UNIVERSITY OF GÖTEBORG

Department of Statistics

ON TESTS OF EQUIVALENCE

by

Sture Holm & Ulla Dahlbom

ON TESTS OF EQUIVALENCE

Sture Holm & Ulla Dahlbom
Department of Statistics
University of Göteborg
Viktoriagatan 13
S-41125 Göteborg
Sweden

ABSTRACT

We will study here a general method for constructing equivalence tests for problems with onedimensional or multidimensional parameter. In the biometric field, the equivalence tests have been studied by many authors under the name of bioequivalence methods. Our general method is closely related to a method for acceptance sampling in the multiparameter case by Berger (1982) and a bioequivalence test method by Schuirmann (1981) for normal distributions and onedimensional parameter. We combine in a general form the ideas of two-sidedness by Schuirmann (1981) and the ideas for multiparameter handling by Berger (1982). We give a number of parametric and nonparametric examples where the general method is used and we illustrate the methods power properties by simulation results.

1. INTRODUCTION

Equivalence tests means tests with the aim to state equivalence between two or several cases. Exact equality can not be 'proved statistically' with a probabilistic protection against erroneous statements in any reasonable case. Thus a suitable setup, which will be used here, is to define a region of approximate equality and to make a test, with some low prescribed level of significance α, of the composite hypothesis that the parameter combination is outside this region. When rejecting this hypothesis, we can make the statement that the parameter combination is inside the region, having the small error α of making this statement wrongly for any parameter combination outside the region.

A common type of equivalence test situation is comparison of bioavailability in pharmaceutics. A new formulation of a drug is compared with a standard formulation in human subjects. For studying the extent of absorption, the areas under the concentration/time curve are then often the basic statistics in the analysis. Some parametric or nonparametric method must be used for evaluating the area under the concentration curve (AUC) from the measures of concentration at a number of times.

A much used type of design of bioequivalence experiment is a two-period crossover design, with some (even) number n of subjects. For the subjects are formed the bioavailability ratios

$$R_i = AUC_i(new) \ / \ AUC_i(standard) \quad i = 1, 2, 3, ..., n$$

or the bioavailability differences

$$D_i = AUC_i(new) - AUC_i(standard) \quad i = 1, 2, 3, ..., n$$

and the analysis is based on either of these sets. The random variables in the set used are usually supposed to be independent.

Since the AUC values are non-negative, it seems more reasonable to use the AUC ratio statistics R_i $i = 1, 2, 3,...., n$ than to use the AUC difference statistics D_i $i = 1, 2, 3, ..., n$. A natural parameter formulation of bioequivalence for the ratio statistics R_i $i= 1, 2, 3,..., n$, is that the expectation of R_i should be in some interval including 1 or that the expectation of the logarithm of R_i should be in some interval including 0. When the difference statistics D_i is used, a natural parameter formulation of bioequivalence is that the expectation of the difference statistics D_i should be in some neighbourhood of 0. For the case of normal distributions there is a method of bioequivalence testing, which seems to be due to Schuirmann. See the abstract Schuirmann (1981). For testing the hypothesis $H_0 : \mu \leq \theta_1$ or $\mu \geq \theta_2$ against the alternative $H_1 : \theta_1 < \mu < \theta_2$ at the level α, the method means making one-sided tests of the two hypotheses $H_{01} : \mu \leq \theta_1$ and $H_{02} : \mu \geq \theta_2$, each one at the level α. The original hypothesis H_0 is rejected only if both these one-sided hypotheses are rejected. It is known that the level of significance is less than or equal to α.

Methods can be given in terms of confidence intervals or in terms of tests. Se related discussions e.g. in Westlake (1972, 1976, 1979), Hauck and Anderson (1984), Steinijans and Diletti (1985) and Kirkwood (1981). In a response to Kirkwood (1981), Westlake suggested use of a 1-2α confidence interval for making bioequivalence statements, which is equivalent to the test method by Schuirmann (1981).

