ELSEVIER

Journal of Statistical Planning and Inference 99 (2001) 41–58

journal of statistical planning and inference

www.elsevier.com/locate/jspi

A nearly unbiased test for individual bioequivalence problems using probability criteria

Weizhen Wang[a], J.T. Gene Hwang[b,*]

[a] *Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, USA*
[b] *Department of Mathematics, Cornell University, Ithaca, NY 14853, USA*

Received 22 October 1999; received in revised form 22 November 2000; accepted 14 December 2000

Abstract

Statistical bioequivalence has recently attracted lots of attention. This is perhaps due to the importance of setting a reasonable criterion on the part of a regulatory agency such as the FDA in the US in regulating the manufacturing of drugs (especially generic drugs). Pharmaceutical companies are obviously interested in the criterion since a huge profit is involved. Various criteria and various types of bioequivalence have been proposed. At present, the FDA recommends testing for average bioequivalence. The FDA, however, is considering replacing average bioequivalence by individual bioequivalence. We focus on the criterion of individual bioequivalence proposed earlier by Anderson and Hauck (J. Pharmacokinetics and Biopharmaceutics 18 (1990) 259) and Wellek (Medizinische Informatik und Statistik, vol. 71, Springer, Berlin, 1989, pp. 95–99; Biometrical J. 35 (1993) 47). For their criterion, they proposed TIER (test of individual equivalence ratios). Other tests were also proposed by Phillips (J. Biopharmaceutical Statist. 3 (1993) 185), and Liu and Chow (J. Biopharmaceutical Statist. 7 (1997) 49). In this paper, we propose an alternative test, called nearly unbiased test, which is shown numerically to have power substantially larger than existing tests. We also show that our test works for various models including 2×3 and 2×4 crossover designs.

MSC: 62F03; 62K99; 62P99

Keywords: Crossover design; Noncentral t-distribution; Power; Type I error

1. Introduction

Statistical bioequivalence has recently become a very active research area. See the book of Chow and Liu (1992) and the survey paper of Berger and Hsu (1996). The idea is to "prove" statistically that two drugs or formulations (called the test drug and the reference drug) are equivalent. At present, to seek the approval of the US FDA and the European Community EC-GCP for a drug through bioequivalence, a

* Corresponding author.
E-mail addresses: wwang@math.wright.edu (W. Wang), hwang@math.cornell.edu (J.T. Gene Hwang).

0378-3758/01/$ - see front matter
PII: S0378-3758(01)00074-X

pharmaceutical company only needs to apply the test drug and the reference drug (usually a brand-name drug) to typically 24 subjects using a 2×2 crossover design and compare the amount of a certain active ingredient in the blood samples. See FDA (1992) and EC-GCP (1993). If the mean AUC (area under the concentration versus time curve) or other pharmacokinetic variables of the two treatments are equivalent statistically, then the two drugs are declared bioequivalent and the test drug can be marketed. Statistically, this is done by rejecting the null hypothesis stating that the two treatments are *not* equivalent. Unlike developing an original drug which typically takes 10 years of clinical trials in several phases and costs more than 200 million dollars, seeking approval in bioequivalence requires only a few months to conduct the experiment and to prepare for an application, costing 2 million dollars or so. The saving in time and money is tremendous, which obviously attracts pharmaceutical companies. Since it helps to reduce the production cost, it should help reduce the medical cost on the part of patients.

However, there is a concern about whether the consumers are well protected. The present FDA guidance (1992) requires only the demonstration of *average bioequivalence*; that is bioequivalence, in terms of averages or expectations of the observed pharmacokinetic variables. However, this causes concern and it has been argued that perhaps not only the expectations, but also the variabilities or even the distributions, should be shown to be equivalent. Bioequivalence in distribution is called the *population bioequivalence*. See Anderson and Hauck (1992), Liu and Chow (1992), Wang (1997), Schall and Luus (1993) and Schall (1995).

Moreover, it has been pointed out that one should show that the two treatments are switchable (see Anderson and Hauck, 1990; Sheiner, 1992). That is, one can expect that the effects are similar when replacing one by the other for any given individual. Because of this, Anderson and Hauck (1990) as well as Wellek (1989, 1993) proposed a procedure which Anderson and Hauck called TIER, *test of individual equivalence ratios*. Later on Phillips (1993) and Liu and Chow (1997) proposed parametric tests. Some details of these procedures are discussed in Section 2. Recently, the FDA (1999) has a draft proposal written which is under circulation to solicit opinions from experts and which proposes to replace the average bioequivalence approach by population and individual bioequivalence approach. The draft suggests considering the moment criteria. See Wang (1999) for a valid level-α test for this problem. Our paper, however, focuses on probability criteria which, as pointed out by a referee, is simpler and easier to interpret.

In this paper, we propose a new procedure, called a *nearly unbiased test* in Section 3. Our test turns out to be much more powerful than TIER and the procedure of Liu and Chow; see Fig. 3. Although Phillips' test has power similar to our test in Fig. 3, Fig. 4 demonstrates a case where our test is much more powerful. By comparing type I errors, it can also be concluded that there are many other cases where our test improves upon other tests substantially. In the cases of Figs. 3 and 4 a 2×2 crossover design with subject effects involving totally 24 subjects is assumed. The setting of the figures are carefully discussed in Example 4.1 of Section 4. Our nearly unbiased test

also applies to more sophisticated designs involving more than two periods such as 2×3 or 2×4 designs. In Section 5, the sample sizes of our test are given, which are about two-thirds or less compared to other tests.

2. Review of previous results

Anderson and Hauck (1990) assume a simplified 2×2 crossover design without period effects

$$Y_{ij} = m_{ij} + e_{ij}, \tag{2.1}$$

where Y_{ij}, the data is the response (in ln scale) of the ith subject, $1 \leqslant i \leqslant n$, and jth formulation, $j = \mathrm{R}$ or T. Here and throughout the paper, R stands for the reference drug and T stands for the test drug. Also in (2.1), m_{ij} is the random formulation effect and e_{ij} represents the within subject variation.

Let

$$Y_i = m_i + e_i,$$

where

$$Y_i = Y_{i\mathrm{T}} - Y_{i\mathrm{R}}, \quad m_i = m_{i\mathrm{T}} - m_{i\mathrm{R}} \quad \text{and} \quad e_i = e_{i\mathrm{T}} - e_{i\mathrm{R}}.$$

Assume that m_i, independent of e_i, and (m_i, e_i), $1 \leqslant i \leqslant n$, are independently identically distributed (i.i.d.). Anderson and Hauck (1990) and Wellek (1989, 1993) proposed to test

$$\mathrm{H}_0\colon\ p \leqslant p_0 \quad \text{vs.} \quad \mathrm{H}_1\colon\ p > p_0, \text{ where } p = P(|m_i| < \Delta) \tag{2.2}$$

and Δ and p_0 $(> \frac{1}{2})$ are prespecified quantities. If the null hypothesis is rejected, individual bioequivalence is then established. The test proposed by Anderson and Hauck (1990) and Wellek (1989, 1993) is named TIER in the former paper. TIER defines the p-value as

$$P(X \geqslant x)$$

where X has a binomial distribution of n trials and with p_0 as the success probability and

$$x = \text{number of } i\text{'s such that } |Y_i| < \Delta.$$

There are two issues regarding TIER. First, is TIER valid? That is, does the test that rejects H_0 if and only if the p-value is less than α has a type I error less than or equal to α? To answer this question, note that TIER is obviously valid for the following hypotheses:

$$\mathrm{H}_0\colon\ p \leqslant p_0 \quad \text{vs.} \quad \mathrm{H}_1\colon\ p > p_0, \text{ where } p = P(|Y_i| < \Delta). \tag{2.3}$$

