Journal of Pharmacokinetics and Biopharmaceutics, Vol. 15, No. 6, 1987

PHARMACOMETRICS

A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability

Donald J. Schuirmann[1]

Received March 2, 1987—Final September 22, 1987

The statistical test of the hypothesis of no difference between the average bioavailabilities of two drug formulations, usually supplemented by an assessment of what the power of the statistical test would have been if the true averages had been inequivalent, continues to be used in the statistical analysis of bioavailability/bioequivalence studies. In the present article, this Power Approach (which in practice usually consists of testing the hypothesis of no difference at level 0.05 and requiring an estimated power of 0.80) is compared to another statistical approach, the Two One-Sided Tests Procedure, which leads to the same conclusion as the approach proposed by Westlake (2) based on the usual (shortest) $1-2\alpha$ confidence interval for the true average difference. It is found that for the specific choice of $\alpha = 0.05$ as the nominal level of the one-sided tests, the two one-sided tests procedure has uniformly superior properties to the power approach in most cases. The only cases where the power approach has superior properties when the true averages are equivalent correspond to cases where the chance of concluding equivalence with the power approach when the true averages are not *equivalent exceeds 0.05. With appropriate choice of the nominal level of significance of the one-sided tests, the two one-sided tests procedure always has uniformly superior properties to the power approach. The two one-sided tests procedure is compared to the procedure proposed by Hauck and Anderson (1).*

KEY WORDS: bioavailability; bioequivalence; hypothesis testing; interval hypotheses.

INTRODUCTION

The statistical issue associated with the analysis of bioavailability/bioequivalence studies that has received the most attention in the pharmaceutical and statistical literature is the question of statistical methods for determining whether two formulations of a drug have been shown to be equivalent with respect to average bioavailability in the population. "Bioavailability," in this context, is to be characterized by one or more

[1] Division of Biometrics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, 5600 Fishers Lane, Rockville, Maryland 20857.

blood concentration profile variables, such as area under the blood concentration-time curve (*AUC*), maximum concentration ($C_{\max}$), etc., and possibly by urinary excretion variables as well.

Hauck and Anderson (1), in an article in which they proposed a new approach to this problem, gave a clear explanation of why the null hypothesis of *no difference* between the two averages, as tested by the "treatments" *F* test from the analysis of variance of a two-treatment (formulation) study, is the wrong statistical hypothesis for assessing the evidence in favor of a conclusion of equivalence. And yet, as Hauck and Anderson note, the test of the hypothesis of no difference is still utilized by many who seek to demonstrate equivalence of two formulations. In most cases those who utilize the test of the hypothesis of no difference supplement it with some assessment of what the power of the test would have been if the averages had been different enough to be considered inequivalent. This Power Approach, as it will be called, has been a standard method in bioequivalence testing, in spite of the fact that it is based on the test of an inappropriate statistical hypothesis.

This article compares this power approach to another method for assessing the equivalence of two formulations which will be called the *Two One-Sided Tests Procedure.* Then the two one-sided tests procedure is compared to the proposed method of Hauck and Anderson.

STATEMENT OF THE PROBLEM

Suppose we have a bioavailability/bioequivalence study in which a test product T and a reference product R are administered. The reference product could be an innovator's product and the test product a potential generic substitute manufactured by a different firm. Alternatively, the test and reference products could be, for example, two different dosage forms of a drug product, both manufactured by the same firm. Let μ_T be the average bioavailability of the test product and μ_R the average bioavailability of the reference product. For purposes of this discussion, we assume that μ_T and μ_R are in fact the *mean* bioavailabilities.

As noted by Hauck and Anderson, the objectives of the statistical analysis may be incorporated into the following statistical hypotheses:

$$\begin{aligned} &H_0: \quad \mu_T - \mu_R \leq \theta_1 \quad \text{or} \quad \mu_T - \mu_R \geq \theta_2 \\ &H_1: \quad \theta_1 < \mu_T - \mu_R < \theta_2 \end{aligned}$$

The structure of these statistical hypotheses is determined by the objective of the analysis. The null hypothesis, H_0, states that μ_T and μ_R are *not* equivalent. The alternative hypothesis, H_1, states that they *are* equivalent. If, on the basis of the results from the study, we may reject H_0, then we may conclude that H_1 is true, i.e., we may conclude that μ_T and μ_R are equivalent. If we do not reject H_0, we do not conclude that H_0 is true. Rather, we say that it has not been shown that H_1 is true. Further studies on the same products could conceivably establish that μ_T and μ_R are equivalent, even though the study in hand does not.

The statistical hypotheses H_0 and H_1 given above will be referred to as the "interval hypotheses." Methods for testing these hypotheses will be called methods for "testing the interval hypothesis H_0." The interval $[\theta_1, \theta_2]$ may be called the "equivalence interval." The limits θ_1 and θ_2 ($\theta_1 < \theta_2$) may be stated as known numbers, expressed in the same units as the bioavailability variable of interest. In other cases, θ_1 and θ_2 may be defined as proportions of the unknown reference mean μ_R. The specification of θ_1 and θ_2 is made by the experts in the fields of biopharmaceutics and medicine (not by the statistician!). For purposes of this discussion, it is assumed that θ_1 and θ_2 are known numbers. A brief discussion of the case where θ_1 and θ_2 are stated as proportions of μ_R appears in Appendix A.

The interval hypotheses presented above do not represent a standard problem in statistical methods. Discussion of the interval hypotheses has not been generally included in the one- or two-semester statistical methods courses that graduate students in the sciences are usually required to take. As noted above, the problem of testing the interval hypothesis H_0 has received attention in the statistical community for many years, but the proposed solutions have not found their way into the standard statistical methods texts.

In comparing the power approach to the two one-sided tests procedure, it is assumed that the data arise from a normal distribution. We also assume that the within-subject (intrasubject) variances of the test and reference products are the same, although this is not a critical assumption. Furthermore, we assume that the study is a *balanced* crossover study. By a balanced study we mean that there is an equal number of subjects in each treatment-administration sequence and there are no missing observations from any subject. The conclusions drawn here concerning the two procedures remain valid if the study is not balanced, but there are certain technical complications involved in the analysis of unbalanced studies. The balance assumption is thus made for simplicity. A brief discussion of unbalanced studies appears in Appendix B. Finally, we also make the assumption that $\theta_1 = -\theta_2$, i.e., that the equivalence interval is symmetric about zero. This assumption is not needed for the assessment of the two one-sided tests procedure, but it is needed for the power approach.

$\bar{X}_T - \bar{X}_R$ is the difference between the observed average bioavailabilities of products T and R, respectively. The precision of $\bar{X}_T - \bar{X}_R$ as an estimator of $\mu_T - \mu_R$ is measured by its standard deviation, which for a balanced study is $\sigma\sqrt{2/n}$, where n is the total number of subjects in the study and σ is the intrasubject (i.e., within-subject) standard deviation of the observations. Since σ is unknown, we estimate it with s, the square root of the "error" mean square from the crossover design analysis of variance. The resulting quantity, $s\sqrt{2/n}$, is called the standard error of $\bar{X}_T-\bar{X}_R$, based on ν degrees of freedom (the number of degrees of freedom associated with the "error" mean square), and is our estimate of the precision with which $\bar{X}_T-\bar{X}_R$ estimates $\mu_T-\mu_R$.

In terms of the data from the bioequivalence study, both of the procedures described depend only on the estimate $\bar{X}_T-\bar{X}_R$, its standard error $s\sqrt{2/n}$, and the degrees of freedom ν. In order to compare these procedures, it will be interesting to examine which pairs of values of $\bar{X}_T-\bar{X}_R$ and $s\sqrt{2/n}$ lead to rejection of the interval hypothesis H_0, and thus to a conclusion of equivalence of μ_T and μ_R.

