Eur J Clin Pharmacol (1983) 24: 127–136

European Journal of
Clinical Pharmacology

Statistical Analysis of Bioavailability Studies: Parametric and Nonparametric Confidence Intervals

V. W. Steinijans and E. Diletti

Biometry Department, Byk Gulden Research Centre, Konstanz, Federal Republic of Germany

Summary. For a two-way cross-over design, which appears to be the most common experimental design in bioavailability studies, 95%-confidence limits for expected bioavailability can be obtained by classical analysis of variance (ANOVA). If symmetry of the confidence interval is desired about zero (differences) or unity (ratios) rather than about the corresponding point estimator, Westlake's modification can be used. Two nonparametric methods and their adaptations to bioavailability ratios are reviewed, one based on Wilcoxon's signed rank test (Tukey), and the other on Pitman's permutation test. The necessary assumptions and the merits of these procedures are discussed. The methods are illustrated by an example of a comparative bioavailability study. A FORTRAN program facilitating the procedures is available from the authors upon request.

Key words: statistical analysis, nonparametric statistical methods; bioavailability, confidence interval, ANOVA

It is widely accepted among biometricians that establishing a confidence interval rather than hypothesis testing is appropriate to the statistical analysis of bioavailability studies (Metzler 1974; Westlake 1972 and 1979; Shirley 1976). In practice, however, publications on comparative bioavailability studies still use classical hypothesis testing for comparison of two or more formulations (Kramer et al. 1977; Rietbrock et al. 1979). Frequently, bioequivalence is accepted if a *t*-test or analysis of variance does not show a significant difference between the formulations. Such a conclusion may be erroneous for various reasons; variability in the physical characteristics of the test product or lack of analytical precision and/or heterogeneity of the group studied may result in overall variability which does not permit products to be distinguished for the chosen sample size. In consequence, products with poor mean bioavailability might pass bioequivalence tests merely because they possess an additional undesirable property – variability (Upton et al. 1980). The lack of power to distinguish between different formulations is characterized by the Type II error. This in turn depends on the residual variance, which should be kept as small as possible (Fluehler et al. 1981). Further, even if so-called parametric confidence intervals are calculated, neither the assumption of a normal distribution of the residual errors common to all subjects, formulations and periods, nor the assumption of the additivity of these factors may be valid. Under these circumstances, the use of so-called nonparametric methods is preferable (Koch 1972; Abt 1977; Royen 1978; Steinijans 1981), i.e. statistical methods which are not dependent on the normal distribution and additional, rather specific assumptions.

It is the objective of this paper to review various statistical procedures for obtaining confidence intervals for a bioavailability characteristic. Discussion is restricted to the two-way cross-over design, which appears to be the commonest design in bioavailability studies (Fluehler et al. 1981). The assumptions of the procedures presented are stated and the appropriate equations for the calculation of 95%-confidence limits are given. In addition, an interactive computer program facilitating all of these procedures has been written in FORTRAN FTN4X (HP 1000) and is available from the authors upon request.

Bioavailability Characteristics

Bioavailability encompasses the rate and extent of drug absorption. The area under the concentration/time curve (AUC) and urinary recovery account for

0031-6970/83/0024/0127/$02.00

Table 1. Two-way cross-over design with a total of $n = 2K$ subjects, i.e. K subjects in each sequence. The response of the kth subject in the period j within the sequence i is denoted by y_{ijk} ($i = 1, 2; j = 1, 2; k = 1, \ldots, K$)

	Period 1	Period 2
Sequence 1	Formulation 1 (Reference) y_{11k} ($k = 1, \ldots, K$)	Formulation 2 (Test) y_{12k} ($k = 1, \ldots, K$)
Sequence 2	Formulation 2 (Test) y_{21k} ($k = 1, \ldots, K$)	Formulation 1 (Reference) y_{22k} ($k = 1, \ldots, K$)

the extent of absorption, while the peak plasma level and the time taken to reach it may characterize the rate of absorption.

Sample size determination and statistical evaluation are usually carried out separately for each characteristic (univariate approach). Vila et al. (1980) considered the three statistical moments AUC, MRT (mean residence time), and VRT (variance of residence time) as a multivariate bioavailability characteristic.