These types of hypotheses were considered in Schall and Luus (1993), Schall (1995), Phillips (1993) and Liu and Chow (1997). It was also shown in Hwang and Wang (1997) that H_0 in (2.2) implies H_0 in (2.3), under the conditions that

$$p_0 \geqslant \tfrac{1}{2} \text{ and } e_i \text{ and } m_i \text{ are unimodal and symmetric.} \tag{2.4}$$

(These conditions obviously apply to the case where m_i and ε_i are normally distributed, an assumption usually assumed in applications.) Since TIER is valid for H_0 in (2.3), TIER is valid for (2.2) under (2.4). Similarly, any test valid for (2.3) is valid for (2.2). Due to this, we shall focus on (2.3).

The other issue relating to TIER is its low power. See Schall and Luus (1993). This concern was shared by Phillips (1993) and Liu and Chow (1997), who took another approach to consider a more parametric model and construct a procedure based on its sufficient statistics. Their procedure has the advantage of handling other effects such as the period effect as well as designs more general than the standard two-sequence, two period crossover design. We shall first deal with the following model without period effect:

$$Y_{ij} = \mu + F_j + S_i + \varepsilon_{ij}, \tag{2.5}$$

where $i = 1, \ldots, n$, when n is the total number of subjects; $j = \mathrm{T}$ or R; μ is the overall mean; F_j is the fixed jth formulation effect with $F_\mathrm{R} + F_\mathrm{T} = 0$; S_i is the random subject effect; ε_{ij} is the random error in observing Y_{ij}. Further, assume that

$$S_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_S^2), \quad \varepsilon_{i\mathrm{T}} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_\mathrm{T}^2) \quad \text{and} \quad \varepsilon_{i\mathrm{R}} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_\mathrm{R}^2) \tag{2.6}$$

are statistically independent.

Note that the above model is closely related to the following model:

$$\begin{aligned} Y_{i\mathrm{T}} &= u_\mathrm{T} + b_{i\mathrm{T}} + e_{i\mathrm{T}}, \\ Y_{i\mathrm{R}} &= u_\mathrm{R} + b_{i\mathrm{R}} + e_{i\mathrm{R}}, \end{aligned} \tag{2.7}$$

where u_j ($j = \mathrm{T}, \mathrm{R}$) is the mean response, b_j is the mean deviation from the population average of a given individual effect, and e_j represents the within-subject variation. Model (2.7) is used in Sheiner (1992), Schall and Luus (1993) and Schall (1995) and is apparently assumed in FDA (1997) (see Schall and Williams, 1996). Note that b_T and b_R may be correlated.

To see the relationship between models (2.5) and (2.7), we shall consider ε_{ij} in (2.5) as the sum of two errors

$$\varepsilon_{ij} = d_{ij} + e_{ij},$$

where e_{ij} is the error incurred in measuring the bioavailabilities and d_{ij} is the deviation of the ith individual subject effect from S_i. Substituting ε_{ij} in (2.5) by $d_{ij} + e_{ij}$, we may write model (2.5) as (2.7) where $u_j = \mu + F_j$, and $b_{ij} = S_i + d_{ij}$. Here $b_{i\mathrm{R}}$ and $b_{i\mathrm{T}}$ are correlated.

Example 2.1. Now return to (2.5) and (2.6). It follows that

$$Y_i = Y_{i\mathrm{T}} - Y_{i\mathrm{R}} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2), \tag{2.9}$$

where

$$\theta = F_{\mathrm{T}} - F_{\mathrm{R}} \quad \text{and} \quad \sigma^2 = \sigma_{\mathrm{R}}^2 + \sigma_{\mathrm{T}}^2.$$

The sufficient statistics $(Y, \hat{\sigma})$ for θ and σ are

$$Y = \sum_{i=1}^{n} Y_i/n \sim \mathrm{N}(\theta, \sigma^2/n) \tag{2.10}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum (Y_i - \bar{Y})^2. \tag{2.11}$$

Note that $\hat{\sigma}^2$ is an unbiased estimator for σ^2 and $(n-1)\hat{\sigma}^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom. Let

$$p = P(|Y_i| < \Delta) = \Phi\left(\frac{\Delta - \theta}{\sigma}\right) - \Phi\left(-\frac{\Delta + \theta}{\sigma}\right), \tag{2.12}$$

where Φ is the cumulative distribution function of a standard normal distribution and Δ is a quantity usually prespecified by a regulatory office. The problem then is to test hypotheses (2.3) or equivalently,

$$\mathrm{H}_0\colon\ \Phi\left(\frac{\Delta - \theta}{\sigma}\right) - \Phi\left(-\frac{\Delta + \theta}{\sigma}\right) \leqslant p_0 \quad \text{vs.} \quad \mathrm{H}_A\colon\ \mathrm{H}_0 \text{ fails.} \tag{2.13}$$

The curve consisting of (θ, σ) such that $p = p_0$, i.e.,

$$\Phi\left(\frac{\Delta - \theta}{\sigma}\right) - \Phi\left(-\frac{\Delta + \theta}{\sigma}\right) = p_0 \tag{2.14}$$

is called *Anderson and Hauck's curve* in this paper. Fig. 1 gives the graph for a specific value of $n = 24$, $\Delta = \ln(1.25)$ and $p_0 = 0.8$ as well as the rejection regions of several tests discussed below. We emphasize here that Anderson and Hauck's curve is on the (θ, σ) plane, i.e., the parameter space, while the others are on the $(Y, \hat{\sigma})$ plane. It seems interesting to put them together for a comparison. We can rewrite (2.14) as $|\theta| = g_{\mathrm{A\text{–}H}}(\sigma)$ where for each σ, $g_{\mathrm{A\text{–}H}}(\sigma)$ is $|\theta|$ where θ is the solution to (2.14). Using this, H_A of (2.13) can be written as

$$|\theta| < g_{\mathrm{A\text{–}H}}(\sigma). \tag{2.15}$$

However, there is no closed-form representation for $g_{\mathrm{A\text{–}H}}$. Consequently, it is very difficult to construct an unbiased test. Phillips (1993) and Liu and Chow (1997) worked with an alternative region, the largest triangle inside Anderson and Hauck's curve. Similar to the two triangles drawn in Fig. 1, the largest triangle, not drawn in Fig. 1, has $(-\Delta, \Delta)$ as its base, but its top touches the highest point of Anderson and Hauck's curve. Liu and Chow (1997) constructed a test by taking the intersection