The use of the type of construction made by Schuirmann is not limited to the case of difference test statistics and normal distributions. The same type of construction can be made for nonparametric test statistics and for all types of bioequivalence formulations for one parameter. Further it is possible to give general methods for construction of equivalence tests valid also for multidimensional parameter cases. This is also related to the work

on multiparameter hypotheses testing and acceptance sampling by Berger (1982).

The aim of this paper is to make a discussion of a general method for constructing equivalence tests. We will prove that the tests constructed with this method have the required level of significance. A number of examples will be given, and the properties will be studied in terms of power functions. We will also discuss the relation of our method to other methods for constructing equivalence tests.

2. A GENERAL CONSTRUCTION

In this section we will give a general construction of equivalence tests with garanteed level of significance. The construction can be used for problems with one or several parameters. The one parameter construction is a special case, but for pedagogical reasons we will treat it sepately before we formulate a theorem for the more general case

THEOREM 1

Let C_1 be the rejection region for a level α test of $H_{01} : \theta \leq \theta_1$ and let C_2 be the rejection for a level α test of $H_{02} : \theta \geq \theta_2$, where $\theta_2 > \theta_1$.

Then a test of $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ with rejection region $C_1 \cap C_2$ has a level of significance $\leq \alpha$.

PROOF: If $\theta \leq \theta_1$ then $P_\theta (C_1 \cap C_2) \leq P_\theta (C_1) \leq \alpha$ and if $\theta \geq \theta_2$ then $P_\theta (C_1 \cap C_2) \leq P_\theta (C_2) \leq \alpha$. $\hspace{2cm}$ Q. E. D.

It should be observed that any type of test could be used in the construction. The test by Schuirmann (1981) is the special case, where a t test is used and the observations are supposed to be normally distributed. But the construction works equally well e.g. for a nonparametric rank test like the Wilcoxon test or for a robust test of Huber type. Also the bootstrap technique could be used for construction of equivalence tests according to the principle given above. In a later section we will discuss properties for different types of one parameter equivalence tests and compare the properties of different tests. The type of construction we have made is however possible to use also for multi-parameter equivalence problems.

THEOREM 2

Suppose that $\Theta_0 \subseteq \Theta$ and $\Theta_\lambda \subseteq \Theta$ for $\lambda \in \Lambda$ are parameter sets such that

$$\bigcup_{\lambda \in \Lambda} \Theta_\lambda = \Theta_0$$

Suppose further that for each λ, C_λ is a critical region for a level α test of $H_\lambda : \theta \in \Theta_\lambda$. Then a test of $H_0 : \theta \in \Theta_0$ with rejection region

$$C_0 = \bigcap_{\lambda \in \Lambda} C_\lambda$$

has size $\leq \alpha$.

PROOF

For each $\theta \in \Theta_0$ there exists at least one λ' such that $\theta \in \Theta_{\lambda'}$ since

$$\bigcup_{\lambda \in \Lambda} \Theta_\lambda = \Theta_0$$

Then $P_\theta (C_0) = P_\theta (\cap C_\lambda) \leq P_\theta (C_{\lambda'}) \leq \alpha.$ Q. E. D.

Like in the one-parameter case, the construction is made such that a rejection in the equivalence test of level α is made, if all the onesided hypotheses in a constructed set of tests are rejected at the level α. In a subsequent section we will consider several examples of the construction.

This type of construction is also used by Berger (1982) for acceptance sampling problems, which have much the same character as equivalence testing problems. He works however essentially with onesided test problems. At the instants when he has equivalence test in some parameters, he does not use the general principle above for the two sides of the hypothesis for those individual parameters. He uses instead a combined test based on a noncentral t distribution in case of normally distributed observations. We will discuss the details later.

It should be observed that the method proposed above is completely general. It is not even neccesary to have a finite number of hypotheses H_λ.

In one of the examples in section 4, there is a natural problem formulation, where we technically have a continuum of hypotheses H_λ, $\lambda \in \Lambda$.