Experimental Design

Bioavailability studies usually follow a balanced, repeated-measurement design, such as a cross-over or a series of Latin squares. The commonly used two-period change-over design is depicted in Table 1.

The following effects are included in the classical ANOVA-model: sequence and period of administration, formulations and subjects (Wallenstein and Fisher 1977; Selwyn et al. 1981). One of the major problems in clinical trials is carry-over effect, which may in addition be confounded with direct effects (Grizzle 1965 and 1974). Carry-over effects, however, are of almost no concern in bioavailability studies. The half-life of a drug is usually known from previous studies, and a washout period of 5 to 6 half-lives should ensure that no measurable residual drug is carried over from one period to the next. This must be verified by taking a blood sample prior to the administration of the second formulation. Possible pharmacodynamic and/or metabolically induced carry-over effects can usually be ruled out after acute administration.

Sample Size Determination

Sample size determination depends on the experimental design and the procedure chosen for the statistical analysis. Hence, if ANOVA is to be used, some a priori knowledge of the residual variance must be available. In addition to the classical approach (Ostle 1966), the nomograms of Fluehler et al. (1981) give for 6 and 12 subjects, respectively, the posterior probability of bioequivalence as a function of the relative difference in formulation means and the coefficient of variation.

Confidence Intervals

Confidence Interval Based on Analysis of Variance (ANOVA)

We assume that bioavailability data have been collected for 2 formulations, a test and a reference preparation, according to a two-way cross-over design, as shown in Table 1. Half the total of n subjects, $K = n/2$, first receive the reference formulation and then, after a suitable washout period, the test formulation (Sequence 1). The remaining $K = n/2$ subjects receive the 2 formulations in reverse order (Sequence 2). The allocation of the subjects to the sequences is random.

If we assume that no carry-over effect exists, then for a particular bioavailability characteristic y, e.g. AUC, the response for the kth subject in the sequence i during the period j can be modelled as follows (Selwyn et al. 1981):

$$y_{ijk} = \mu + \pi_j + \tau_l + s_{ik} + e_{ijk} \quad (i = 1,2; \quad j = 1,2; \quad l = 1,2; \quad k = 1,\ldots,K), \tag{1}$$

where π_j and τ_l are fixed effects associated with period and formulation (treatment). For the sake of simplicity, it is assumed that $\tau_1 = -\tau_2$ and $\pi_1 = -\pi_2$, so that μ becomes the overall mean. s_{ik} is associated with the random subject effect, and e_{ijk} denotes the random error term. It is assumed that the $\{s_{ik}\}$ and $\{e_{ijk}\}$ are all independently and normally distributed with means 0 and variances σ_s^2 and σ_e^2, respectively. The assumptions made about s_{ik} and e_{ijk} imply that each observation y_{ijk} has the variance $\sigma_s^2 + \sigma_e^2$, and that the two observations on the same individual have the covariance σ_s^2 and hence the correlation $\sigma_s^2/(\sigma_s^2 + \sigma_e^2)$. Observations made on different subjects are independent. The analysis of variance is given by Selwyn et al. (1981; p 13, Table 1).

The expected means of the two formulations are denoted by μ_l ($l = 1, 2$). The two-sided 95%-confidence limits for the expected mean difference $\delta = \mu_2 - \mu_1 = \tau_2 - \tau_1$ are calculated as follows:

$$d.. \pm t(n-2; 0.975)\ \sqrt{2\ \mathrm{MSE}/n}, \tag{2}$$

where $\bar{d}.. = (\bar{y}_{12.} + \bar{y}_{21.})/2 - (\bar{y}_{11.} + \bar{y}_{22.})/2$ is the estimated mean difference between Formulation 2 (Test) and Formulation 1 (Reference). MSE is the corresponding mean square for error obtained in the ANOVA, and $t(n-2;\ 0.975)$ denotes the 97.5-percentile of the t-distribution with $n-2$ degrees of freedom, e.g. $t(10;\ 0.975) = 2.2281$.