THE TWO ONE-SIDED TESTS PROCEDURE

The Two One-Sided Tests Procedure, as its name implies, consists of decomposing the interval hypotheses H_0 and H_1 into two sets of one-sided hypotheses

$$H_{01}\colon\quad \mu_T-\mu_R\le\theta_1$$

$$H_{11}\colon\quad \mu_T-\mu_R>\theta_1$$

and

$$H_{02}\colon\quad \mu_T-\mu_R\ge\theta_2$$

$$H_{12}\colon\quad \mu_T-\mu_R<\theta_2$$

The two one-sided tests procedure consists of rejecting the interval hypothesis H_0, and thus concluding equivalence of μ_T and μ_R, if and only if both H_{01} *and* H_{02} are rejected at a chosen *nominal level of significance α.* The logic underlying the two one-sided tests procedure is that if one may conclude that $\theta_1<\mu_T-\mu_R$, and may also conclude that $\mu_T-\mu_R<\theta_2$, then it has in effect been concluded that $\theta_1<\mu_T-\mu_R<\theta_2$.

Under the normality assumption that has been made, the two sets of one-sided hypotheses will be tested with ordinary one-sided t tests. Thus it will be concluded that μ_T and μ_R are equivalent (for a balanced study) if

$$t_1=\frac{(\bar{X}_T-\bar{X}_R)-\theta_1}{s\sqrt{2/n}}\ge t_{1-\alpha(\nu)} \quad\text{and}\quad t_2=\frac{\theta_2-(\bar{X}_T-\bar{X}_R)}{s\sqrt{2/n}}\ge t_{1-\alpha(\nu)}$$

where, once again, s is the square root of the "error" mean square from the crossover design analysis of variance. $t_{1-\alpha(\nu)}$ is the point that isolates probability α in the upper tail of the Student's t distribution with ν degrees of freedom, where ν is the number of degrees of freedom associated with the "error" mean square.

The two one-sided tests procedure turns out to be operationally identical to the procedure of declaring equivalence only if the ordinary $1-2\alpha$ (not $1-\alpha$) confidence interval for $\mu_T-\mu_R$ is completely contained in the equivalence interval $[\theta_1, \theta_2]$. For this reason, it is sometimes referred to as the confidence interval approach. In this form, it has been recommended by Westlake (2).

THE POWER APPROACH

The Power Approach is an ad hoc method of testing the interval hypothesis H_0 which has been a standard method until recently. The power approach consists of testing the Hypothesis of No Difference

$$H_0'\colon\quad \mu_T-\mu_R=0$$

$$H_1'\colon\quad \mu_T-\mu_R\neq 0$$

at the 0.05 level of significance, using a standard two-sided t test. If the bioavailability/bioequivalence study is a two-treatment study, then this t test corresponds to the "treatments" F test in the crossover design analysis of variance (if there are more than two products in the study, the t test and the treatments F test are *not* the same).

Under the Power Approach, if the hypothesis of no difference H_0' is rejected, then the interval hypothesis H_0 is not rejected, i.e., one does not conclude that μ_T and μ_R are equivalent. If the hypothesis of no difference H_0' is not rejected, then the question arises of what the Power of the test of H_0' would have been if $\mu_T-\mu_R$ had in fact been equal to θ_2, i.e., if $\mu_T-\mu_R$ had been large enough for the means to be considered inequivalent. This power depends on the true value of σ. Arbitrary convention has dictated that this power should be at least 0.80 before failure to reject the hypothesis of no difference may be taken as evidence that μ_T and μ_R are equivalent. There exists a value $\sigma_{0.80}$ such that if $\sigma \leq \sigma_{0.80}$, then the power of the test of no difference (H_0') to detect $\mu_T-\mu_R=\theta_2$ (or $-\theta_2$) is greater than or equal to 0.80. Unfortunately, since σ is unknown, it cannot be compared to $\sigma_{0.80}$, nor can the actual power be calculated. The best that can be managed is to estimate the unknown σ by s. The power approach then consists of rejecting the interval hypothesis H_0, and thus concluding that μ_T and μ_R

are equivalent, if

$$-t_{0.975(\nu)} \leq \frac{\bar{X}_T - \bar{X}_R}{s\sqrt{2/n}} \leq t_{0.975(\nu)} \quad \text{and} \quad s \leq \sigma_{0.80}$$

That is, if the hypothesis of no difference H_0' is not rejected and the power of the test to detect $\mu_T - \mu_R = \theta_2$ is estimated to have been at least 0.80. Note that this procedure is based on estimated power, since the true power is a function of the unknown σ.

The logic underlying the power approach is that if $\mu_T - \mu_R$ had actually been as large as θ_2 (or as small as $\theta_1 = -\theta_2$) it is likely (provided the estimate of this likelihood is at least 80%) that the hypothesis of no difference H_0' would have been rejected. Thus if H_0' is not rejected, it is concluded that $|\mu_T - \mu_R|$ was not as large as θ_2, i.e., the interval hypothesis H_0 is rejected.

The two one-sided tests procedure depends on the choice of the nominal level of significance α. In the case of the power approach, it is of course possible to carry out the test of the hypothesis of no difference at a level other than 0.05 and/or to require an estimated power other than 0.80, but this is virtually never done. We consider only the power approach with the test carried out at the 0.05 level and the required power 0.80.

COMPARISON OF THE REJECTION REGIONS FOR THE TWO PROCEDURES

The set of values of $\bar{X}_T - \bar{X}_R$ and $s\sqrt{2/n}$ which lead to rejection of the interval hypothesis H_0, and thus to the conclusion that μ_T and μ_R are equivalent, is called the *rejection region* for the procedure. Figure 1 presents the values of $\bar{X}_T - \bar{X}_R$ and $s\sqrt{2/n}$ leading to rejection of the interval hypothesis H_0 using the power approach, for the specific example of $\theta_2 = 20$ units and the degrees of freedom $\nu = 10$. Any values of $\bar{X}_T - \bar{X}_R$ and $s\sqrt{2/n}$ falling in the illustrated triangle lead to a conclusion that μ_T and μ_R are equivalent. One of the most notable aspects of this figure is its flatness. For $s\sqrt{2/n} > 6.44$, no value of $\bar{X}_T - \bar{X}_R$ leads to rejection of the interval hypothesis H_0, not even $\bar{X}_T - \bar{X}_R = 0$. This corresponds to the estimated power being less than 0.80. For $s\sqrt{2/n} \leq 6.44$, the figure has a very interesting shape, in that as the precision (as estimated by $s\sqrt{2/n}$) of $\bar{X}_T - \bar{X}_R$ as an estimate of $\mu_T - \mu_R$ improves, $\bar{X}_T - \bar{X}_R$ must be *closer* to zero, until in the limit as $s\sqrt{2/n}$ approaches zero, $\bar{X}_T - \bar{X}_R$ must be zero if we are to conclude $-20 \leq \mu_T - \mu_R \leq 20$. So if we were to observe $\bar{X}_T - \bar{X}_R = 10$ and $s\sqrt{2/n} = 6$, we would conclude $-20 \leq \mu_T - \mu_R \leq 20$ using the power approach, but if we were to observe $\bar{X}_T - \bar{X}_R = 10$ and $s\sqrt{2/n} = 2$, we would *not* conclude $-20 \leq \mu_T - \mu_R \leq 20$! It should be apparent that this is an illogical property for the procedure to have. Surely if the estimate is close enough to zero,