3. CROSSOVER DESIGNS

In bioequivalence testing, the crossover design is a much used experimental design. By using each experimental unit for two treatments A and B we get a within subject comparison of these two treatments. The random allocation of the order of using treatments A and B in subjects gives the chance to eliminate possible additive time effects. It is to be noted however, that if there are time effects, then there is a location difference between the distributions for the means of the differences between the results with treatment B and A in the two cases of order of treatment AB and BA. This means that the error estimate in a normal model should not be taken from a standard deviation of the full set of BA differences directly. It should be taken from a pooled standard deviation using the internal standard deviation estimates within the two groups having the order of treatment AB and BA respectively. Thus in a simple case with onedimensional observations and n experimental units for each of the treatment order cases AB and BA, there are 2(n-1) degrees of freedom in the pooled estimate. In simple normal models this is handled in a standrard way. It is slightly more complicated if we consider nonparametric methods for the tests. We will discuss these problems in more detail later.

4. SOME EXAMPLES

In order to illustrate the general method of construction, we will give here some simple examples of equivalence tests. Some of those examples will be studied further concerning power and other properties in subsequent sections.

EXAMPLE 1   Suppose that we have n subjects with independent observations and for the subject number i we have two independent components X_i and Y_i, which are observations of two types of effects, e.g. in form of differences or ratios. Suppose further that the X_i:s are normally distributed with standard deviation σ_1 and that the Y_i:s are normally distributed with standard deviation σ_2. If the observation pairs are obtained in a crossover design, the number n should be even, and there should be n/2 subjects for each ordering AB and BA of the treatments A and B. If there are time effects present, there are different expectations for the X_i and Y_i differences between treatment B and A for the two orderings AB and BA. After balancing out the time effects we have however estimates of the pure B-A difference parameters, which we denote μ_1 and μ_2. If the equivalence statement we want to possibly make is $\mu_{11} < \mu_1 < \mu_{12}$ and $\mu_{21} < \mu_2 < \mu_{22}$, and if we use the level of significance α , then we can make ordinary t tests of the four preliminary hypotheses

$$H_{11} : \mu_1 \leq \mu_{11} \qquad H_{12} : \mu_1 \geq \mu_{12} \qquad H_{21} : \mu_2 \leq \mu_{21} \qquad H_{22} : \mu_2 \geq \mu_{22},$$

at the same (technical) level α. The equivalence statement will be made only if all four hypotheses are rejected. Then the preliminary hypothesis H_{11} will be rejected if

$$\bar{X} \geq \mu_{11} + t_{1-\alpha} \, S_X / \sqrt{n}$$

where $t_{1-\alpha}$ is the $1 - \alpha$ fractile in the t distribution with n-2 degrees of freedom and S_X is a pooled estimate of the standard deviation σ_X of the X

differences. In the same way the other preliminary hypotheses H_{12}, H_{21} and H_{22} will be rejected if respectively

$$\overline{X} \leq \mu_{12} - t_{1-\alpha} \; S_X / \sqrt{n}$$

$$\overline{Y} \geq \mu_{21} + t_{1-\alpha} \; S_Y / \sqrt{n}$$

$$\overline{Y} \leq \mu_{22} - t_{1-\alpha} \; S_Y / \sqrt{n} \; .$$

Observe that using these preliminary test results for the eqivalence test means that the equivalence statement is made only if the rectangula

$$(\overline{X} - t_{1-\alpha} \; S_X / \sqrt{n} \,, \overline{X} + t_{1-\alpha} \; S_X / \sqrt{n} \,) \times (\overline{Y} - t_{1-\alpha} \; S_X / \sqrt{n} \,, \overline{Y} + t_{1-\alpha} \; S_X / \sqrt{n} \,)$$

lies completely inside the rectangula $(\mu_{11} \,, \mu_{12}) \times (\mu_{21} \,, \mu_{22})$.