The 95%-confidence limits for the expected mean difference between the two formulations are then expressed as a fraction of the mean for the reference formulation. In order to do so, the unknown expected mean of the reference formulation has to be approximated by its sample mean $\bar{y}_{\text{Reference}} = \bar{y}_{\text{Formulation 1}} = (\bar{y}_{11.} + \bar{y}_{22.})/2$ (Westlake 1972). The ratio $\bar{y}_{\text{Test}}/\bar{y}_{\text{Reference}} = \bar{y}_{\text{Formulation 2}}/\bar{y}_{\text{Formulation 1}}$ serves as a point estimator of the bioavailability ratio μ_2/μ_1, and

$$(\bar{y}_{\text{Test}} \pm t(n-2;\ 0.975)\ \sqrt{2\ \text{MSE}/n})/\bar{y}_{\text{Reference}} \qquad (3)$$

serves as its approximate 95%-confidence limits.

If, for example, log AUC rather than AUC is chosen as the bioavailability characteristic (Metzler 1974; Westlake 1973 and 1976; Steinijans et al. 1982), then Expression (2) yields 95%-confidence limits of the logarithm of the expected bioavailability ratio. Taking antilogs thus provides 95%-confidence limits of the bioavailability ratio itself, a point estimator of which is the geometric mean of the individual ratios $\text{AUC}_{\text{Test}}/\text{AUC}_{\text{Reference}}$.

The presentation given so far assumes that $K = n/2$ subjects are randomly assigned to each of the Sequences 1 and 2. In the case of dropouts, unequal numbers of subjects in the two sequences may occur, designated by n_1 and n_2, respectively. This leads to modifications, which were presented by Grizzle (1965 and 1974). As pointed out by Abt (1977), the expected mean square for periods then depends on $(n_1 - n_2)(\tau_1 - \tau_2)$. Hence, for $n_1 \neq n_2$, it is not possible to test the null-hypothesis of no period effect. See also Grieve (1982).

Westlake's Modification of the ANOVA-Based Confidence Interval

It is apparent from Expression (2) that the conventional 95%-confidence interval for the expected mean difference is symmetrical about the estimated mean difference $\bar{d}.. = \bar{y}_{\text{Test}} - \bar{y}_{\text{Reference}}$, and not symmetrical about zero. Similarly, Expression (3) shows that the conventional 95%-confidence interval for the bioavailability ratio is symmetrical about the point estimator $\bar{y}_{\text{Test}}/\bar{y}_{\text{Reference}}$, and not about unity. Conventional confidence intervals thus reflect the direction in which the sample difference or ratio has been found.

With particular reference to comparative bioavailability trials, Westlake (1972, 1976 and 1979) shifted the emphasis from estimation to decision making. If the 95%-confidence limits fall within acceptable limits, for example as recommended by a regulatory agency, then the test formulation will be accepted, and if not it will be rejected. As acceptable limits are usually given in a symmetrical form, say 0.8 to 1.2, the use of a confidence interval symmetrical about zero for differences, or about unity for ratios, has been proposed by Westlake. Hence, the conventional 95%-confidence interval for the expected mean difference $\mu_2 - \mu_1$, which is symmetrical about its point estimator $\bar{d}..$, cf. Expression (2), and thereby symmetrical about the underlying t-distribution, is replaced by a 95%-confidence interval symmetrical about zero. This involves the selection of two constants k_1 and k_2, such that Westlake's condition is satisfied:

$$k_1 + k_2 = 2\,(\bar{y}_{\text{Reference}} - \bar{y}_{\text{Test}})/\ \sqrt{2\ \text{MSE}/n} \qquad (4.1)$$

and

$$\int_{k_2}^{k_1} f_{t\,(n-2)}(s)\,ds = 0.95 \qquad (4.2)$$

Equation (4.2) states that the interval (k_2, k_1) includes 95% of the mass of the t-distribution with n-2 degrees of freedom. The 95%-confidence limits for the expected mean difference between test and reference formulation, $\mu_2 - \mu_1$, are then given by

$$\bar{d}.. + k_2\ \sqrt{2\ \text{MSE}/n} \text{ and } \bar{d}.. + k_1\ \sqrt{2\ \text{MSE}/n}, \qquad (5)$$

which is Westlake's analogue of Expression (2). The symmetry of the confidence limits (5) about zero is ensured by the condition stated in Equation (4.1). As before, approximate 95%-confidence limits for the bioavailability ratio μ_2/μ_1 are given by

$$(\bar{y}_{\text{Test}} + k_i\ \sqrt{2\ \text{MSE}/n})/\bar{y}_{\text{Reference}}\ (i = 1, 2), \qquad (6)$$

which is Westlake's analogue of Expression (3).