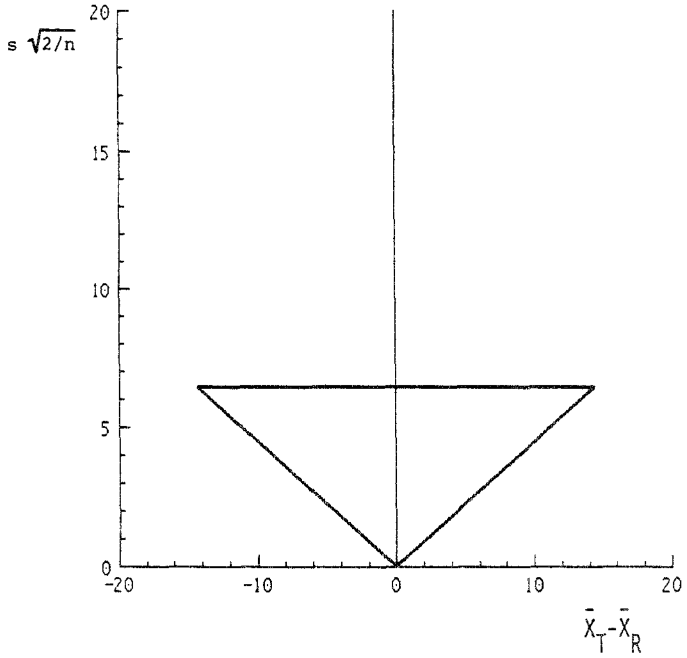

Fig. 1. Rejection region for the power approach, for the example of $\theta_2 = -\theta_1 = 20$ units and $\nu = 10$. Any study for which the pair $(\bar{X}_T - \bar{X}_R, s\sqrt{2/n})$ falls in the illustrated triangle results in the conclusion that μ_T and μ_R are equivalent, using the power approach.

for a given level of precision, for us to consider μ_T and μ_R equivalent, then the same estimate with better precision should also enable us to declare equivalence. So this lack of what might be called convexity of the rejection region is a problem with the power approach. (Stated simply, a rejection region is called convex if it does not get wider as $s\sqrt{2/n}$ increases.)

Figure 2 presents the values of $\bar{X}_T - \bar{X}_R$ and $s\sqrt{2/n}$ leading to rejection of the interval hypothesis H_0 using the two one-sided tests procedure, for the example of $\theta_2 = 20$ units, the degrees of freedom $\nu = 10$, and nominal level $\alpha = 0.05$. For values of $s\sqrt{2/n}$ higher than $\theta_2 / t_{0.95(10)} = 11.04$, no value of $\bar{X}_T - \bar{X}_R$ leads to rejection of the interval hypothesis H_0. The interpretation of this is that the estimated precision of $\bar{X}_T - \bar{X}_R$ as an estimate of $\mu_T - \mu_R$ is too poor to make a reliable determination of whether $-20 \leq \mu_T - \mu_R \leq 20$. For $s\sqrt{2/n} \leq 11.04$, the smaller $s\sqrt{2/n}$ is, the wider is the interval of $\bar{X}_T - \bar{X}_R$ values leading us to conclude that $-20 \leq \mu_T - \mu_R \leq 20$. That is, as the precision (estimated by $s\sqrt{2/n}$) of $\bar{X}_T - \bar{X}_R$ as an estimate of $\mu_T - \mu_R$ improves, $\bar{X}_T - \bar{X}_R$ can be farther from zero and we will still conclude $-20 \leq \mu_T - \mu_R \leq 20$. Thus the rejection region for the two one-sided tests

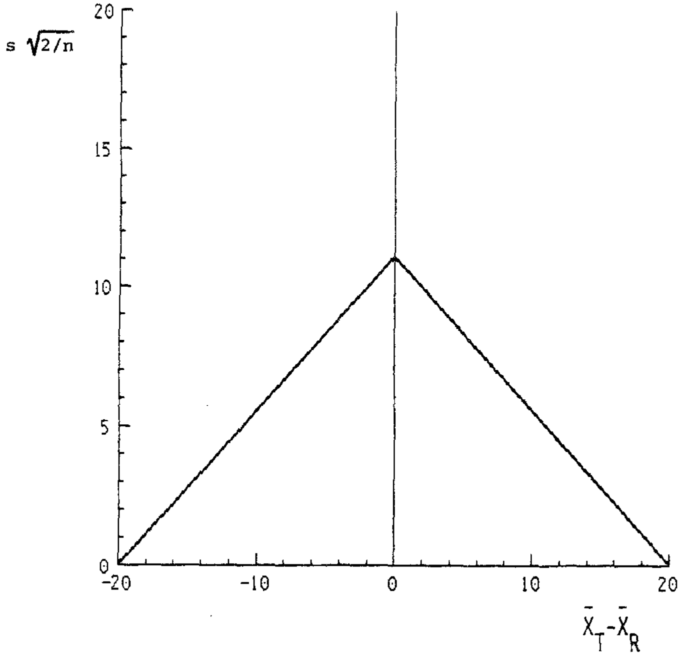

Fig. 2. Rejection region for the two one-sided tests procedure, for the example of $\theta_2 = -\theta_1 = 20$ units, $\nu = 10$, and nominal $\alpha = 0.05$.

procedure has the "convexity" property that was lacking in the case of the power approach. In general the rejection region for the two one-sided tests procedure has the same basic triangular shape as illustrated in Fig. 2, with the peak of the triangle occurring at $s\sqrt{2/n} = \theta_2/t_{1-\alpha(\nu)}$.

Figure 3 presents the rejection regions for the power approach (Fig. 1) and the two one-sided tests procedure with nominal $\alpha = 0.05$ (Fig. 2) on the same graph. It is seen that most values of $(\bar{X}_T - \bar{X}_R, s\sqrt{2/n})$ leading to rejection of the interval hypothesis H_0 using the power approach also lead to rejection using the two one-sided tests procedure, but not all. A small area of values in the upper corners of the power approach rejection region do not lead to rejection using the two one-sided tests procedure, $\alpha = 0.05$. For the rejection region of the two one-sided tests procedure to completely contain the rejection region of the power approach, we would have to do the two one-sided tests at a nominal level of about 0.20.

It should be apparent from examination of Figs. 1-3 that the two one-sided tests procedure is superior to the power approach as a test of the interval hypothesis H_0. The shape of the rejection region of the power approach is simply not that of a sensible test. However, some may not be

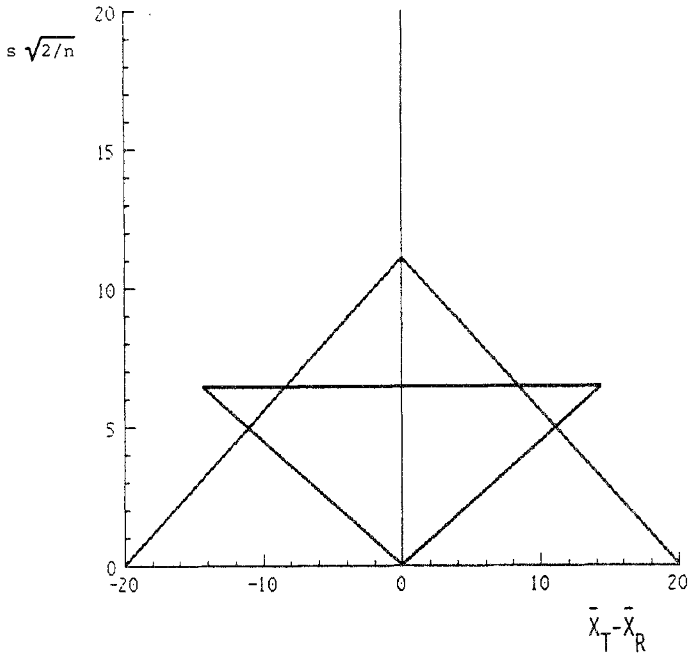

Fig. 3. Comparison of the rejection regions for the power approach (Fig. 1) and the two one-sided tests procedure with nominal $\alpha = 0.05$ (Fig. 2), for the example of $\theta_2 = -\theta_1 = 20$ units and $\nu = 10$.

convinced by this intuitive argument, so we examine the actual probability characteristics of the two procedures.

COMPARISON OF THE PROBABILITY CHARACTERISTICS OF THE TWO PROCEDURES

In order to compare the probability characteristics of the two one-sided tests procedure to the power approach, it is necessary to introduce a measure that indexes the *sensitivity* of a bioequivalence study. This measure, which shall be called ∇, is the ratio of the width of the equivalence interval to the true standard deviation of our estimator of $\mu_T - \mu_R$. For a balanced study, this is given by

$$\nabla = \frac{\theta_2 - \theta_1}{\sigma\sqrt{2/n}} = \frac{2\theta_2}{\sigma\sqrt{2/n}} \qquad \text{if } \theta_1 = -\theta_2$$

If ∇ is high, it means that the precision (as measured by $\sigma\sqrt{2/n}$) of our estimate of $\mu_T - \mu_R$ compares favorably to the width of the equivalence interval, i.e., we have a sensitive study. If ∇ is small, we have an insensitive study. The performance of both of the test procedures depends on ∇.