A naive simple way of generating an equivalence statement with garanteed level of significance would be to make an ordinary rectangular confidence set for $(\mu_1 \,, \mu_2)$ with confidence coefficient $1-\alpha$, and to check if this rectangula lies completely inside the rectangula $(\mu_{11} \,, \mu_{12}) \times (\mu_{21} \,, \mu_{22})$. This means substituting $t_{1-\alpha}$ in our method by $t_{1-\alpha'}$, where $1-2\alpha' = \sqrt{1-\alpha}$. For instance for $n = 20$ and $\alpha = 0.05$, the t value for our method would be 1.73, while the t value for the naive simple method would be 2.43. This indicates that our method has much higher power. Power properties will be studied further in a following section.

EXAMPLE 2   Suppose that we have observations for n individuals, and that the observation on individual i consists of a pair (X_i , Y_i) of possibly dependent normal random variables with some mean vector (μ_1 , μ_2) and unknown covariance matrix Σ. The pairs corresponding to different individuals are supposed to be independent. We consider here a model without time effects in the crossover design. Again our aim is to get equivalence statements of the type that $\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ lies inside a rectangula (μ_{11} , μ_{12}) $\times$ (μ_{21} , μ_{22}). A test with level α of the preliminary hypothesis can be based on the t statistic

$$t_{11} = (\bar{X} - \mu_1) \sqrt{n} / S_X$$

like in the case with independence within the pairs (X_i , Y_i). See e.g. Morrison (1967).   The critical value is the   1-α fractile in the t distribution with   n-1   degrees of freedom. The other preliminary hypotheses are treated in the same way, and the rule for the equivalence statement is exactly the same as in the independence case. The power properties however depend on the covariance between the X and Y variables. There is a symmetry in the independence case, which is missing in the general case. Also the power of the case with dependent X and Y variables will be studied in a following section.

EXAMPLE 3   Suppose that the two components X_i and Y_i for each of n individuals are independent measurements of two characteristics, where the same measurement method is used in both cases. It might then be reasonable to assume that both components have the same unknown variance σ^2. Suppose further that X_i and Y_i are normally distributed with parameters μ_1 and μ_2 and that the equivalence statement we want to make is of the form $\mu_1^2 + \mu_2^2 < r^2$. In this case the hypothesis to test is

$$H_0 : \mu_1^2 + \mu_2^2 \geq r^2$$

and we could write the hypothesis parameter set as a union of the sets

$$\{ (\mu_1, \mu_2) : \mu_1 \cos \varphi + \mu_2 \sin \varphi \geq r \}$$

for all φ, $0 \leq \varphi < 2\pi$. For each of these sets, the corresponding hypothesis H_φ can be tested by a simple t test. Observe however that there is information on the common variance σ^2 in both the X and Y sample. The estimate of σ^2 is $S^2 = \frac{1}{2}(S_X^2 + S_Y^2)$, and the degrees of freedom are $2(n-1)$. The preliminary hypothesis H_φ is rejected if

$$\bar{X} \cos \varphi + \bar{Y} \sin \varphi \leq r - t_{1-\alpha} S / \sqrt{n}$$

i.e. if the mean point $(\bar{X}, \bar{Y})$ is at a distance from the hypothesis parameter set boundary of at least $t_{1-\alpha} S / \sqrt{n}$. This means that the equivalence statement $\mu_1^2 + \mu_2^2 < r^2$ will be made only if the sphere with centre in the mean point $(\bar{X}, \bar{Y})$ and radius $t_{1-\alpha} S / \sqrt{n}$ lies completely inside the sphere $\mu_1^2 + \mu_2^2 < r^2$.

This method can be compared to the naive method of making a $1-\alpha$ confidence set for (μ_1, μ_2) i form of a sphere, and to see if this falls completely inside $\mu_1^2 + \mu_2^2 < r^2$. In that case the radius of the sphere is $(2 v_{1-\alpha} / n)^{1/2} S$. For instance for 20 observations the radius in our method is $0.387 S$, while the radius in the naive method is $0.570 S$. Power calculations and power comparisons will be made in a following chapter.