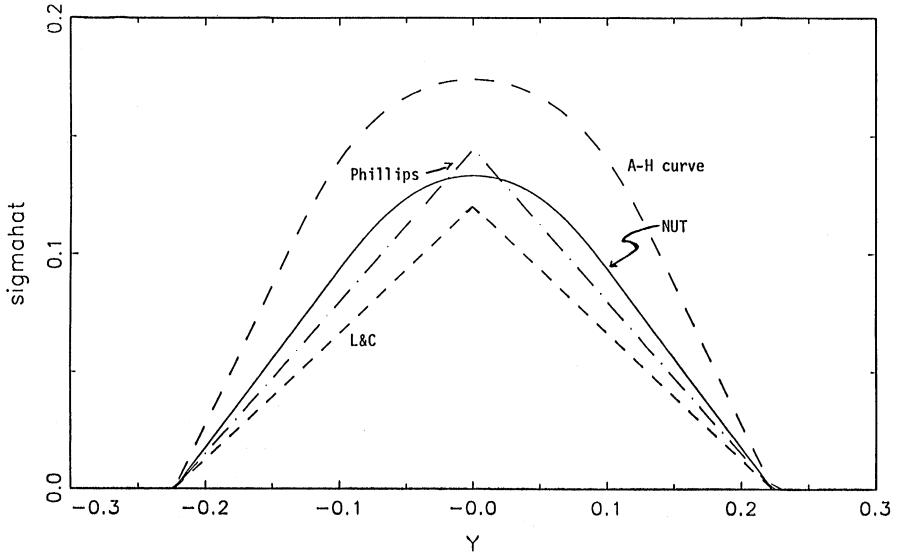

Fig. 1. Anderson and Hauck's (A–H) curve, rejection regions of Phillips' test, the nearly unbiased test (NUT) and Liu and Chow's test (L&C) with $n = 24$, $v = 23$, $p_0 = 0.8$, $\Delta = \ln(1.25)$, $\alpha = 0.05$. The same Δ and α are used in all figures below. Note that A–H curve is on the (θ, σ) plane but is superimposed with other regions which are on the $(Y, \hat{\sigma})$ plane.

of the rejection regions of two tests each valid for testing against the region above the line containing one side of the largest triangle. Their test is obviously valid for testing against the complement of the largest triangle and hence valid for (2.13). This approach has the advantage of being strictly valid. However, the disadvantage is that its size could be much less than the nominal level and it results in a test whose power is low. Phillips' rejection region, however, has a similar form as Liu and Chow's (1997) but uses properly adjusted cutoff point so that its type I error at $\theta = 0$ on the Anderson and Hauck's curve equals α. Consequently, Phillips' rejection region represented by the larger triangle of the two in Fig. 1 properly contains the rejection region of Liu and Chow (1997), represented by the smaller triangle in Fig. 1. We shall discuss a test to be described in the next section, whose size is numerically demonstrated to be α and which is more powerful than the previous tests.

3. Construction of the proposed test

We shall describe our test for (2.13) based on the canonical form statistics Y and $\hat{\sigma}$ which are statistically independent with the distributions

$$Y \sim \mathrm{N}(\theta, \tau^2), \quad \frac{v\hat{\sigma}^2}{\sigma^2} \sim \chi_v^2 \tag{3.1}$$

where $v > 0$ is the degrees of freedom. Obviously, Example 2.1 reduces to this form with $v = n - 1$ and $\tau^2 = \sigma^2/n$. Later in Section 4 we shall discuss other models which can be reduced to this form as well.

As suggested by (2.15), we focus on tests with a rejection region of the form

$$|Y| < T(\hat{\sigma}), \tag{3.2}$$

where T is a function at our disposal. The tests of Phillips (1993) and Liu and Chow (1997) use a linear function T; we, however, work with a nonlinear function. Below we summarize the results, the proofs of which are straightforward and can be found in Lemma 7.2.2 of Wang (1995).

Lemma 1. *The power function, i.e., the probability that* (3.2) *holds is nonincreasing in* $|\theta|$. *Furthermore, if*

$$\tau = r\sigma \tag{3.3}$$

and

$$T(\hat{\sigma})/\hat{\sigma} \quad \textit{is nondecreasing in } \hat{\sigma} \tag{3.4}$$

for a fixed constant r, then the power function is nonincreasing in σ *when* $\theta = 0$.

Note that τ/σ is usually a constant ($=1/n^{1/2}$ in Example 2.1, for instance), and hence (3.3) is satisfied for some r.

This lemma guarantees that (3.2) defines a size-α test provided that the supremum of the type I error at Anderson and Hauck's curve is equal to α. Additionally, if the type I error for every (θ, σ) on the curve is α, then the test is *unbiased*, i.e.,

$$\text{its power} \leqslant \alpha \ \forall(\theta, \sigma) \text{ satisfying } H_0$$
$$\text{its power} \geqslant \alpha \ \forall(\theta, \sigma) \text{ satisfying } H_1.$$

Furthermore, since Anderson and Hauck's curve is compact, the calculations of type I error of (3.2) based on numerical integration on a set dense enough on the curve provide decisive evidence on whether the test is of size-α. Below the evidence shows that our test is always valid and is nearly unbiased. An unbiased test or a nearly unbiased test is desirable, since its rejecting probability (or type I error) achieves or nearly achieves the largest allowable level under H_0 and consequently its power or probability of rejection under alternative hypothesis often is larger than other valid tests. This phenomenon is demonstrated in the figures to be shown later.

Now return to the construction of the nearly unbiased test. The problem is how to choose $T(\cdot)$. The following theorem can be used to guide us in choosing $T(\cdot)$. Below, we shall use z_{p_0} to denote the p_0 quantile of a standard normal random variable hence $\Phi(z_{p_0}) = p_0$, and use $t_{\alpha,v}(\eta)$ to denote the α quantile of a noncentral t distribution with v degrees of freedom and noncentrality η.

Theorem 1. *Assume that* $T(\cdot)$ *is continuously differentiable for* $\hat{\sigma} \geqslant 0$; (θ, σ) *is on the Anderson and Hauck's curve. If* (3.3) *holds for a finite positive constant r, then*

the probability of (3.2) *approaches* α, $0 < \alpha < 1$, *as* $\sigma \to 0$ *if and only if* (i) *and* (ii) *below hold*:

(i) $\lim_{\hat{\sigma}\to 0} T(\hat{\sigma}) = \Delta$;

(ii) $\lim_{\hat{\sigma}\to 0} T'(\hat{\sigma}) = rt_{\alpha,v}(-z_{p_0}/r)$.

Proof. See the appendix.

Actually, condition (3.3) of the theorem can be replaced by a weaker condition

$$\lim_{\sigma\to 0} \frac{\tau}{\sigma} = r. \tag{3.5}$$

Now we return to the construction of our proposed test. Note that the alternative space in (2.13) has the form

$$\Phi\left(\frac{\Delta-\theta}{\sigma}\right) - \Phi\left(-\frac{\Delta+\theta}{\sigma}\right) > p_0. \tag{3.6}$$

It then seems reasonable to estimate θ and σ by Y and $\hat{\sigma}$. The rejection region becomes

$$\Phi\left(\frac{\Delta-Y}{\hat{\sigma}}\right) - \Phi\left(-\frac{\Delta+Y}{\hat{\sigma}}\right) > K. \tag{3.7}$$

Theorem 2. *Inequality* (3.7) *can be rewritten in form of* (3.2), *i.e.*,

$$|Y| < T_N(\hat{\sigma}). \tag{3.8}$$

If $K \geqslant \frac{1}{2}$, *then* T_N *is a nonincreasing function, and conditions* (i) *and* (ii) *of Theorem* 1 *are satisfied if one chooses*

$$K = \Phi(-rt_{\alpha,v}(-z_{p_0}/r)). \tag{3.9}$$

Proof. See the appendix.