The probability of rejecting the interval hypothesis H_0 (and thus declaring the average bioavailabilities to be equivalent) when μ_T and μ_R are not in fact equivalent is largest when $\mu_T - \mu_R$ is on the edge of the equivalence interval, i.e., when $\mu_T - \mu_R = \theta_2$ or θ_1 $(= -\theta_2)$. The symmetry of the procedures makes these probabilities equal. The probability of rejecting the interval hypothesis H_0 when $\mu_T - \mu_R = \pm\theta_2$ will be called the true level of significance of the procedure, representing the maximum probability of saying the means are equivalent if in fact they are not equivalent. (See Appendix C: Method of Obtaining the Figures.)

Figure 4 presents the true level of significance of the power approach for degrees of freedom $\nu = 10$, as a function of ∇. The true level of the power approach in this example rises to a peak of 0.060 at about $\nabla = 6.334$, and then falls again, so that for large values of ∇, i.e., for very sensitive studies, the true level is virtually zero. At first it might seem that this is a desirable property, since there would be virtually no chance of saying the means are equivalent if in fact they are not, but this is in fact not desirable. This is because if one does not take as much of a chance as is tolerable of

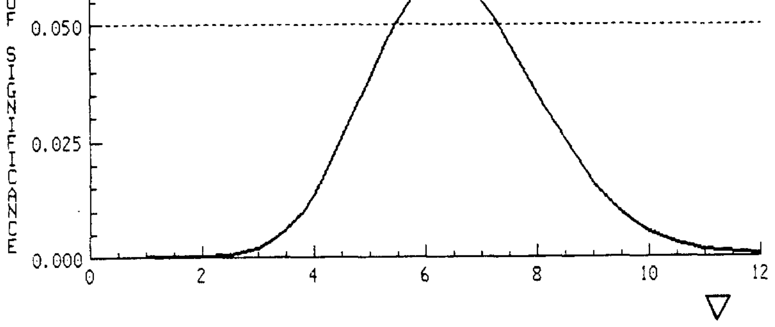

Fig. 4. True level of significance of the power approach, as a function of ∇ (see text), for the example of $\nu = 10$. The true level is the probability of concluding that μ_T and μ_R are equivalent if in fact $\mu_T - \mu_R = \pm\theta_2$, i.e., the maximum probability of *incorrectly* concluding equivalence. The dashed line corresponding to a true level of 0.05 is included for reference.

saying the means are equivalent when they are not, then there will be little or no chance of concluding that the means are equivalent when they are.

A study with 10 degrees of freedom for error would be fairly small. For example, a two-treatment crossover study with 12 subjects. As an example of a larger study, Fig. 5 presents results for the power approach for 40 degrees of freedom. We see that the true level of significance, as a function of ∇, rises to a higher peak of 0.096 at ∇ equal to about 6.214 and then falls sharply as ∇ continues to increase.

In the power approach, the test of the hypothesis of no difference H_0' is carried out at a level of 0.05. Many people are somehow convinced that for this reason, the level of the power approach as a test of the interval hypothesis H_0 is also 0.05. However, we have seen in Figs. 4 and 5 that this is not the case, unless ∇ just happens to equal one of the two values of ∇ for which the true level is 0.05. Other users of the power approach believe that the true level of the test is 0.20, which is just one minus the required power, 0.80. We have seen that this is also not the case (at least not for finite degrees of freedom). The maximum level of significance of the power approach does increase with increasing degrees of freedom. Table I presents the maximum level of the power approach for a range of degrees of freedom. As degrees of freedom continue to increase, the maximum level will approach 0.20.

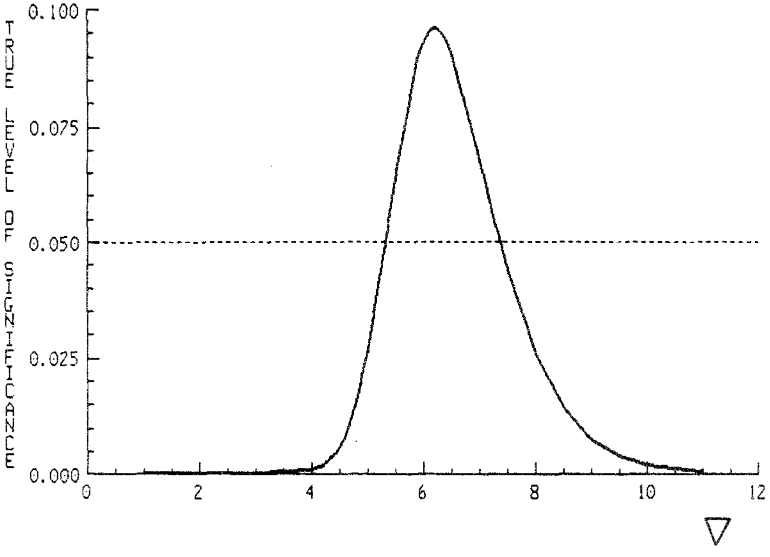

Fig. 5. True level of significance of the power approach, as a function of ∇, for the example of $\nu = 40$.

Table I. Maximum Probability of Concluding That μ_T and μ_R Are Equivalent, If in Fact They Are Not Equivalent, Using the Power Approach, for Several Values of the Degrees of Freedom ν

ν	Maximum probability	ν	Maximum probability
10	0.0605	30	0.0884
16	0.0722	40	0.0958
20	0.0779	50	0.1016
26	0.0847	100	0.1188

The results in Table I establish the maximum probability of saying the means are equivalent, if in fact they are not, that one must be willing to live with to use the power approach. Since the within-subject standard deviation σ, and thus ∇, is unknown, we cannot rule out the possiblity that these maximum levels will be achieved. On the other hand, if our study is more sensitive than we had planned, we will be in the range of large ∇ where the power approach is increasingly conservative, hardly a desirable property of a testing procedure.

In Figs. 4 and 5 we examined the probability of concluding that the means are equivalent when in fact they differ by enough ($\mu_T - \mu_R = \theta_2$) to be considered inequivalent, for the case of the power approach. Figure 6 presents the corresponding probabilities for the two one-sided tests pro-

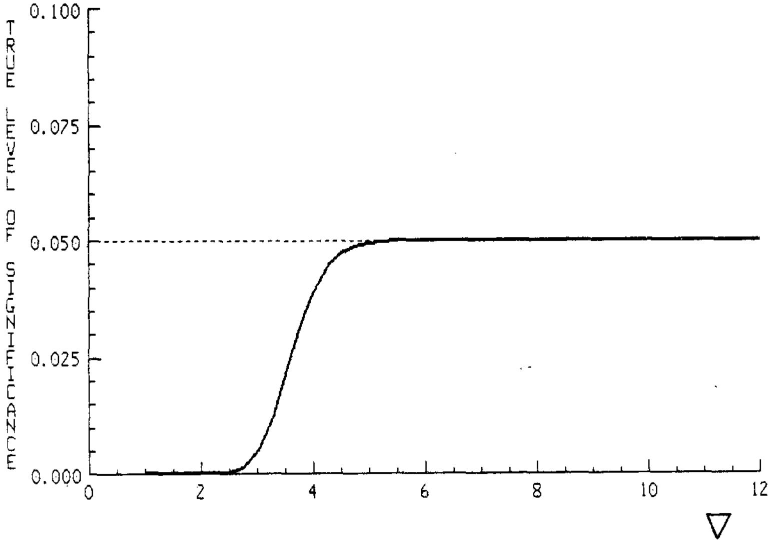

Fig. 6. True level of significance of the two one-sided tests procedure, with nominal level $\alpha = 0.05$, as a function of ∇, for the example of $\nu = 40$.

cedure for the example of nominal level 0.05 and 40 degrees of freedom. For very small values of ∇ (i.e., for very insensitive studies) the test is noticeably conservative, i.e., the true level of the test is substantially less than the nominal level of the one-sided tests. As ∇ increases, the true level becomes indistinguishable from the nominal level. The true level never exceeds the nominal level, so the nominal level of the one-sided tests may be set to the maximum probability of incorrectly concluding equivalence that can be tolerated.