EXAMPLE 4   When there are doubts about the normality of the observations we can use some nonparametric method like a one sample Wilcoxon test for the two preliminary tests in each component e.g. in example 1. The statistical model in this case is that the X and Y distributions are symmetric around some points. These symmetry points μ_1 and μ_2 are also the parameters to use in the hypotheses formulations like in example 1. In testing for instance the preliminary hypothesis $H_{11} : \mu_1 \leq \mu_{11}$ we can use the translated X observations $X_i - \mu_{11}$ in an ordinary one sample Wilcoxon test of symmetry around 0 against the alternative of symmetry around a positive value, i.e. a onesided alternative. The other preliminary tests are treated analogously.

If we do not have to consider time effects in a crossover design all preliminary tests are easily handled. But if we have to consider possible time effects the problem becomes a little more complicated. Then we have also to estimate the time effects beside making a nonparametric test for the effect parameter. This can be handled e.g. by estimating the time effect by an Hodges-Lehmann estimate or a median difference estimate, adjusting the series by this estimate and make an ordinary one sample Wilcoxon test for the effect parameter. The distribution for this Wilcoxon test statistic under the hypohesis is then however not the ordinary one.

We are not going to treat this more complicated nonparametric methods in more detail in the present paper, but we have the intention to do it elsewhere.

5. POWER PROPERTIES

In this section we will demonstrate the power properties for the suggested methods by presenting simulated power functions for some examples.

Our first power function simulation is concerned with the methods for normally distributed observations, given in examples 1 and 2 of section 4. In example 1 we supposed that the two types of observations were independent and in example 2 we allowed for dependence. But the method works in the same way in the two cases. In Table 1 on the next page we give the simulated power function for correlations $\rho = 0.0$, 0.5 and -0.5 for the special case of 12 observations with standard deviation 1.0 and equivalence region in form of a square with side 2.0. In the simulation with sample size 10000 we have supposed a model without time effects in the crossover design. Thus the standard deviation is estimated with 11 degrees of freedom. The result would not be changed very much if we had a situation, where we considered time effect, and had an estimate of the standard deviation with 10 degrees of freedom. The same random variables are used with different translations to get the different power function results, which explains the regular behaviour. There is a symmetry around the diagonal which is not used in the simulation. Thus there is a slight random lack of symmetry. The essential effect of dependence is a bigger or smaller rounding off in the corners of the equivalence square. For the same sample size and other standard deviations and size of equivalence square, the results are obtainable from the same simulation, if the ratios between the sides of the equivalence square and the standard deviations are the same.

Table 1. Power of a two-dimensional equivalence test based on t statistics for 12 observations with standard deviation 1.0. The equivalence region is (-2.0,2.0) (-2.0,2.0) and the level of significance 0.05.

Correlation ρ = 0.0

Expectation μ_1	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
Expectation μ_2											
0.0	1.00	1.00	1.00	1.00	0.99	0.94	0.82	0.62	0.38	0.16	0.05
0.2	1.00	1.00	1.00	1.00	0.99	0.94	0.82	0.62	0.38	0.16	0.05
0.4	1.00	1.00	1.00	1.00	0.99	0.94	0.82	0.62	0.38	0.16	0.05
0.6	1.00	1.00	1.00	1.00	0.99	0.94	0.82	0.62	0.38	0.16	0.05
0.8	0.99	0.99	0.99	0.99	0.98	0.93	0.81	0.62	0.38	0.16	0.05
1.0	0.95	0.95	0.95	0.95	0.94	0.89	0.78	0.59	0.37	0.16	0.05
1.2	0.83	0.83	0.83	0.83	0.82	0.78	0.68	0.52	0.32	0.14	0.04
1.4	0.61	0.61	0.61	0.61	0.60	0.58	0.50	0.38	0.24	0.10	0.03
1.6	0.37	0.37	0.37	0.37	0.37	0.35	0.31	0.24	0.15	0.07	0.02
1.8	0.16	0.16	0.16	0.16	0.16	0.15	0.13	0.10	0.06	0.03	0.01
2.0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.03	0.02	0.01	0.00