As a technical remark, the condition $K \geqslant \frac{1}{2}$ is satisfied if $p_0 \geqslant \frac{1}{2}$. Rejection region (3.7) with K defined in (3.9) is the proposed test of this paper. By Theorem 2 and Lemma 1, whether the test is a size-α or unbiased test depends on its type I error on Anderson and Hauck's curve. We shall show that type I error of the test on Anderson and Hauck's curve is nearly nominal and hence it will be called the *nearly unbiased test*. The test will be numerically shown to improve upon the other tests significantly.

We are thankful to a referee who pointed out to us that it is possible to find the p-value corresponding to the procedure by solving α in the following equation:

$$\Phi\left(\frac{\Delta-Y}{\hat{\sigma}}\right) - \Phi\left(-\frac{\Delta+Y}{\hat{\sigma}}\right) = \Phi(-rt_{\alpha,v}(-z_{p_0}/r))$$

which results in the p-value

$$F_{v,-z_{p_0}/r}\left(\frac{-1}{r}\Phi^{-1}\left(\Phi\left(\frac{\Delta-Y}{\hat{\sigma}}\right) - \Phi\left(-\frac{(\Delta+Y)}{\hat{\sigma}}\right)\right)\right).$$

Here $F_{v,-z_0/r}$ represents the cumulative distribution function of a t-distribution with v degrees of freedom and $-z_{p_0}/r$ noncentrality parameter.

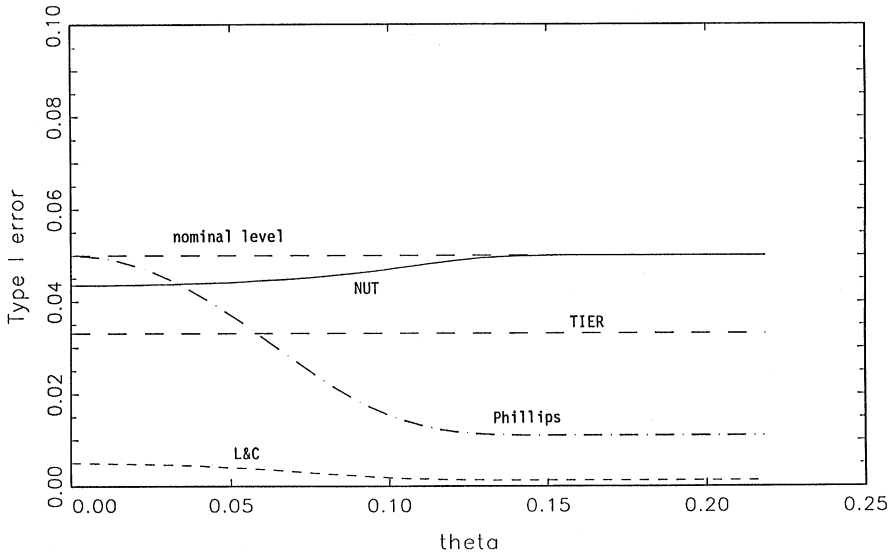

Fig. 2. Type I error on Anderson and Hauck's curve with $n=24$, $v=23$, $p_0=0.8$ and $\Delta=\ln(1.25)\approx 0.22$. Note that the fact that $p_0>0.5$ creates a restriction that $\theta\leqslant\Delta\approx 0.22$.

Fig. 1 gives the boundaries of the rejection regions of the nearly unbiased test and other tests for $n=24$. TIER cannot be plotted since it depends on the data in a more complicated way. Note that the nearly unbiased test has a rejection region containing Liu and Chow's and hence is uniformly more powerful. However, the rejection regions of the nearly unbiased test and Phillips' test do not contain each other. Hence, one cannot draw the conclusion that one uniformly dominates the other in power. A similar picture was drawn for $n=48$, although it is not reported here.

Fig. 2 reports type I error of the four tests on Anderson and Hauck's curve. Since for a fixed $p_0=0.8$, the point (θ,σ) of the curve is uniquely determined by θ, the pictures are plotted against θ. Note that both TIER and Liu and Chow's test have low type I error. For TIER, it is due to discreteness; for Liu and Chow's test, it is due to the shrinking of the alternative region to a triangle region. Phillips' test, although having a correct α-level at $\theta=0$, has small type I error for $|\theta|$ even slightly larger than 0.05. The nearly unbiased test, however, has a type I error much closer to the nominal level 0.05. It is always bounded by 0.05 with the supremum equal to 0.05. Therefore, the test has size 0.05 by Lemma 1 and Theorem 1. The minimal type I error 0.043 attained at $\theta=0$, is close to the nominal level too. Hence Lemma 1 implies that the test is nearly unbiased, justifying the name of our proposed test. As demonstrated below, the closer the type I error is to the nominal level, the more powerful the test will be. As a technical note, calculations leading to Fig. 2 and all other figures of this paper are based on Simpson's method of numerical integration in Gauss. We also calculate the type I error for many other choices of n, p_0 and α including the combinations of

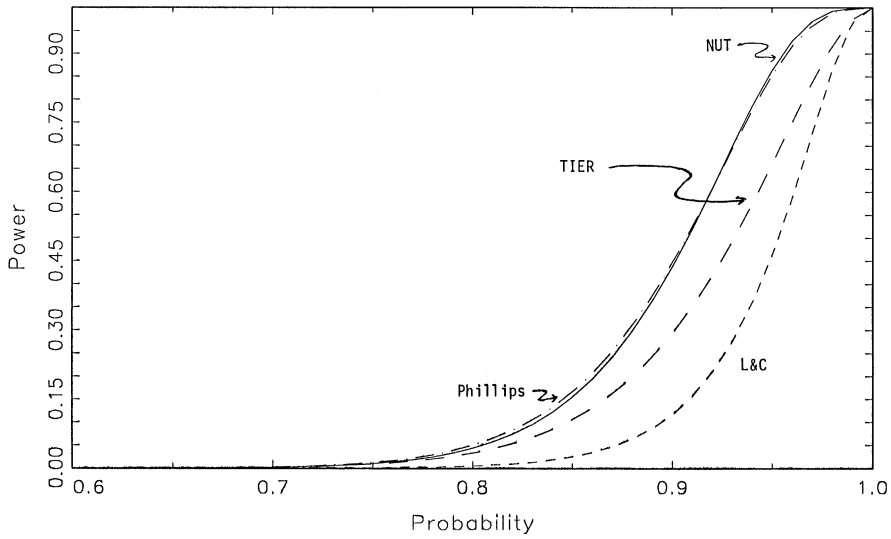

Fig. 3. Power functions of the nearly unbiased test (NUT), Phillips' test, TIER and Liu and Chow's test (L&C) with $n = 24$, $v = 23$, $p_0 = 0.8$, and $\theta = 0$.