In Figs. 4 and 5 and Table I it was noted that the behavior of the true level of significance of the power approach as a function of ∇ depended strongly on the degrees of freedom, ν. This is much less so for the two one-sided tests procedure. The rise of the true level from virtually zero for very small values of ∇ to virtually equal to the nominal level α for larger values of ∇ tends to be more abrupt for larger degrees of freedom. However, the basic shape is the same for all degrees of freedom, and once ∇ reaches around 5 or 6, the true level is practically indistinguishable from the nominal level, for nominal levels of 0.05 or more.

From these results, we can see that the two one-sided tests procedure permits us to control the probability of declaring the average bioavailabilities to be equivalent when they are, in fact, not equivalent. If α is the highest probability of this error we can tolerate, then by setting the nominal level of the two one-sided tests at α we are assured that the true level will not exceed α. Furthermore, the true level will be virtually equal to α for ∇ greater than 5 or so, i.e., if the study has a reasonable degree of precision. We still must be concerned that the study has sufficient sensitivity, but we need not be concerned, as was the case with the power approach, that the study has *too much* sensitivity.

In Figs. 4–6 the probability of concluding equivalence was plotted as a function of ∇, for a specific value of $\mu_T-\mu_R$, namely, $\mu_T-\mu_R=\theta_2$. In order to examine the probabilities of concluding that the means are equivalent when in fact they *are* equivalent, we plot the probabilities of concluding equivalence as a function of $\mu_T-\mu_R$, for particular values of ∇.

Figure 7 presents the probabilities of rejecting the interval hypothesis H_0, and thus concluding equivalence, for $\nabla=4$, which would correspond to a relatively insensitive study (to try to express this in more familiar terms, in a crossover study with 12 subjects and an equivalence criterion with θ_2 approximately equal to 20% of μ_R, a ∇ of 4 would roughly correspond to a *within*-subject coefficient of variation of 25%. With 24 subjects, it would roughly correspond to a within-subject *CV* of 36%). Figures 7–10 are for the case of $\nu=40$. Results for other degrees of freedom would be qualitatively similar. Both the two one-sided tests procedure with a nominal level of 0.05 and the power approach are illustrated.

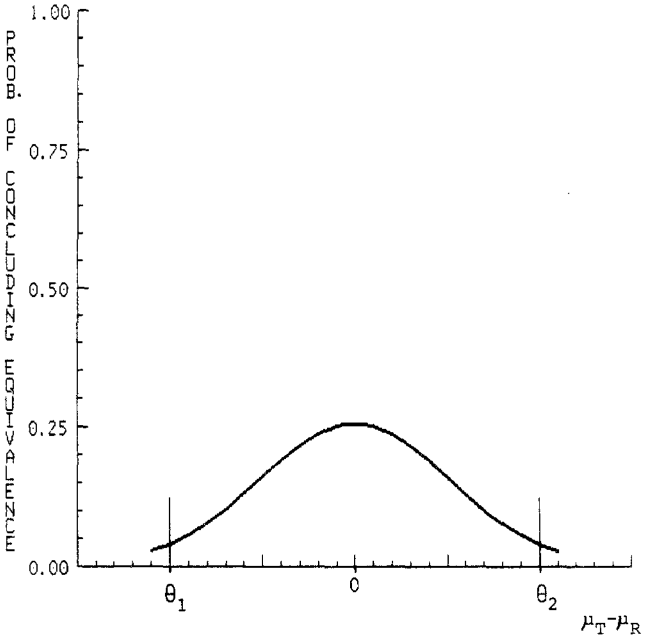

Fig. 7. Probability of concluding that μ_T and μ_R are equivalent, as a function of $\mu_T - \mu_R$, for the power approach (dashed line), and for the two one-sided tests procedure with nominal level $\alpha = 0.05$ (solid line), for the case of $\nabla = 4$, $\nu = 40$. Note that in this case the probabilities associated with the power approach are so low that they do not show up on the graph.

Looking first at the probabilities for the two one-sided tests procedure, $\alpha = 0.05$, we see that the probabilities at the edges of the equivalence interval ($\mu_T - \mu_R = \pm\theta_2$) are somewhat less than 0.05, as we already observed in Fig. 6. As $\mu_T - \mu_R$ moves towards zero, the probabilities increase to a peak of a little less than 0.25 at $\mu_T - \mu_R = 0$. Having only a 25% chance of concluding that the means are equivalent when in fact they are equal is a bit disappointing. But now consider the probabilities associated with the power approach. These probabilities are not zero, but they are so low that they do not show up on the graph. This is because with a study of this sensitivity ($\nabla = 4$) there is virtually no chance that the estimated power will be 0.80 or more.

In Fig. 8 we have the case of $\nabla = 8$. Here, the probability characteristics of the power approach do not look bad, but the probabilities for the two one-sided tests procedure, $\alpha = 0.05$, are uniformly better.

Figure 9 presents the case of $\nabla = 16$, an example of a very precise, sensitive study. For the power approach, the probability at the midpoint $\mu_T - \mu_R = 0$ is reasonably high, 0.95 to be exact. But for $\mu_T - \mu_R$ equal to

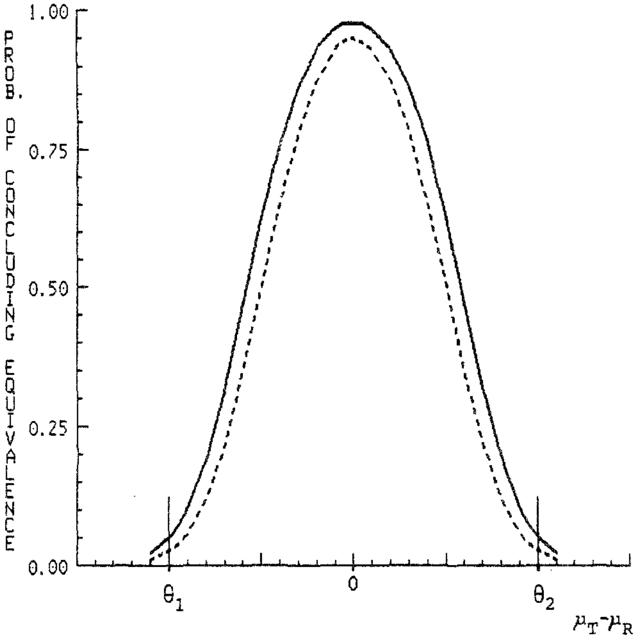

Fig. 8. Probability of concluding that μ_T and μ_R are equivalent, as a function of $\mu_T-\mu_R$, for the power approach (dashed line), and for the two one-sided tests procedure with nominal level $\alpha=0.05$ (solid line), for the case of $\nabla=8$, $\nu=40$.

about ±$0.6\theta_2$, the probabilities for the power approach have dropped almost to zero, even though we are still well within the equivalence interval. This is an illustration of the conservatism of the power approach for very large ∇, and results from the fact that for sensitive studies (large ∇) even very small true differences between μ_T and μ_R will be detected by the t test of the hypothesis of no difference H_0'.