Correlation ρ = 0.5

Expectation μ_1	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
Expectation μ_2											
0.0	1.00	1.00	1.00	1.00	0.98	0.94	0.82	0.62	0.38	0.17	0.05
0.2	1.00	1.00	1.00	1.00	0.98	0.94	0.82	0.62	0.38	0.17	0.05
0.4	1.00	1.00	1.00	1.00	0.98	0.94	0.82	0.62	0.38	0.17	0.05
0.6	1.00	1.00	1.00	0.99	0.98	0.94	0.82	0.62	0.38	0.17	0.05
0.8	0.99	0.99	0.99	0.98	0.97	0.93	0.82	0.62	0.38	0.17	0.05
1.0	0.95	0.95	0.95	0.94	0.93	0.90	0.80	0.61	0.38	0.17	0.05
1.2	0.83	0.83	0.83	0.83	0.82	0.80	0.72	0.56	0.36	0.16	0.04
1.4	0.61	0.61	0.61	0.61	0.61	0.60	0.56	0.45	0.31	0.15	0.03
1.6	0.37	0.37	0.37	0.37	0.37	0.37	0.35	0.30	0.22	0.11	0.02
1.8	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.14	0.11	0.06	0.01
2.0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.03	0.01

Correlation ρ = -0.5

Expectation μ_1	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
Expectation μ_2											
0.0	1.00	1.00	1.00	1.00	0.99	0.94	0.83	0.63	0.38	0.15	0.04
0.2	1.00	1.00	1.00	1.00	0.99	0.94	0.83	0.63	0.38	0.15	0.04
0.4	1.00	1.00	1.00	1.00	0.99	0.94	0.83	0.63	0.38	0.15	0.04
0.6	1.00	1.00	1.00	1.00	0.99	0.94	0.83	0.62	0.38	0.15	0.04
0.8	0.99	0.99	0.99	0.99	0.98	0.93	0.82	0.61	0.37	0.15	0.04
1.0	0.95	0.95	0.95	0.95	0.93	0.89	0.78	0.58	0.34	0.13	0.03
1.2	0.83	0.83	0.83	0.83	0.82	0.77	0.67	0.48	0.27	0.10	0.02
1.4	0.61	0.61	0.61	0.61	0.60	0.56	0.47	0.33	0.18	0.06	0.01
1.6	0.37	0.37	0.37	0.37	0.36	0.33	0.27	0.18	0.09	0.02	0.00
1.8	0.16	0.16	0.16	0.16	0.16	0.14	0.10	0.06	0.03	0.00	0.00
2.0	0.05	0.05	0.05	0.05	0.05	0.04	0.03	0.02	0.01	0.00	0.00

If the sides of the equivalence square are larger or slightly smaller than the ones used, the shape of the power function in the corners of the equivalence square will be essentially the same. If the size of the equivalence square is much smaller than the one used in the simulation, the power function will be much changed. This occurs for a decrease of about 30 % and more. With such a decrease of the equivalence square also follows a considerable decrease of the maximal obtainable power within the equivalence square. Already for 50 % smaller sides, the small maximal power begins to make the equivalence test worthless in practice. In the following Table 2 is given a simulation result for 50 % smaller sides of the equivalence square in the case of independent observations. In all other respects the simulation is made like the one for $\rho = 0.0$ in Table 1. The maximal power is almost 80 %. It rapidly decreases with further decrease of the size of the equivalence square.

Table 2. Power of a two-dimensional equivalence test based on t statistics for 12 observations with standard deviation 1.0. The equivalence region is $(-1.0,1.0)\times(-1.0,1.0)$, the level of significance is 0.05 and the correlation between the components is 0.0.