In order to facilitate the computation of k_1 and k_2 according to (4.2), Spriet and Beiler (1978) provided useful tables with n-2 (degrees of freedom) and $k_1 + k_2$ derived from (4.1) as entries. It should be pointed out that for $\bar{y}_{\text{Reference}} < \bar{y}_{\text{Test}}$, $k_1 + k_2$ becomes negative. In this case, $-(k_1 + k_2)$ serves as the table entry, and the resulting value will be $-k_1$ and not k_2.

Comparison of Expression (5) and (2) shows that in essence only the 95-percentiles of the t-distribution have been changed from the central values

Table 2. Nonparametric 1-α confidence interval of expected bioavailability ratio based on Wilcoxon's signed rank test (Tukey)

Number of subjects	Number of Walsh averages	1-α confidence interval of expected bioavailability ratio		
		Index of ordered geometric Walsh average		
n	$\frac{n(n+1)}{2}$	C_α	$\frac{n(n+1)}{2}+1-C_\alpha$	1-α
6	21	1	21	0.9688
7	28	3	26	0.9531
8	36	4	33	0.9609
9	45	6	40	0.9609
10	55	9	47	0.9512
11	66	11	56	0.9580
12	78	14	65	0.9575
13	91	18	74	0.9521
14	105	22	84	0.9506
15	120	26	95	0.9521
16	136	30	107	0.9557
17	153	35	119	0.9552
18	171	41	131	0.9517
19	190	47	144	0.9506
20	210	53	158	0.9516
21	231	59	173	0.9540
22	253	66	188	0.9538
23	276	74	203	0.9516
24	300	82	219	0.9509

$\pm\ t(n\text{-}2, 0.975)$ to k_2 (<0) and k_1, respectively. From this, it follows immediately that for $\bar{y}_{Test} \neq \bar{y}_{Reference}$, Westlake's 95%-confidence interval is always longer than the conventional 95%-confidence interval. In fact, the confidence coefficient for Westlake's interval is always greater than 95% (Westlake 1976). Only within the class of 95%-confidence intervals symmetrical about zero does Westlake's condition provide the shortest confidence intervals possible.

If log-transformed data (indicated by a subscript log) are used to construct a 95%-confidence interval of the bioavailability ratio μ_2/μ_1, and if this interval is supposed to be symmetrical about unity, then the following condition (Westlake 1976) must hold in addition to (4.2):

$$\text{antilog}(\bar{d}_{log}+k_2\sqrt{2\,MSE_{log}/n}) + \text{antilog}(\bar{d}_{log}+k_1\sqrt{2\,MSE_{log}/n})=2. \quad (4.3)$$

Antilog($\bar{d}_{log}$) is the geometric mean of the individual ratios $y_{Test}/y_{Reference}$, and the 95%-confidence limits for μ_2/μ_1 are obtained by multiplying the geometric mean by the antilogs of $k_2\sqrt{2\,MSE_{log}/n}$ and $k_1\sqrt{2\,MSE_{log}/n}$, respectively. If natural logarithms are used, antilog stands for the exponential function.

Confidence Interval Based on the Paired t-Test

If no period effect is postulated, the full ANOVA model given by Equ. (1) reduces to

$$y_{ijk}=\mu+\tau_l+s_{ik}+e_{ijk}\ (i=1,2;\ l=1,2;\ k=1,\ldots,K). \quad (7)$$

The one degree of freedom previously associated with periods increases the degrees of freedom for error by 1 to n-1.

Intraindividual differences $d_{1k}=y_{12k}-y_{11k}$ ($k=1,\ldots,K$) and $d_{2k}=y_{21k}-y_{22k}$ ($k=1,\ldots,K$) in Sequences 1 and 2, respectively, have the expectation $E[d_{ik}]=\mu_2-\mu_1$ and the variance $V[d_{ik}]=2\sigma_e^2$ ($i=1,2;\ k=1,\ldots,K$).