Figure 10 presents the probabilities for $\nabla=6.214$ (this was the value of ∇ at which the true level of the power approach reached its peak in Fig. 5). Comparing first the two one-sided tests procedure, $\alpha=0.05$, to the power approach, we see that the height of the curve at the boundaries $\pm\theta_2$ is virtually equal to 0.05 for the two one-sided tests procedure, $\alpha=0.05$ (solid line), but is equal to about 0.096 for the power approach (dashed line). For values of $\mu_T-\mu_R$ toward the outsides of the equivalence interval, the probability of concluding equivalence is higher for the power approach, but for $\mu_T-\mu_R$ in about the middle one-third of the interval, the probability is higher for the two one-sided tests procedure, $\alpha=0.05$. This is the only

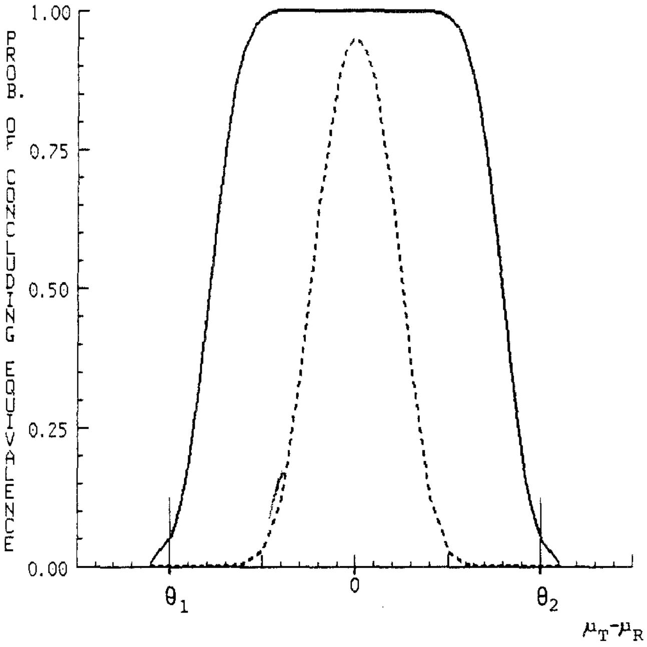

Fig. 9. Probability of concluding that μ_T and μ_R are equivalent, as a function of $\mu_T - \mu_R$, for the power approach (dashed line), and for the two one-sided tests procedure with nominal level $\alpha = 0.05$ (solid line), for the case of $\nabla = 16$, $\nu = 40$.

case we have considered where the two one-sided tests procedure with nominal $\alpha = 0.05$ is not uniformly superior to the power approach. However, since the true level of significance of the power approach in this case is about 0.096, it is not acceptable if we are only willing to take a 0.05 chance of saying the means are equivalent if in fact they are not equivalent. On the other hand, if we can tolerate a true level of around 0.096, then we should carry out the two one-sided tests at a nominal level of 0.096. With this in mind, the third curve in Fig. 10 presents the probabilities for the two one-sided tests procedure with a nominal α of 0.095 (a "rounder" number than 0.096). We see now that the two one-sided tests procedure, $\alpha = 0.095$, is as good or better than the power approach over the entire equivalence interval. Thus with appropriate choice of the nominal level α, the two one-sided tests procedure always has superior probability characteristics to the power approach.

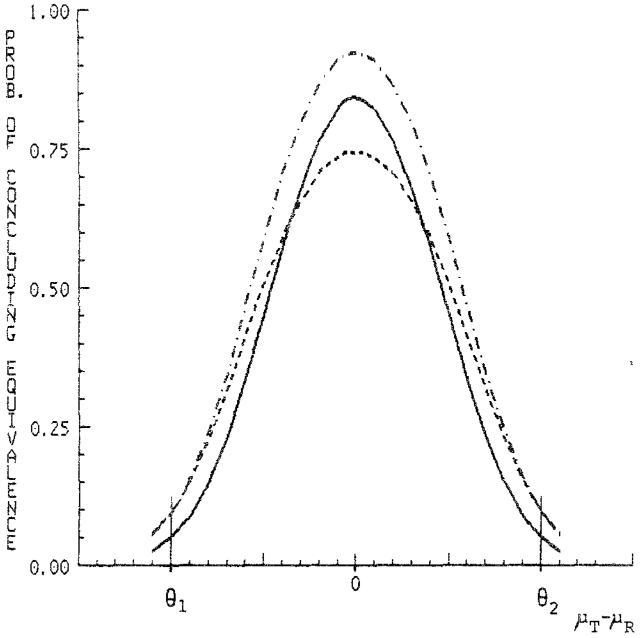

Fig. 10. Probability of concluding that μ_T and μ_R are equivalent, as a function of $\mu_T-\mu_R$, for the power approach (dashed line), the two one-sided tests procedure with nominal level $\alpha=0.05$ (solid line), and the two one-sided tests procedure with nominal level $\alpha=0.095$ (line — · —), for the case of $\nabla=6.214$, $\nu=40$.

COMPARISON OF THE TWO ONE-SIDED TESTS PROCEDURE TO THE PROCEDURE PROPOSED BY HAUCK AND ANDERSON

Figures 11 and 12 present values of $\bar{X}_T-\bar{X}_R$ and $s\sqrt{2/n}$ leading to rejection of the interval hypothesis H_0, and thus to a conclusion of equivalence of μ_T and μ_R, using the procedure proposed by Hauck and Anderson (1), for the example of $\alpha=0.05$, $\theta_2=20$ units, and $\nu=10$. The corresponding rejection region for the two one-sided tests procedure, $\alpha=0.05$, is included for reference. Hauck and Anderson's procedure (1) is the "central t approximation" examined in their earlier paper (3) in the statistical literature.

Based on that part of the picture that is shown in Fig., 11, Hauck and Anderson's procedure appears to have possible advantages over the two one-sided tests procedure, since inclusion of the extra part of the rejection region for the larger values of $s\sqrt{2/n}$ removes some of the conservatism

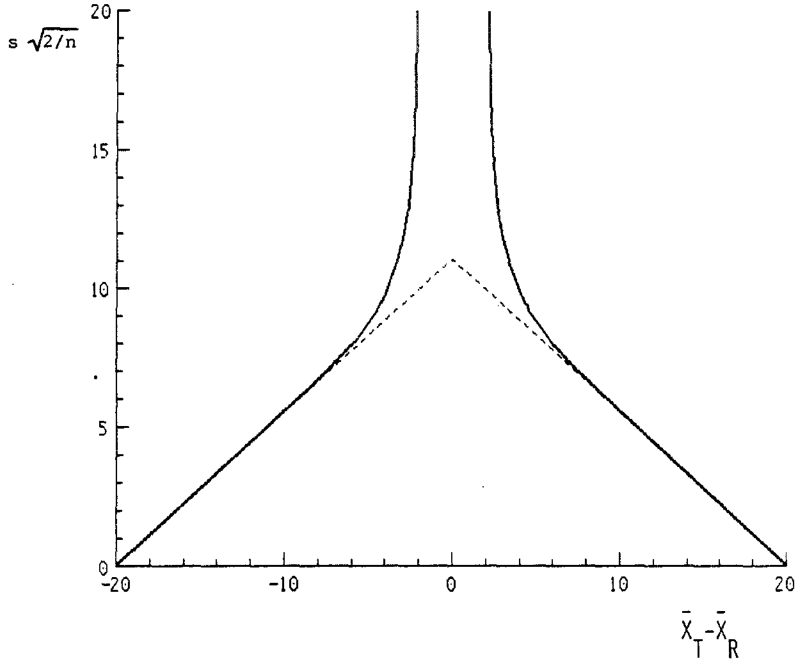

Fig. 11. Part of the rejection region for the procedure proposed by Hauck and Anderson (1), for the example of $\theta_2 = -\theta_1 = 20$ units, $\nu = 10$, and nominal $\alpha = 0.05$, illustrated for $s\sqrt{2/n}$ ranging from 0 to 20 units (solid line). The rejection region for the two one-sided tests procedure with nominal level $\alpha = 0.05$, $\nu = 10$, is included for reference (dashed line).

seen in the two one-sided tests procedure for moderately small values of ∇. When the whole picture is examined, however, there are some concerns. For $s\sqrt{2/n}$ greater than about 20, the boundaries of the Hauck and Anderson region start to spread apart, so the region has the nonconvex shape that was worrisome in the case of the power approach. Indeed, even Fig. 12 does not show the whole picture, since the rejection region for Hauck and Anderson's procedure goes on forever, getting wider and wider with increasing $s\sqrt{2/n}$. Eventually, the region is such that we would conclude that μ_T and μ_R are equivalent for values of $\bar{X}_T - \bar{X}_R$ that lie *outside* of the equivalence interval $[-20, 20]$, as has been noted by Rocke (4).