Expectation 1	0.0	0.2	0.4	0.6	0.8	1.0
Expectation 2						
0.0	0.79	0.72	0.55	0.35	0.15	0.04
0.2	0.72	0.65	0.50	0.31	0.13	0.04
0.4	0.45	0.49	0.38	0.24	0.10	0.03
0.6	0.33	0.30	0.24	0.15	0.07	0.02
0.8	0.14	0.13	0.10	0.06	0.03	0.01
1.0	0.05	0.05	0.03	0.02	0.01	0.004

As mentioned in the discussion of the examples in section 4, the simple bioequivalence test obtained by checking if an ordinary confidence rectangula with confidence degree 1 - α falls completely inside the equivalence rectangula, is less efficient than ours. For a comparison we give

simulated values of its power function in Table 3. The simulation is made in the same way as for our method in the case of independent observations. It can be seen that the simple confidence set method has considerably less power than our method.

Table 3. Power of an equivalence test based on the ordinary rectangular confidence set for 12 normally distributed observations with standard deviation 1.0 and correlation 0 between the components. The equivalence region is (2.0,2.0) (2.0,2.0) and the level of significance is 0.05.

Expectation μ_1	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
Expectation μ_2											
0.0	1.00	1.00	1.00	0.97	0.91	0.79	0.58	0.35	0.16	0.05	0.01
0.2	1.00	1.00	1.00	0.97	0.91	0.79	0.58	0.35	0.16	0.05	0.01
0.4	0.99	0.99	0.99	0.96	0.91	0.78	0.58	0.35	0.16	0.05	0.01
0.6	0.98	0.98	0.98	0.95	0.89	0.77	0.57	0.35	0.16	0.05	0.01
0.8	0.92	0.92	0.92	0.89	0.84	0.72	0.53	0.33	0.15	0.05	0.01
1.0	0.79	0.79	0.79	0.77	0.73	0.62	0.46	0.28	0.13	0.04	0.01
1.2	0.57	0.57	0.57	0.56	0.53	0.46	0.34	0.21	0.10	0.03	0.00
1.4	0.34	0.34	0.34	0.34	0.32	0.27	0.21	0.13	0.06	0.02	0.00
1.6	0.16	0.16	0.16	0.15	0.14	0.12	0.09	0.05	0.03	0.01	0.00
1.8	0.05	0.05	0.05	0.05	0.05	0.04	0.03	0.02	0.01	0.00	0.00
2.0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00

In example 3 of section 4, we had a spherical equivalence set and a method based on an infinite number of preliminary hypotheses. Also in this case there is a similar but less efficient simple method based on an ordinary spherical confidence set. The following Table 4 gives simulated power function values for our method and this simple method. The power functions have a spherical symmetry and they are determined by the radius from the center of the equivalence sphere only. Also here, our method has a considerably higher power than the simple method based disectly on an ordinary confidence set for the twodimensional parameter.

Table 4. Power function for our method and the simple confidence set method for a circular equivalence set with 12 observations on pairs of independent normal random variables with variance 1. The intended level of significance is 0.05, and the number of replicates is 10000.

Radius 2.0

Expectation distance from center	Power for our method	Power for confidence set method
0.0	1.00	1.00
0.2	1.00	1.00
0.4	1.00	0.99
0.6	1.00	0.97
0.8	0.99	0.90
1.0	0.94	0.75
1.2	0.82	0.52
1.4	0.59	0.29
1.6	0.35	0.12
1.8	0.15	0.04
2.0	0.047	0.005

Radius 1.0

Expectation distance from center	Power for our method	Power for confidence set method
0.0	0.76	0.32
0.1	0.75	0.31
0.2	0.70	0.27
0.3	0.60	0.22
0.4	0.50	0.16
0.5	0.40	0.11
0.6	0.29	0.07
0.7	0.19	0.042
0.8	0.12	0.022
0.9	0.070	0.009
1.0	0.036	0.003