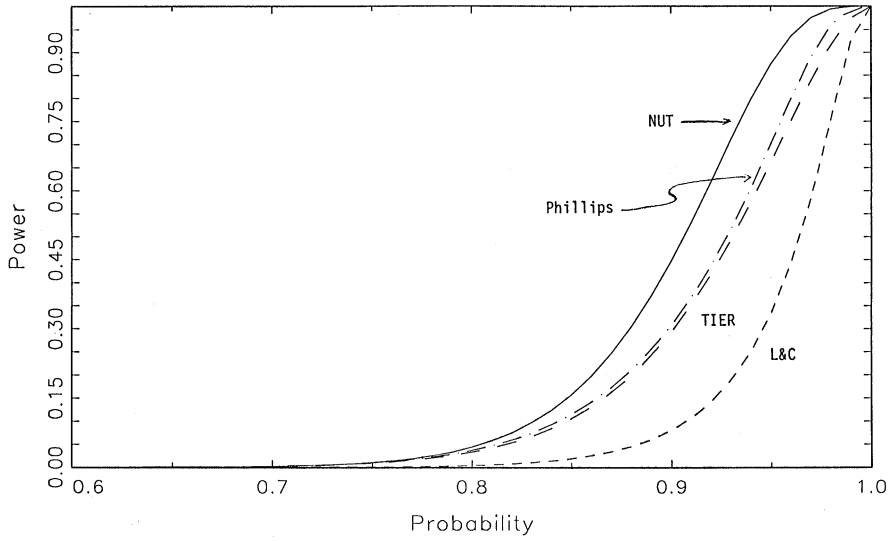

Fig. 4. Power functions of the nearly unbiased test (NUT), Phillips' test, TIER and Liu and Chow's test (L&C) with $n = 24$, $v = 23$, $p_0 = 0.8$, and $\theta = 0.05$.

$n = 16, 20, 24, 28, 32, 48$, $p_0 = \frac{2}{3}, \frac{3}{4}, 0.8$, and $\alpha = 0.05$ and 0.1. All results show that the proposed test is valid and is nearly unbiased.

Figs. 3 and 4 report the power of various tests, plotting against the alternative p, defined in (2.12) for a fixed θ. As is demonstrated, the power of our proposed test is

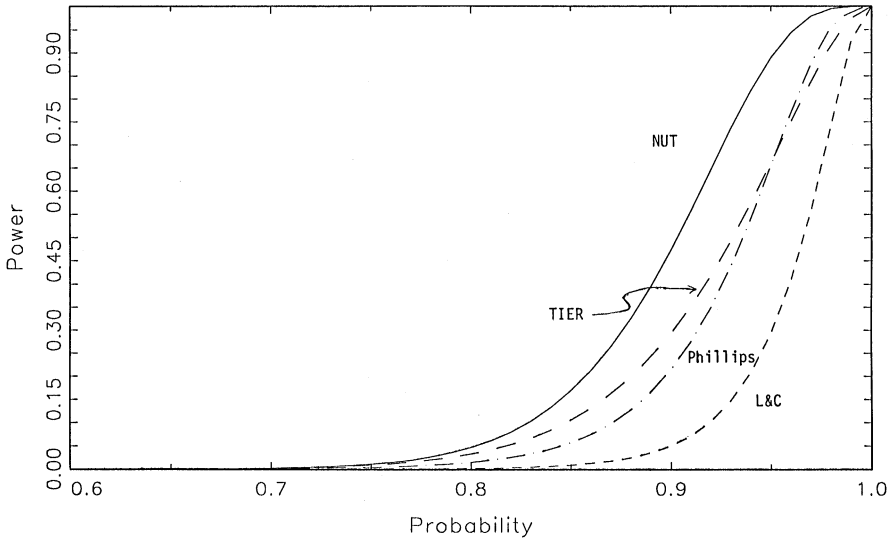

Fig. 5. Power functions of the nearly unbiased test (NUT), Phillips' test, Liu and Chow's test (L&C), and TIER when $n = 24$, $v = 23$, $p_0 = 0.8$, and $\theta = 0.1$.

much higher than TIER and Chow and Liu's test in both cases. More specifically, in Fig. 3, when $\theta = 0$, the powers at $p = 0.95$ of the nearly unbiased test, TIER, and Liu and Chow's test, all with nominal level 0.05, are 0.86, 0.66, and 0.46, respectively. Phillips' test has a power similar to our test for $\theta = 0$. Since $p_0 = 0.8$, the power at $p = 0.8$ is the type I error which for the nearly unbiased test and Phillips' test is close to 0.05. In contrast, the type I errors of the other two tests are lower, which apparently result in their low power.

It is somewhat unexpected that Liu and Chow's test is less powerful than TIER. This, however, is due to the fact that the former has a type I error much smaller than TIER. See Fig. 2. In some cases though, Liu and Chow's test does have a larger power than TIER. The figure corresponding to $n = 21$ and $\theta = 0$, which we have plotted but did not report, shows that Liu and Chow's test has a higher power than TIER if and only if p is sufficiently large. Although Phillips' test performs practically the same as our test when $\theta = 0$, our test has a substantially higher power when $\theta = 0.05$, as shown in Fig. 4. For $p = 0.95$, the powers of the nearly unbiased test, Phillips' test, TIER and Liu and Chow's test are 0.87, 0.7, 0.68 and 0.33 approximately. Note apparently the reason that the nearly unbiased test is more powerful than Phillips' test is that the type I error of the nearly unbiased test is higher than that of Phillips' test for most of $|\theta|$, especially when $|\theta|$ is large. See Fig. 2. Fig. 2 also shows that when $\theta = 0.1$, the type I error of Phillips' test is low, 0.015. Consequently, it is not surprising that the power of Phillips' test is low, sometimes lower than TIER as shown in Fig. 5. Using the same reasoning and judging from Figs. 2 and 7, it is expected that the nearly unbiased test will be most powerful for most θ among all the tests considered in this paper, especially for large $|\theta|$. This indeed turns out to be so in Figs. 3–6.

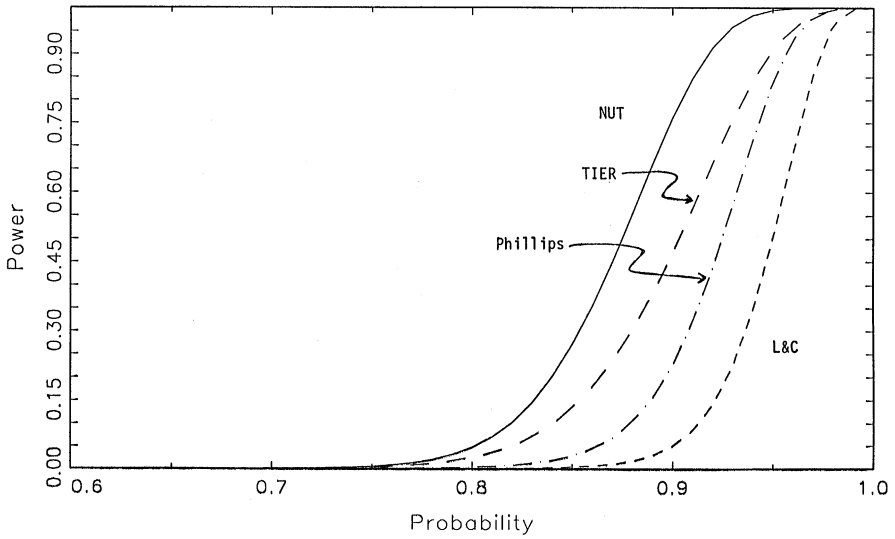

Fig. 6. Power functions of the nearly unbiased test (NUT), Phillips' test, TIER and Liu and Chow's test (L&C) with $n = 48$, $v = 47$, $p_0 = 0.8$, and $\theta = 0.05$.