Estimating $V[d_{ik}]$ by $s_d^2=\frac{1}{n-1}\sum_{i=1}^{2}\sum_{k=1}^{K}(d_{ik}-d_{\cdot\cdot})^2$, and recalling that σ_e^2 is estimated by MSE, the 95%-confidence limits for the expected mean difference in formulations, which are given by Expression (2) for the full ANOVA model, now become:

$$\bar{d}_{\cdot\cdot}\pm t(n-1;0.975)\,s_d/\sqrt{n} \quad (8)$$

This, however, is the well-known expression for two-sided 95%-confidence limits of the expected mean difference based on the paired t-test. As before, 95%-confidence limits for the bioavailability ratio μ_2/μ_1 are given by

$$(\bar{y}_{Test}\pm t(n\text{-}1;0.975)\,s_d/\sqrt{n})/\bar{y}_{Reference}. \quad (9)$$

The procedure for log-transformed data is also straightforward.

Nonparametric Confidence Interval Based on Wilcoxon's Signed Rank Test (Tukey)

Some of the assumptions of the ANOVA model, such as additivity of period, subject and formulation effect, or homogeneity of variances for subjects and residuals, respectively, are neither obvious nor easily verifiable. A statistical method to obtain a confidence interval under less restrictive assumptions is based on Wilcoxon's signed rank test, which is the nonparametric analogue of the paired t-test. Again, we assume that no period effect is present, and the intraindividual differences "test minus reference" are denoted simply by d_i ($i=1,\ldots,n$), irrespective of the sequence of administration.

The model (Hollander and Wolfe 1973; pp 26–33) is

$$d_i=\delta+e_i\ (i=1,\ldots,n), \quad (10)$$

Table 3. Nonparametric 1-α confidence interval of expected bioavailability ratio based on Pitman's permutation test

Number of subjects	Number of averages	1-α confidence interval of expected bioavailability ratio		
		Index of ordered average		
n	2^n-1	K_α	2^n-K_α	1-α
6	63	1	63	0.9688
7	127	3	125	0.9531
8	255	6	250	0.9531
9	511	12	500	0.9531
10	1023	25	999	0.9512
11	2047	51	1997	0.9502
12	4095	102	3994	0.9502

where the e's denote the random error terms, and δ is the expected difference between formulations. The e's are assumed to be mutually independent, each e coming from a continuous distribution which is symmetrical about zero. However, it is no longer assumed that all e's must necessarily come from the same distribution. In other words, the random fluctuation may be different for each subject. This seems to be a realistic assumption, particularly in view of the large inter- and intraindividual variations known for such drugs as theophylline.

Under the above assumptions, a nonparametric confidence interval is obtained as follows. Let

$$a_{ij}=(d_i+d_j)/2\ (i\leq j;\ i,j=1,\ldots,n) \tag{11}$$

denote the $n(n+1)/2$ arithmetic Walsh averages of the d_i $(i=1,\ldots,n)$, and let $a(1),\ldots,a(n(n+1)/2)$ denote their ordered values. For $1-\alpha\geq 0.95$, the $1-\alpha$ confidence interval (δ_L, δ_U) is given by

$$\delta_L=a(C_\alpha),\ \delta_U=a(n(n+1)/2+1-C_\alpha), \tag{12}$$

where $C_\alpha=n(n+1)/2+1-t(\alpha/2;\ n)$, and $t(\alpha/2;\ n)$ is the critical point of the Wilcoxon sum T^+ of positive ranks; for example, $t(0.0425/2;\ 12)=65$, i.e. $P(T^+\geq 65)<0.25$, hence $C_\alpha=14$ and $P(a(14)<\delta<a(65))=0.9575>0.95$. For relevant sample sizes, $6\leq n\leq 24$, the values of C_α and $n(n+1)/2+1-C_\alpha$ are given in Table 2 (cf. Hollander and Wolfe 1973, pp 269–271; Wilcoxon et al. 1973, pp 247–259; Geigy, Wissenschaftliche Tabellen 1980 p 163).

The above result is due to Tukey, who showed that if there are no ties among the d's and none of the d's is zero, then A^+, the number of positive arithmetic Walsh averages, is equal to the positive rank sum T^+ (Hollander and Wolfe 1973, pp 35–39).

