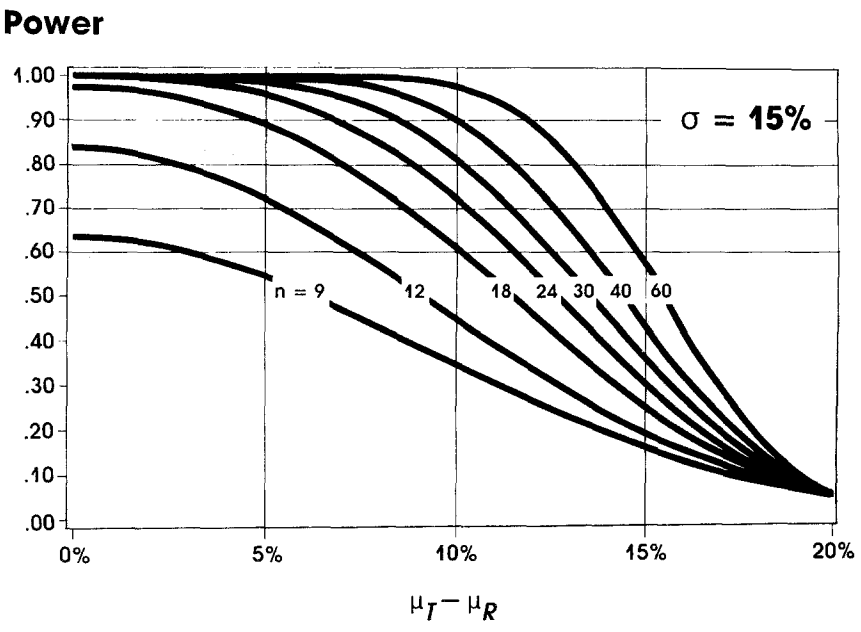

Fig. 2. Power curves for the two one-sided tests procedure with coefficient of variation 15% and samples of $n = 9$, 12, 18, 24, 30, 40, and 60.

Certain characteristics of the power function are apparent from the graphs. Each curve attains maximum power at 0% difference, and falls to the test's true significance level of less than 0.05 at ±20%. Second, as σ increases, the power decreases. Third, as the sample size increases, the power increases for every value of the difference in means.

It may be disturbing to experimenters that the probability of rejecting bioequivalence can be large when the observed mean difference between two treatments is less than 20%. This happens when σ is large, the true difference is large (say 15–20%) or the sample is small; that is, bioequivalence may be rejected when there is insufficient information in the data to distinguish between equivalence and nonequivalence.

DETERMINING SAMPLE SIZE

The sample size for a bioequivalence study is usually determined by demanding that the analysis of variance F test for treatment have 80% power for detecting a 20% difference in treatment means, calculating the power from a formula such as Westlake's (7). When the two one-sided tests procedure is the test of bioequivalence, the sample size should be determined

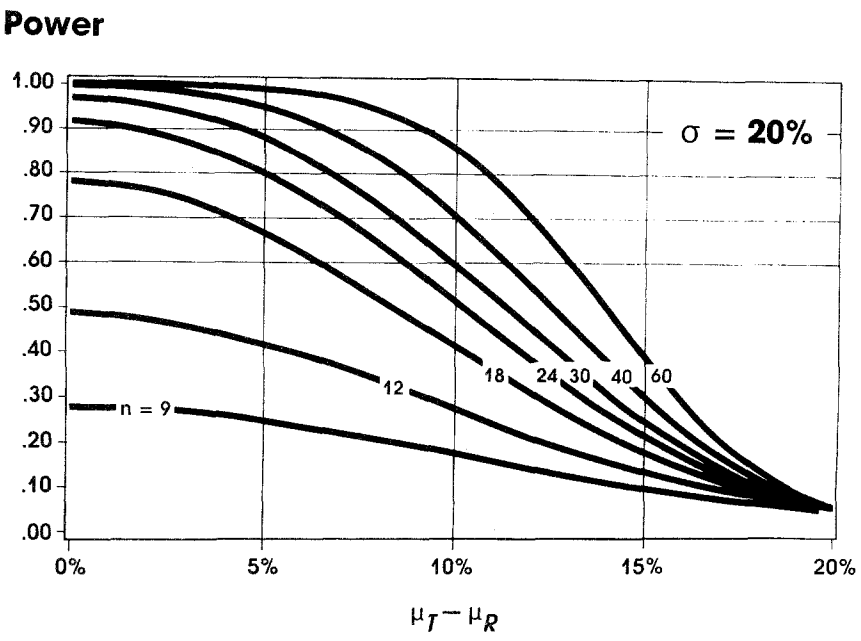

Fig. 3. Power curves for the two one-sided tests procedure with coefficient of variation 20% and samples of $n = 9$, 12, 18, 24, 30, 40, and 60.

from its power function. Ideally, sample size calculations would take into account the costs of experimentation and of failure to demonstrate bioequivalence. A simple approach is to require that the sample be large enough that, based on estimates of $\mu_T - \mu_R$ and σ, the experiment will lead to a declaration of bioequivalence with high probability. By this criterion, if the true mean difference is unlikely to be more than 10%, σ is no more than 15% of the reference mean, and a 20% chance of failing to demonstrate bioequivalence is acceptable, then from Fig. 2, 30 subjects will be enough.

Table I contains sample sizes based on Westlake's formula and the criterion just described. For the two one-sided tests procedure, the sample size is determined by demanding that the probability of demonstrating bioequivalence be a specified value, π, when σ and $\mu_T - \mu_R$ are as shown in the other columns.