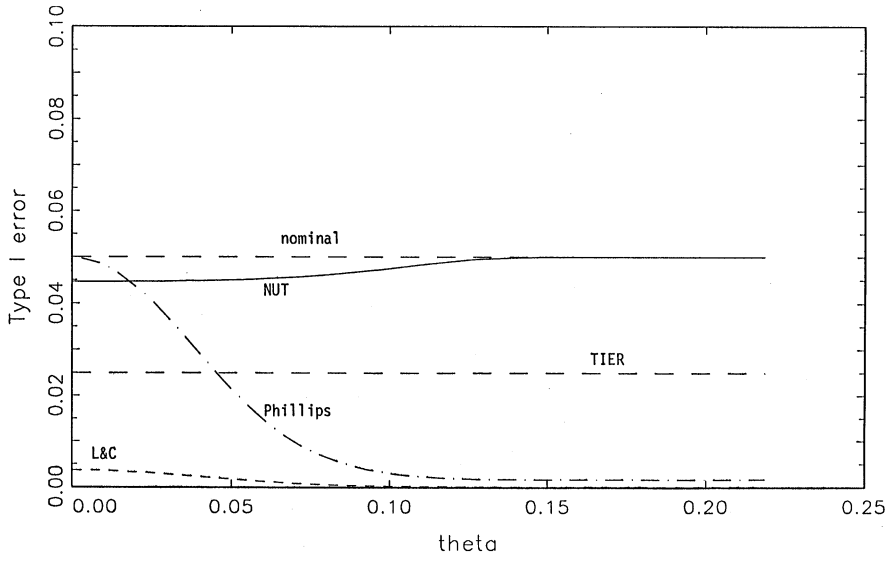

Fig. 7. Type I error on Anderson and Hauck's curve with $n = 48$, $v = 47$, $p_0 = 0.8$ and $\Delta = \ln 1.25 \approx 0.22$. As in Fig. 2, θ is bounded by $\Delta \approx 0.22$.

What about a larger n? Fig. 6 demonstrates similarly that, for $n = 48$, $\theta = 0.05$, the nearly unbiased test substantially improves upon all other tests. When $p = 0.95$, its power is more than 0.99, whereas the power of the second most powerful test, TIER, is 0.91. Fig. 6 also shows that for $\theta = 0.05$, both Phillips' test and Liu and Chow's

test have lower power than TIER. This can be explained by Fig. 7, where it is shown that, at $\theta = 0.05$, Phillips's test and Liu and Chow's test have much smaller type I error than TIER. Also as a referee points out that as n increases from 24 to 48, type I errors of TIER, Phillips and Liu and Chow are getting smaller. This is perhaps due to asymptotic inconsistency of these procedures.

Our test, however, behaves well asymptotically, as stated below. This is due to the fact that our nonlinear rejection region approaches the correct alternative space.

Theorem 3 (Consistency). *Assume that* $0 < \alpha < 1$, $0 < p_0 < 1$. *Then as* $r \to 0$ *and* $v \to \infty$, *the power of the nearly unbiased test* (3.7) *and* (3.9) *approaches one as long as* (θ, σ) *is in the alternative space of* (2.3).

Proof. The proof is omitted.

Note that in Example 2.1, $r = 1/n^{1/2}$ and $v = n - 1$. Hence in this example and all the other examples we know the assumptions $r \to 0$ and $v \to \infty$ are equivalent to the typical asymptotic assumption $n \to \infty$.

4. Other more sophisticated designs

Our test applies to other designs involving more periods. It applies to virtually any design in statistics, so long as the design involves two treatments, so that the problem makes sense. We shall, however, demonstrate the applicability in the next three examples, two of which are the designs referred to in FDA (1997).

Example 4.1 (2×2 *crossover design with period effects*). The model assumed is

$$Y_{ijk} = \mu + F_{(j,k)} + p_j + S_{ik} + \epsilon_{ijk}, \tag{4.1}$$

where Y_{ijk}, $i = 1, \ldots, n_k$, $k = 1, 2$ and $j = 1, 2$, represent the response of the ith subject in the kth sequence for the jth period. In the first sequence, $k = 1$, n_1 subjects are applied to the test drug (T) and after a washout period are applied to the reference drug (R). Conversely for the second sequence, RT is applied to n_2 subjects.

The formulation effects $F_{(j,k)}$ are assumed to depend on the drug received and not on the sequence or the period and hence

$$F_{(1,1)} = F_{(2,2)} = F_{\mathrm{T}} \quad \text{and} \quad F_{(1,2)} = F_{(2,1)} = F_{\mathrm{R}}$$

and without loss of generality $F_{\mathrm{T}} + F_{\mathrm{R}} = 0$. The notation used here seems somewhat different from (2.5), but it generalizes model (2.5) as well as (2.7) by adding the period effects. The notation here is more consistent with the next examples.

It is assumed that S_{ik}, ϵ_{i11}, ϵ_{i22}, ϵ_{i12}, and ϵ_{i21} are all independently normally distributed with mean zero and variances σ_S^2, σ_{T}^2, σ_{T}^2, σ_{R}^2, and σ_{R}^2, respectively. Let

$Y_{ik} = Y_{i1k} - Y_{i2k}$. The sufficient statistics are

$$\bar{Y}_{\cdot 1} = \frac{1}{n_1}\sum_{1}^{n_1} Y_{i1} \quad \text{and} \quad \bar{Y}_{\cdot 2} = \frac{1}{n_2}\sum_{1}^{n_2} Y_{i2}$$

and

$$\hat{\sigma}^2 = \sum_{k=1}^{2}\sum_{i=1}^{n_k}(Y_{ik} - \bar{Y}_{\cdot k})^2/(n_1 + n_2 - 2).$$

Also, let

$$Y = \frac{\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}}{2}$$

which does not depend on the period effects. It seems reasonable that the hypothesis should involve neither period effects nor subject effects. Hence consider testing

$$\mathrm{H}_0\colon\ p_G \leqslant p_0 \text{ and } \mathrm{H}_A\colon\ p_G > p_0, \text{ where } p_G = P(|F_{\mathrm{T}} + \epsilon_{i\mathrm{T}k} - (F_{\mathrm{R}} + \epsilon_{i\mathrm{R}k})| < \Delta).$$

Note that the difference inside the $P(\cdot)$ would equal $Y_{i\mathrm{T}k} - Y_{i\mathrm{R}k}$, if the period effects were missing. Now,

$$p_G = \Phi\left(\frac{\Delta - \theta}{\sigma}\right) - \Phi\left(-\frac{\Delta + \theta}{\sigma}\right),$$

with $\theta = F_{\mathrm{T}} - F_{\mathrm{R}}$ and $\sigma^2 = \sigma_{\mathrm{R}}^2 + \sigma_{\mathrm{T}}^2$. Obviously, Y and $\hat{\sigma}^2$ satisfy the canonical form (3.1) with $\tau^2 = (1/n_1 + 1/n_2)\sigma^2/4$ and $v = n_1 + n_2 - 2$, and hence $r = \tau/\sigma = \frac{1}{2}(1/n_1 + 1/n_2)^{1/2}$. Our proposed test is then (3.7) and (3.9).