The 1-α confidence interval (12) consists of those values δ_0 for which the two-sided α-level signed rank test accepts the hypothesis $\delta=\delta_0$. Under the above assumptions, we can control the coverage probability to be 1-α without knowledge of the underlying error distribution. Thus (δ_L, δ_U) provides a distribution-free confidence interval for δ over a large class of populations.

Hodges and Lehmann proposed the median of the arithmetic Walsh averages as a point estimator of δ (Hollander and Wolfe 1973, pp 33–35). This estimator, say $\hat{\delta}$, is natural insofar as the shifted sample $d_i-\hat{\delta}$ $(i=1,\ldots,n)$ appears (when viewed by the signed rank test statistic T^+) to come from a population with the median 0. In addition, $\hat{\delta}$ will be relatively insensitive to outliers, which is not the case for the arithmetic mean of the d_i $(i=1,\ldots,n)$.

The modification of the Tukey method for ratios instead of differences is done directly by taking the logarithm of the bioavailability characteristic. For example, consider $r_i=AUC_i(Test)/AUC_i(Reference)$ and $d_i=\log r_i=\log AUC_i(Test)-\log AUC_i(Reference)$. Model (10) then becomes

$$r_i=\rho f_i\ (i=1,\ldots,n), \tag{13}$$

with $\delta=\log\rho$ and $e_i=\log f_i$ $(i=1,\ldots,n)$, and the hypothesis $\rho=1$ corresponding to $\delta=0$.

Arithmetic Walsh averages $a_{ij}=(d_i+d_j)/2$ are replaced by geometric Walsh averages $g_{ij}=\sqrt{r_i r_j}$, and hence $a_{ij}=\log g_{ij}$ $(i\leq j;\ i,j=1,\ldots,n)$. Due to the monotonicity of the logarithmic transformation, point estimator and confidence limits, which are order statistics of Walsh averages, are directly transferable to geometric instead of arithmetic Walsh averages, i.e.

$$\hat{\rho}=\text{median}\{g_{ij}: i\leq j;\ i,j=1,\ldots,n\} \tag{14}$$

and

$$\rho_L=g(C_\alpha),\ \rho_U=g(n(n+1)/2+1-C_\alpha). \tag{15}$$

Hence, instead of calculating distribution-free confidence limits for the expected difference in log-transformed AUC's first, and then taking antilogs of these limits, a distribution-free confidence interval of the bioavailability ratio is directly obtainable from geometric Walsh averages. Naturally, in the case "$n(n+1)/2$ even", the median is obtained by geometric instead of arithmetic interpolation.

As the above confidence interval is based on the discrete distribution of T^+, the coverage probability is usually greater than 0.95 (see Table 2). Hence, the confidence interval is somewhat longer than neces-

Table 4. Demographic data and sequence of administration of 12 healthy male volunteers

Subject	Age [years]	Weight [kg]	Height [cm]	Sequence of administration (R = Reference, T = Test)	
1	40	84	175	T	R
2	44	76	179	T	R
3	28	74	173	R	T
4	38	70	180	R	T
5	29	71	189	T	R
6	34	74	168	T	R
7	28	70	189	R	T
8	32	78	178	T	R
9	42	67	168	R	T
10	49	72	172	T	R
11	40	74	180	R	T
12	40	74	182	R	T
Median	39	74	179		
Minimum	28	67	168		
Maximum	49	84	189		

Table 5. Area under the concentration/time curve, AUC, after administration of 385.6 mg theophylline in a sustained release preparation under reference condition (fasted) and test condition (standard breakfast)

Subject	AUC [mg/l h]		
	Reference	Test	Ratio
1	136.0	135.7	1.00
2	152.6	155.3	1.02
3	123.1	148.9	1.21
4	77.0	81.2	1.05
5	115.7	139.2	1.20
6	72.0	91.7	1.27
7	116.4	118.7	1.02
8	151.1	133.2	0.88
9	118.9	115.6	0.97
10	156.1	150.3	0.96
11	222.4	223.9	1.01
12	158.1	154.1	0.97
Geometric mean	127.7	133.1	1.04

sary. To account for this, a simple interpolation has been suggested (Steinijans 1981).