In Table I, the samples derived by the usual ANOVA method are considerably smaller than those based on the power of the two one-sided tests procedure, except for the smaller values of σ and $\mu_T - \mu_R$. This means that for larger values of these parameters there is a high probability of not demonstrating bioequivalence, even though the products are bioequivalent.

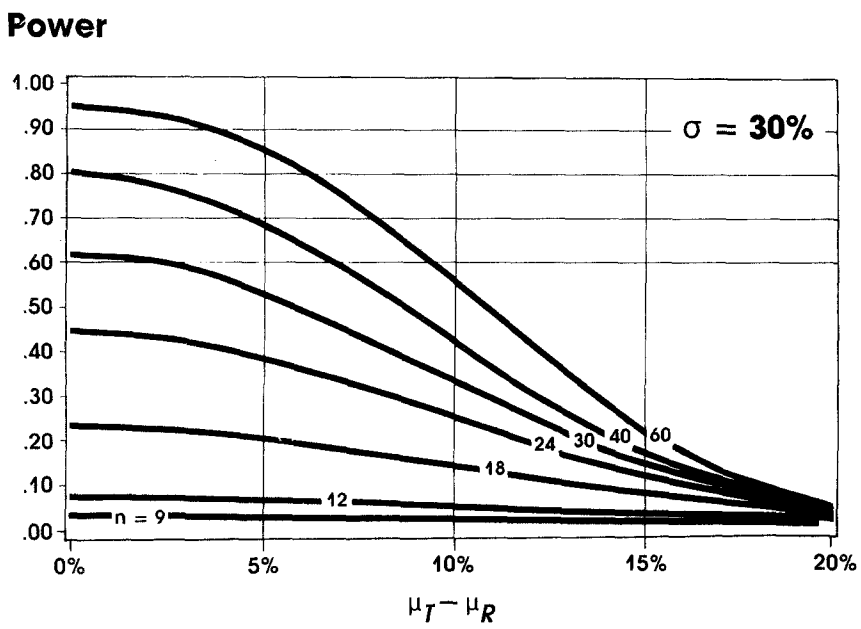

Fig. 4. Power curves for the two one-sided tests procedure with coefficient of variation 30% and samples of $n = 9$, 12, 18, 24, 30, 40, and 60.

Table I. Comparison of Sample Sizes Derived From Analysis of Variance and From the Power Function of the Two One-Sided Tests Procedure

Sample size (ANOVA)			Sample size (Two One-Sided Tests Procedure)			
σ (%)	π (%)		$\mu_T - \mu_R$ 0%	5%	10%	15%
10	70	7	6	7	12	40
20		18	16	20	40	152
30		38	34	42	87	341
10	80	7	7	8	14	51
20		18	19	24	51	200
30		38	40	52	113	447
10	90	7	8	10	19	70
20		18	24	33	70	276
30		38	51	71	156	618

For the largest values of σ and $\mu_T - \mu_R$, the required samples are so large that bioequivalence experiments of those sizes would never be undertaken.

It could be argued that the comparison of the two methods is "unfair," since the newer method allows adjustment of two additional parameters, and that the sample size for the ANOVA method can be increased by demanding higher power at 20% or 80% power at a smaller difference. But Schuirmann's (2) discussion makes clear that because of the nonsensical shape of the rejection region for the power rule, strengthening the power requirement may well increase the probability of *rejecting* bioequivalence. The superiority of the two one-sided tests procedure stems from its logical coherence rather than its parametrization.

DISCUSSION

The power rule (2) required that for a declaration of bioequivalence an analysis of variance show no difference in treatment means, and that the power of detecting a 20% difference be at least 80%. The best strategy for demonstrating bioequivalence with the power rule was to use the minimum sample size that would produce 80% power; a larger sample increased the risk of detecting an unimportant difference that would surely exist. With the two one-sided tests procedure, power, the a priori probability of demonstrating the bioequivalence of bioequivalent products, always increases with sample size. Power is used only in planning the experiment, not as part of the statistical test. An experimenter with accurate estimates of the mean and variance of the difference in treatment means can choose a sample size that reduces the likelihood of rejecting bioequivalence to near zero. In practice, the statistical parameters are not known with certainty, since then experimentation would be irrelevant, so sample size calculations must be based on estimates.

In applying a formal test of bioequivalence, regulators have no interest in power, but in evaluating drug development as a whole, their concern with rational experimentation and safety may lead to participation in sample size determination. The power function of the two one-sided tests procedure provides a basis for the quantification of the risks in bioequivalence testing.

REFERENCES

1. Bioequivalence Task Force. Report on recommendations from the bioequivalence hearing conducted by the Food and Drug Administration, September 29-October 1, 1986.
2. D. J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokin. Biopharm.* **15**:657-680 (1987).
3. W. W. Hauck and S. Anderson. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *J. Pharmacokin. Biopharm.* **12**:83-91 (1984).

4. R. E. Odeh and D. B. Owen. *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*, Marcel Dekker, New York, 1980.
5. D. B. Owen, A special case of a noncentral *t*-distribution. *Biometrika* **52:**437-446 (1965).
6. SAS. *SAS User's Guide: Basics*, Version 5, SAS Institute, Cary, NC, 1985.
7. W. J. Westlake. The design and analysis of comparative blood-level trials. In J. Swarbrick (ed.), *Current Concepts in the Pharmaceutical Sciences, Dosage Form Design and Bioavailability*, Lea & Febiger, Philadelphia, 1973, pp. 149-179.