Recall that the two one-sided tests procedure (which as noted before is identical to the procedure, proposed by Westlake (2), of concluding equivalence if and only if the usual (shortest) $1-2\alpha$ confidence interval for $\mu_T - \mu_R$ is contained within the equivalence interval) consisted of testing the two one-sided hypotheses H_{01} and H_{02}. If we let p_1 be the p value associated with the test of H_{01} and p_2 be the p value associated with the

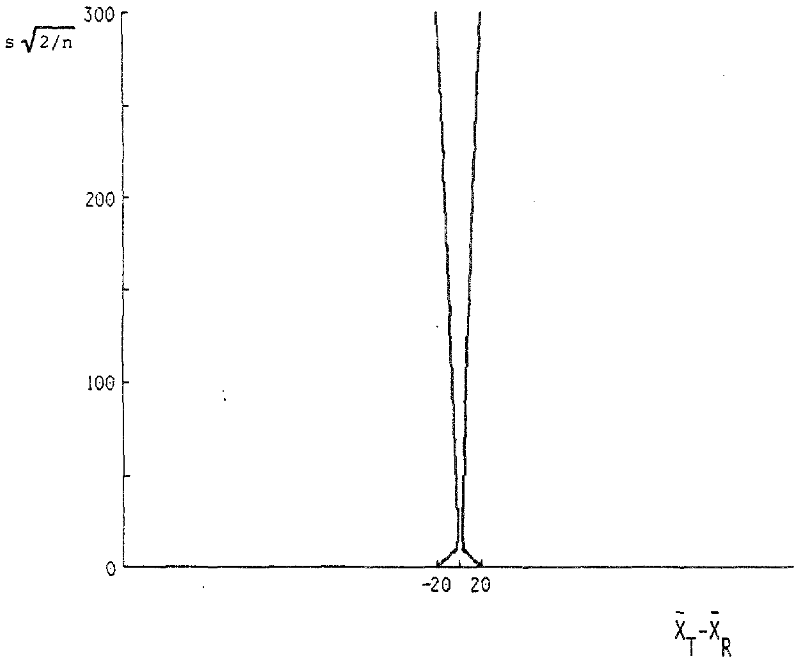

Fig. 12. Part of the rejection region for the procedure proposed by Hauck and Anderson (1), for the example of $\theta_2=-\theta_1=20$ units, $\nu=10$, and nominal $\alpha=0.05$, illustrated for $s\sqrt{2/n}$ ranging from 0 to 300 units.

test of H_{02}, then Anderson and Hauck (3) have pointed out that the p value for the test of the interval hypothesis H_0 associated with the two one-sided tests procedure is the larger of p_1 and p_2, i.e., $\max(p_1, p_2)$, while for the procedure proposed by Hauck and Anderson the p value is $|p_1-p_2|$. Thus, to take a hypothetical example, it would be possible to have a study for which p_1 was 0.44 and p_2 was 0.40, in which case the p value associated with the two one-sided tests procedure would be 0.44 while the p value associated with Hauck and Anderson's procedure would be 0.04. So a study that was completely inadequate to establish that $\theta_1<\mu_T-\mu_R$ ($p_1=0.44$), and completely inadequate to establish that $\mu_T-\mu_R<\theta_2$ ($p_2=0.40$), somehow is adequate to establish that $\theta_1<\mu_T-\mu_R<\theta_2$, based on the outcome of Hauck and Anderson's procedure.

Hauck and Anderson's procedure is always more powerful than the corresponding two one-sided tests procedure, i.e., there is always higher probability of concluding equivalence with Hauck and Anderson's procedure (the difference in power between the two procedures becomes negligible as ∇ becomes large). However, this is true when μ_T and μ_R are not equivalent as well as when they are equivalent, and indeed another drawback of Hauck and Anderson's procedure is that the true level of significance may exceed the nominal level α, particularly for small ν.

In the opinion of the author, the procedure proposed by Hauck and Anderson cannot be acceptable as it stands, with its open-ended rejection region that permits rejection of H_0, for some values of $\bar{X}_T - \bar{X}_R$, no matter how large $s\sqrt{2/n}$ is. There is such a thing as an inadequate study, and when a study is inadequate there should be *no* chance of concluding that μ_T and μ_R are equivalent. On the other hand, Hauck and Anderson's procedure does offer genuine advantages over the two one-sided tests procedure for values of ∇ that are moderately small, say ∇ approximately in the range of 2 or 3 up to 5. The best procedure to use may therefore turn out to be a compromise between the two procedures. This is a topic for further research.

ACKNOWLEDGMENT

The author would like to acknowledge the helpful suggestions of the referees.

APPENDIX A: PROPORTIONAL EQUIVALENCE CRITERIA

Until now, for purposes of this discussion, it has been assumed that θ_1 and θ_2 are known numbers, expressed in the same units as the bioavailability variable (*AUC*, $C_{\max}$, etc.) of interest. However, a more common situation in practice is for θ_1 and θ_2 to be expressed as proportions of the unknown reference average μ_R. Using the example of $\theta_1 = -0.20\mu_R$ and $\theta_2 = 0.20\mu_R$ (the common "±20%" criteria), the interval hypotheses would be stated as

$$H_0:\quad \mu_T - \mu_R \leq -0.20\mu_R \quad \text{or} \quad \mu_T - \mu_R \geq 0.20\mu_R$$

$$H_1:\quad -0.20\mu_R < \mu_T - \mu_R < 0.20\mu_R$$

which, if $\mu_R > 0$, may be restated as

$$H_0:\quad \mu_T/\mu_R \leq 0.80 \quad \text{or} \quad \mu_T/\mu_R \geq 1.20$$

$$H_1:\quad 0.80 < \mu_T/\mu_R < 1.20$$

The problem, therefore, is no longer stated in terms of the difference of μ_T and μ_R, but rather in terms of the ratio of μ_T and μ_R.

Hauck and Anderson (1) noted that it is often deemed appropriate to assume that the statistical assumptions of normality, homogeneous variance, etc. are satisfied for the logarithmically transformed variables. This is based on theoretical arguments involving pharmacokinetic compartmental models

(5) and also on the empirical observation that observed distributions of bioavailabilty variables are often skewed, with a long "tail" of higher values. If the statistical assumptions that have been made are in fact true on the logarithmic scale, then the interval hypotheses, for the example of the ±20% criteria, may be restated as

$$H_0: \quad \eta_T - \eta_R \leq \log(0.8) \quad \text{or} \quad \eta_T - \eta_R \geq \log(1.2)$$

$$H_1: \quad \log(0.8) < \eta_T - \eta_R < \log(1.2)$$

where η_T and η_R are the true test and reference means, respectively, of the logarithmically transformed variables. Logarithms to the base 10 or natural logarithms may be used.

Under this circumstance, the two one-sided tests procedure or Hauck and Anderson's procedure would be carried out as before, with log (0.8) taking the role of θ_1 and log (1.2) taking the role of θ_2. All of the results cited earlier concerning the probability characteristics of the two one-sided tests procedure still apply.

In the case of the power approach, there is an additional difficulty, if the approach is to be based on the test of the hypothesis of no difference

$$H_0': \quad \eta_T - \eta_R = 0$$

$$H_1': \quad \eta_T - \eta_R \neq 0$$

The difficulty is that since $\log(0.8) \neq -\log(1.2)$, the estimated power at $\eta_T - \eta_R = \log(0.8)$ will not be the same as the estimated power at $\eta_T - \eta_R = \log(1.2)$. If one was determined to use the power approach in the case of logarithmically transformed variables, one way to do it would be to base the approach on the test of

$$H_0'': \quad \eta_T - \eta_R = (\log(0.8) + \log(1.2))/2$$

$$H_1'': \quad \eta_T - \eta_R \neq (\log(0.8) + \log(1.2))/2$$

That is, the hypothesis H_0'' that the difference of means on the log scale is equal to the midpoint of the equivalence interval [log (0.8), log (1.2)] (where, once again, we are considering the example of ±20% equivalence criteria). However, persons who use the power approach seem loath to abandon the hypothesis of no difference. Once again, the concept of equality is confused with the concept of equivalence. In any event, even if the power approach in this context is based on H_0'' instead of H_0', it still has the unfavorable properties presented before.