Example 4.2 (*Two sequence dual crossover design* (*TRT,RTR*)). This case involves two sequences and three periods. In the first sequence, n_1 subjects go through TRT, i.e., applying the treatment drug, waiting for the washout period, applying the reference drug, waiting for the washout period before the final application of the test drug. Similarly, in the second sequence, RTR are applied to n_2 subjects. This design is one of the two designs suggested in FDA (1997). The observed response at the kth sequence, jth period and the ith subject satisfies

$$Y_{ijk} = \mu + F_{(j,k)} + p_j + S_{ik} + \epsilon_{ijk}, \tag{4.2}$$

a model similar to (4.1) except that there are three periods corresponding to $j = 1, 2$ and 3. Furthermore,

$$F_{12} = F_{21} = F_{32} = F_{\mathrm{R}},$$
$$F_{11} = F_{22} = F_{31} = F_{\mathrm{T}}$$

and $F_{\mathrm{R}} + F_{\mathrm{T}} = 0$; hence, the drug effects do not depend on the sequence or the period.

Define

$$D_{ik} = \begin{cases} \frac{1}{2}(Y_{i11} + Y_{i31}) - Y_{i21} & \text{for } k = 1, \\ Y_{i22} - \frac{1}{2}(Y_{i12} + Y_{i32}) & \text{for } k = 2. \end{cases}$$

Let σ_R^2 and σ_T^2 be the variance of ϵ corresponding to the reference and the treatment. We shall assume that $\sigma_R^2 = \sigma_T^2$, which is denoted by σ_ϵ^2. To reduce this model to the canonical form (3.1), let $D_{\cdot k}$ denote the average of D_{ik} for a fixed k, and

$$Y = \tfrac{1}{2}(D_{\cdot 1} + D_{\cdot 2}). \tag{4.3}$$

Under the normal assumption of ϵ's with equal variances, (4.3) has a normal distribution $N(\theta, \tau^2)$ with

$$\theta = F_T - F_R, \quad \tau^2 = r^2\sigma^2$$

and $r = (\frac{3}{16}(1/n_1 + 1/n_2))^{1/2}$, where $\sigma^2 = 2\sigma_\epsilon^2$. Let $\hat{\sigma}_\epsilon^2$ be the mean square error using model (4.2) with fixed S_{ik} and $\hat{\sigma}^2 = 2\hat{\sigma}_\epsilon^2$. Hence $v\hat{\sigma}^2/\sigma^2 \sim \chi_v^2$ with $v = 2(n_1 + n_2) - 3$. We can then apply our procedure (3.7) and (3.9) to conduct the test for (2.13) with σ^2 being the variance of the difference between the observations corresponding to the treatment and the reference after taking away the period effects.

Example 4.3 (*Four-period crossover design* (*TRTR, RTRT*)). Similarly, we can consider the four-period crossover region (TRTR, RTRT). Here the observations Y_{ijk} are similar to Example 4.2, except that instead of three there are four periods corresponding to $j = 1, 2, 3, 4$ and also the order of treatments are arranged according to TRTR or RTRT for sequences 1 and 2, respectively. This is the recommended design in FDA (1997). Consequently, it is assumed that

$$F_{12} = F_{21} = F_{32} = F_{41} = F_R,$$
$$F_{11} = F_{22} = F_{31} = F_{42} = F_T$$

and $F_R + F_T = 0$.

To reduce to the canonical form (3.1), we let

$$G_{ik} = Y_{i1k} - Y_{i2k}, \quad H_{ik} = Y_{i3k} - Y_{i4k}$$

for $1 \leqslant i \leqslant n_k$ and $k = 1, 2$. Then

$$Y = \tfrac{1}{4}(G_{\cdot 1} - G_{\cdot 2} + H_{\cdot 1} - H_{\cdot 2}) \sim N(\theta, \tau^2)$$

where $\theta = F_T - F_R$, $\tau^2 = r^2\sigma^2$, $\sigma^2 = \sigma_R^2 + \sigma_T^2$ and $r = (1/n_1 + 1/n_2)^{1/2}/8^{1/2}$.

If equal variance $\sigma_R^2 = \sigma_T^2 = \sigma_\epsilon^2$ is assumed, let $\hat{\sigma}_\epsilon^2$ be the mean square error under model (4.2) but with four periods and with subject effects being held fixed. Then Y and $\hat{\sigma}^2 = 2\hat{\sigma}_\epsilon^2$ satisfy (3.1) with $v = 3(n_1 + n_2) - 4$ degrees of freedom. We may then use (3.7) and (3.9) to conduct a test for (2.13). The sample size comparisons of several tests are provided in the next section.

If the assumption of $\sigma_R^2 = \sigma_T^2$ is not made, define

$$\hat{\sigma}^2 = \sum_{k=1}^{2} \sum_{i=1}^{n_k} [(G_{ik} - G_{\cdot k})^2 + (H_{ik} - H_{\cdot k})^2]/[2(n_1 + n_2) - 4].$$

Then Y and $\hat{\sigma}^2$ satisfy (3.1). However, the price to pay is that the degrees of freedom v drop to $2(n_1 + n_2) - 4$.

Table 1
Total number of subjects, n, required for the nearly unbiased test with respect to designs described in Examples 4.1–4.3 with $\alpha = 0.05$

Power at p	p_0	p	2×2	2×3	2×4
0.8	2/3	0.90	16	10	8
		0.95	10	6	6
	3/4	0.90	30	20	14
		0.95	16	10	8
0.9	2/3	0.90	20	12	8
		0.95	12	8	6
	3/4	0.90	40	24	16
		0.95	20	12	8
Degrees of freedom			$n-2$	$2n-3$	$3n-4$

5. Sample sizes

It seems useful to calculate n, the total number of subjects needed to achieve a specific power. Below the number of subjects are taken to be even so that one can assign randomly half of the number of subjects to R or T. Table 1 below gives the sample sizes where 2×2, 2×3 and 2×4 refers to models assumed in Examples 4.1–4.3 where σ_R and σ_T are assumed to be equal. In comparing these n's with those of Liu and Chow (1997, Table 2) and Anderson and Hauck (1990, Table III), we see that the number of subjects needed using our test is always much less and is typically two-thirds or less.

Appendix

Proof of Theorem 1. By symmetry of (3.2), it suffices to consider $\theta > 0$.

First of all, we shall derive an approximate expression for Anderson and Hauck's curve as $\sigma \to 0$. Below we assume that (θ, σ) is on the curve. Hence,

$$\Phi\left(\frac{\Delta - \theta}{\sigma}\right) - \Phi\left(-\frac{\Delta + \theta}{\sigma}\right) = p_0.$$

As $\sigma \to 0$, $-(\Delta + \theta)/\sigma \to -\infty$, and hence the last displayed equation equals approximately

$$\Phi\left(\frac{\Delta - \theta}{\sigma}\right) = p_0,$$

which gives

$$\frac{\Delta - \theta}{\sigma} \to z_{p_0}, \quad \text{i.e., } \theta \approx \Delta - \sigma z_{p_0} \tag{A.1}$$

where $z_{p_0} = \Phi^{-1}(p_0)$.