6. OTHER METHODS FOR TWO SIDED PROBLEMS

Berger (1982) has used another method than ours for handling the test problem with two-sided hypothesis for the different components of the parameter. Denoting a component parameter by θ_i, the component hypothesis might be written $\theta_i \leq \theta_{i1}$ or $\theta_i \geq \theta_{i2}$. In our method the two subhypotheses would be tested at the same level α. For the case when the observations related to this parameter component are normally distributed with expectation θ_i and unknown standard deviation σ_i, Berger (1982) has a method based on the noncentral t distribution. He assumes an upper bound $\sigma_i \leq \sigma_{i0}$ of the unknown standard deviation. The combined hypothesis $\theta_i \leq \theta_{i1}$ or $\theta_i \geq \theta_{i2}$ is rejected if $| T | < a$ for a suitable a, where T is given by $T = (X_i - \theta_{i0}) / (S_i / \sqrt{m})$ and $\theta_{i0} = (\theta_{i1} + \theta_{i2}) / 2$. Under the null hypothesis T has a noncentral t distribution, and a is determined by a table of that distribution. Table 5 gives simulated power function values for our method and the method based on noncentral t distribution. The table shows that if the true standard deviation is closely below the assumed upper bound, the method based on noncentral t is more powerful than ours. If however the true standard deviation is a little below the boundary, the methods have quite similar power functions and if the true standard deviation is considerably smaller than the bound, our method is much more powerful. If the true standard deviation is above the assumed bound, the method based on noncentral t does not keep the level of significance at its prerequired value, while our method does for all values of the true standard deviation σ.

Table 5. Power β_1 of our test and power β_2 of a test based on noncentral t distribution. Intended level of significance 0.05, equivalence interval (-0.05;0.02) and nine observations on normal random variables. There is power function symmetry around the point -0.015. The number of replicates is 10000.

Standard deviation σ	0.03		0.04		0.05	
Expectation μ	β_1	β_2	β_1	β_2	β_1	β_2
-0.015	.849	.350	.496	.350	.260	.369
-0.010	.809	.325	.477	.341	.252	.350
-0.005	.682	.243	.423	.305	.235	.308
0.000	.540	.140	.346	.228	.182	.249
+0.005	.375	.058	.241	.154	.150	.200
+0.010	.230	.029	.161	.088	.114	.141
+0.015	.101	.005	.101	.050	.077	.092
+0.020	.049	.001	.049	.023	.047	.051

Standard deviation σ	0.06		0.07	
Expectation μ	β_1	β_2	β_1	β_2
-0.015	.130	.387	.036	.350
-0.010	.119	.355	.038	.354
-0.005	.111	.343	.037	.333
0.000	.088	.273	.031	.307
+0.005	.062	.239	.023	.270
+0.010	.048	.183	.016	.226
+0.015	.054	.144	.021	.178
+0.020	.036	.106	.010	.140

7. ACKNOWLEDGEMENT

REFERENCES

Berger R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. Technometrics, 24, 295-300.

Hauck W, W. & Andersson S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. Journal of Phamacokinetics and Biopharmaceutics, 12, 83-91.

Kirkwood T. B. L. (1981). Bioequivalence testing - a need to rethink. (With a response by W. J. Westlake.) Biometrics, 37, 589-594.

Morrison D. F. (1967). Multivariate statistical methods. McGraw-Hill.

Schuirmann D. L. (1981). On hypothesis testing to dewtermine if the mean of a normal distribution is contained in a known interval.
Biometrics, 37, 617.

Stenijans V. W. & Dilette E. (1985). Generalization of distributions-free confidence intervals for bioavailability ratios.
Eur. J. Clin. Pharmacol., 28, 85-88.

Westlake W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. J. of Pharmaceutical Sciences, 61, 1340-1341.

Westlake W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. Biometrics, 32, 741-744.

Westlake W. J. (1979). Statistical aspects of comparative bioavailability trials. Biometrics, 35, 273-280.

1990:1   Holm, S.   Abstract bootstrap confidence
                            intervals in linear models