Under the assumption that the variances of the test and reference formulations are the same on the log scale (which corresponds to the assumption that the two formulations have comparable coefficients of variation on the original scale), the mean difference $\eta_T - \eta_R$ is in fact equal to the logarithm of μ_T/μ_R, the ratio of means on the original scale (6). If the variances are not equal, then $\eta_T - \eta_R$ is *not* equal to $\log \mu_T/\mu_R$. On the other hand, $\eta_T - \eta_R$ is equal to the logarithm of the ratio of medians of T and R whether the variances are equal or not. Thus, in the case of logarithmically transformed variables, more care than usual should be given to the question of what is meant by "average" bioavailability (i.e., mean or median), and if interest lies in the means, some attention should be paid to the assumption of equal variances on the log scale.

The question of what to do if the equivalence criteria are stated in proportional terms, but the assumptions of normality and additivity of the statistical model are thought to be satisfied on the *original* scale, is outside the scope of this discussion. However, we note that Locke (7) has described an exact method for obtaining a confidence interval (or confidence set) for μ_T/μ_R under these circumstances.

APPENDIX B. UNBALANCED CROSSOVER STUDIES

The assumption was made that the bioavailability/bioequivalence study under consideration was a *balanced* crossover study, that is

1. There is an equal number of subjects in each treatment-administration sequence.
2. There are no missing observations from any subject.

All of the results cited above concerning the properties of the two one-sided tests procedure and the power approach are equally true for unbalanced crossover studies. If we let *Est.* be the estimator of $\mu_T - \mu_R$, and SE be the standard error of the estimator, then the two one-sided tests procedure utilizes the two test statistics

$$t_1 = \frac{Est. - \theta_1}{\mathrm{SE}} \quad \text{and} \quad t_2 = \frac{\theta_2 - Est.}{\mathrm{SE}}$$

In the case of balanced studies, the estimator *Est.* is in the fact the difference of the *observed* means, $\bar{X}_T - \bar{X}_R$, and therefore the standard error SE is equal to $s\sqrt{2/n}$. In the case of unbalanced studies, the best unbiased (least squares) estimator of $\mu_T - \mu_R$ is *not*, in general, the difference of observed means.

For the special case of a two-treatment, two-period crossover study in which n_1 subjects receive the test formulation in period one and the reference formulation in period two, while n_2 subjects receive the reference formulation in period one and the test formulation in period two, the unbiased estimator is given by

$$Est. = \frac{(\bar{X}_{T1} + \bar{X}_{T2})}{2} - \frac{(\bar{X}_{R1} + \bar{X}_{R2})}{2}$$

where

$\bar{X}_{\mathrm{T1}}$ = the observed mean of the n_1 observations on the test formulation in period one.

$\bar{X}_{\mathrm{T2}}$ = the observed mean of the n_2 observations on the test formulation in period two.

$\bar{X}_{\mathrm{R1}}$ = the observed mean of the n_2 observations on the reference formulation in period one.

$\bar{X}_{\mathrm{R2}}$ = the observed mean of the n_1 observations on the reference formulation in period two.

The standard error of this estimator is

$$\mathrm{SE} = s\sqrt{\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

where as before s is the square root of the "error" mean square from the crossover design analysis of variance, based on ν degrees of freedom.

Note that if $n_1 = n_2$, these formulas reduce to *Est.* $= \bar{X}_{\mathrm{T}} - \bar{X}_{\mathrm{R}}$ and $\mathrm{SE} = s\sqrt{2/n}$ (where n = total number of subjects $= n_1 + n_2$) as before.

In the case of a study with more than two treatments (formulations) and/or periods, the formulas for *Est.* and SE will depend on the particular pattern of unbalance, and can be very complicated. Usually a computer routine will be needed to obtain them.

APPENDIX C. METHOD OF OBTAINING THE FIGURES

The probabilities illustrated in Figs. 4–10 were obtained by numerical integration on the joint probability distribution of $\bar{X}_{\mathrm{T}} - \bar{X}_{\mathrm{R}}$ and $s\sqrt{2/n}$. Under the assumptions made, $\bar{X}_{\mathrm{T}} - \bar{X}_{\mathrm{R}}$ has a normal distribution with mean $\mu_{\mathrm{T}} - \mu_{\mathrm{R}}$ and variance $\sigma^2(2/n)$. The distribution of $s\sqrt{2/n}$ is related to the χ^2 distribution by the fact that $\nu(s\sqrt{2/n})^2/\sigma^2(2/n)$ has a χ^2 distribution with ν degrees of freedom. Furthermore, $\bar{X}_{\mathrm{T}} - \bar{X}_{\mathrm{R}}$ and $s\sqrt{2/n}$ are statistically independent. From these facts, the joint probability distribution of $\bar{X}_{\mathrm{T}} - \bar{X}_{\mathrm{R}}$ and $s\sqrt{2/n}$ may be obtained.

The probability of rejection, for either procedure, depends on $\mu_{\mathrm{T}} - \mu_{\mathrm{R}}$, ∇, and ν, and in the case of the two one-sided tests procedure, on the nominal level of significance α. It actually depends on $\mu_{\mathrm{T}} - \mu_{\mathrm{R}}$ only through the quantity g, defined by the relationship

$$\begin{aligned}\mu_{\mathrm{T}} - \mu_{\mathrm{R}} &\equiv \frac{\theta_1 + \theta_2}{2} + g\frac{\theta_2 - \theta_1}{2}\\ &= g\theta_2 \quad \text{for the case } \theta_1 = -\theta_2\end{aligned}$$

The true level of significance of the procedure is obtained by letting $\mu_T - \mu_R = \theta_2$, i.e., $g = 1$.

For each given set of values of g, ∇, ν, and α, the joint probability distribution was integrated over the appropriate rejection region, i.e., regions like Fig. 2 for the two one-sided tests procedure, regions like Fig. 1 for the power approach.

The boundaries of the rejection region for Hauck and Anderson's procedure, pictured in Figs. 11 and 12, were obtained by solving the equation

$$F_\nu\left(\frac{1+c}{w}\right) - F_\nu\left(\frac{1-c}{w}\right) = 0.05$$

numerically for c ($c > 0$) for a series of values of w, where $F_\nu(\cdot)$ is the cumulative distribution function of the Student's t distribution with ν degrees of freedom. For each pair of c and w so obtained, $(-c, w)$ and (c, w) are points on the boundary of the rejection region for the case of $\theta_2 = -\theta_1 = 1$ unit. $(-20c, 20w)$ and $(20c, 20w)$ are the corresponding boundary points for the example of $\theta_2 = -\theta_1 = 20$ units.

REFERENCES

1. 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).
2. W. J. Westlake. Response to T. B. L. Kirkwood: Bioequivalence testing—a need to rethink. *Biometrics* **37:**589-594 (1981).
3. S. Anderson and W. W. Hauck. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Comm. Stat. A* **12:**2663-2692 (1983).
4. D. M. Rocke. On testing for bioequivalence. *Biometrics* **40:**225-230 (1984). (See also Correspondence in *Biometrics* **41:**561-563 (1985).)
5. 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 and Febiger, Philadelphia, 1973, pp. 149-179.
6. D. Mandallaz and J. Mau. Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics* **37:**213-222 (1981).
7. C. S. Locke. An exact confidence interval from untransformed data for the ratio of two formulation means. *J. Pharmacokin. Biopharm.* **12:**649-655 (1984).