We shall prove the "only if" part first under (3.5). Let Z be a standard normal random variable independent of $\hat{\sigma}$. Now probability of (3.2) equals

$$P(-T(\hat{\sigma}) < Y < T(\hat{\sigma})) = P\left(\frac{-T(\hat{\sigma}) - \theta}{\tau} < Z < \frac{T(\hat{\sigma}) - \theta}{\tau}\right) \tag{A.2}$$

which is assumed to approach α as $\sigma \to 0$. Note that as $\sigma \to 0$, $\tau \to 0$ and from (A.1), $\theta \to \Delta$. Furthermore,

$$\frac{-T(\hat{\sigma}) - \theta}{\tau} \leqslant \frac{-\Delta}{\tau} \to -\infty,$$

and hence (A.2) is equivalent to $P(Z < (T(\hat{\sigma}) - \theta)/\tau)$. Since $0 < \alpha < 1$, it follows that condition (i) holds. Let $T'(0)$ denote the quantity on the right-hand side of assumption (ii). Hence the probability approaches

$$P\left(Z < \frac{\Delta + T'(0)\hat{\sigma} - (\Delta - \sigma z_{p_0})}{\tau}\right) = P\left(Z < \frac{\sigma z_{p_0} + T'(0)\hat{\sigma}}{\tau}\right)$$
$$= P\left(\frac{Z - z_{p_0}/r}{\hat{\sigma}/\sigma} < \frac{T'(0)}{r}\right) = \alpha.$$

(Here we ignore the error term. To show that the error term actually approaches zero, we may use the mean value theorem. We also need to introduce a bound $\hat{\sigma} < B\sigma$ and let $B \to \infty$.)

By definition of $t_{\alpha,v}(\eta)$,

$$P\left(\frac{Z - z_{p_0}/r}{(\chi_v^2/v)^{1/2}} < t_{\alpha,v}(-z_{p_0}/r)\right) = \alpha,$$

implying condition (ii).

The "if" part follows similarly to the above argument and its proof is omitted.

Proof of Theorem 2. Let

$$G(Y, \hat{\sigma}) = \Phi\left(\frac{\Delta - Y}{\hat{\sigma}}\right) - \Phi\left(-\frac{\Delta + Y}{\hat{\sigma}}\right).$$

For any fixed $\hat{\sigma}$, note that $G(Y, \hat{\sigma})$ is strictly decreasing in $|Y|$. Hence there exists one nonnegative number, say $T_N(\hat{\sigma})$, such that

$$G(Y, \hat{\sigma}) > K \text{ if and only if } |Y| < T_N(\hat{\sigma}). \tag{A.3}$$

Note that $T_N(\hat{\sigma})$ is uniquely defined, although it has to be taken to be zero for some $\hat{\sigma}$ and K. Now $K \geqslant \frac{1}{2}$ implies that

$$T_N \leqslant \Delta. \tag{A.4}$$

To see that this is true, if $T_N > \Delta$, then $G(\Delta, \hat{\sigma}) > K$ by (A.3). However $G(\Delta, \hat{\sigma}) = \Phi(0) - \Phi(-2\Delta/\hat{\sigma}) < \frac{1}{2} \leqslant K$, a contradiction. To argue that T_N is nonincreasing, we may concentrate on $|Y| < \Delta$ by (A.4). For such Y's, $G(Y, \hat{\sigma})$ is obviously a decreasing function of $\hat{\sigma}$. Hence for a larger $\hat{\sigma}$, (3.8) is harder to be satisfied, implying that the boundary $T_N(\hat{\sigma})$ for $|Y|$ gets smaller.

Suppose that K is chosen as in (3.9). Since

$$\Phi\left(\frac{\Delta - T_N(\hat{\sigma})}{\hat{\sigma}}\right) - \Phi\left(-\frac{\Delta + T_N(\hat{\sigma})}{\hat{\sigma}}\right) = \Phi(-rt_{\alpha,v}(-z_{p_0}(r))$$

for all small $\hat{\sigma}$. Letting $\hat{\sigma}$ go to zero, one may show that conditions (i) and (ii) in Theorem 1 are met.

References

Anderson, S., Hauck, W.W., 1990. Consideration of individual bioequivalence. J. Pharmacokinetics Biopharmaceutics 18, 259–273.

Anderson, S., Hauck, W.W., 1992. Types of bioequivalence and related statistical considerations. Int. J. Clinical Pharmacology, Therapy Toxicology 30, 181–187.

Berger, R., Hsu, J., 1996. Bioequivalence trials, intersection-union tests, and equivalence confidence sets. Statistical Sci. 11, 283–319.

Chow, S.C., Liu, J.P., 1992. Design and Analysis of Bioavailability and Bioequivalence Studies. Marcel Dekker, Inc., New York.

EC-GCP, 1993. Biostatistical methodology in clinical trials in applications for marketing authorization for medical products. CPMP Working Party on Efficiency of Medical Products, Draft Guideline Edition, Brussels.

FDA, 1992. FDA guidance on statistical procedures for bioequivalence studies using a standard two-treatment crossover design. Division of Bioequivalence, Office of Generic Drugs, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Rockville, MD.

FDA, 1999. Draft guidance on Average, Population, and Individual Approaches to Establishing Bioequivalence U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), August 1999, BP.

Hwang, J.T., Wang, W., 1997. The validity of the test of individual equivalence ratios (TIER). Biometrika 84 (4), 893–900.

Liu, J.P., Chow, S.C., 1997. A two one-sided tests procedure for assessment of individual bioequivalence. J. Biopharmaceutical Statist. 7, 49–61.

Liu, J.P., Chow, S.C., 1992. On Assessment of Bioequivalence in Variability of Bioavailability Studies. Commun. Statist. Part A 21, 2591–2607.

Phillips, K.F., 1993. A log-normal model for individual bioequivalence. J. Biopharmaceutical Statist. 3, 185–201.

Schall, R., 1995. Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. Biometrics 51, 615–626.

Schall, R., Luus, H.G., 1993. On population and individual bioequivalence. Statist. Med. 12, 1109–1124.

Schall, R., Williams, R.L., 1996. Towards a practical strategy for assessing individual bioequivalence. J. Pharmacokinetics Biopharmaceutics 24(1) 133–149 (for FDA individual bioequivalence working group).

Sheiner, L.B., 1992. Bioequivalence revisited. Statist. Med. 11, 1777–1788.

Wang, W., 1995. On assessment of bioequivalence. Ph.D. thesis, Cornell University.

Wang, W., 1997. Optimal unbiased tests for equivalence in intra-subject variability. J. Amer. Statist. Assoc. 92, 1163–1170.

Wang, W., 1999. On testing of individual bioequivalence. J. Amer. Statist. Assoc. 94, 880–887.

Wellek, S., 1989. Vorschläge zur Reformulierung der statistischen Definition von Bioäquivalenz, Medizinische Informatik und Statistik, Vol. 71, Springer, Berlin, Heidelberg, pp. 95–99.

Wellek, S., 1993. Basing the analysis of comparative bioavailability trials on an individualized statistical definition of equivalence. Biometrical J. 35 (1), 47–55.