Nonparametric Confidence Interval Based on Pitmans's Permutation Test

The model based on the permutation test is again given by Eq. (10), but with the distinction that the e's need not necessarily come from continuous distributions. It is still assumed that the distributions of the stochastically independent e's are symmetrical about zero.

Under the hypothesis $\delta = 0$, the 2^n permutations of the signs of the observed differences produce a discrete uniform distribution with point probability 2^{-n}. Consequently, the permutation test permits a closer approximation of the exact 95%-level than the signed rank test. For example, if $n = 12$, the actual confidence coefficient for the signed rank test is 0.9575 (cf. Table 2), whereas that for the permutation test is 0.9502 (cf. Table 3).

The 1-α confidence interval is derived from order statistics of the 2^n-1 averages of up to all n differences observed (Royen 1978). More precisely, let $\{i_1, \ldots, i_M\}$ denote any nonempty subset of the index set $\{1, \ldots, n\}$, and let A denote the set of corresponding arithmetic averages of observed differences d_i $(i = 1, \ldots, n)$:

$$A = \left\{ \frac{1}{M} \sum_{m=1}^{M} d_{i_m} : \{i_1, \ldots, i_M\} \subset \{1, \ldots, n\} \right\}.$$

Let $a(1), \ldots, a(2^n\text{-}1)$ denote the ordered elements of A. The 1-α confidence limits for δ are then given by

$$\delta_L = a(K_\alpha), \quad \delta_U = a(2^n - K_\alpha), \tag{16}$$

where K_α is chosen such that $K_\alpha / 2^n \leq \alpha/2$.

For example, if $n = 12$, there are $2^{12} - 1 = 4095$ averages of observed differences: the 12 differences themselves, $\binom{12}{2} = 66$ averages of 2 distinct differences each, $\binom{12}{3} = 220$ averages of 3 distinct differences each, etc. From $102/2^{12} = 0.0249 < 0.05/2$, $K_\alpha = 102$ follows. Hence, the 95.02% confidence limits are given by the 102nd and the 3994th of the total of 4095 elements in A.

As with all randomization tests, it is not possible to tabulate a test statistic with its corresponding significance points, since the distribution of such a statistic is totally dependent on the observed data set. Only the indices K_α and 2^n-K_α of those ordered averages which determine the 1-α confidence limits can be given. For $n > 12$, the computing expenditure of the permutation test becomes prohibitive for routine use. Hence, in Table 3, the above indices are only given for $6 \leq n \leq 12$.

The modification of the above procedure to ratios instead of differences is the same as in the case of the signed rank test.

Example

The demographic data of 12 healthy male volunteers who participated in a cross-over study to investigate the influence of food intake on the bioavailability of

Table 6. Point estimate and 95%-confidence limits of bioavailability ratio for data given in Table 5

	Statistical method	Point estimate	95%-confidence limits	Exact level of confidence
Normal distribution	Paired *t*-test	1.03	0.97, 1.09	
	ANOVA	1.03	0.97, 1.10	
	Westlake		0.92, 1.08	≥ 0.95
Lognormal distribution	Paired *t*-test	1.04	0.97, 1.12	
	ANOVA	1.04	0.97, 1.12	
	Westlake		0.89, 1.11	≥ 0.95
Distribution-free (nonparametric) ratio analysis	Signed rank test (Tukey)	1.02	0.97, 1.11	0.9575
	Pitman's permutation test	1.04	0.97, 1.12	0.9502

Table 7. Analysis of variance for data given in Table 5

Source of variation	Degrees of freedom	Sum of squares	Mean-square	F-Test
Formulations	1	97.61	97.61	1.071 n.s.
Periods	1	0.88	0.88	0.010 n.s.
Subjects	11	31168.08	2833.46	31.094 $p < 0.001$
Error	10	911.25	91.13	
Total	23	32177.82		

Westlake's condition: $k_1 + k_2 = -2.07 \Rightarrow k_1 = 1.83$, $k_2 = -3.90$

Table 8. Analysis of variance for data given in Table 5 after logarithmic transformation (log = ln, antilog = exp)

Source of variation	Degrees of freedom	Sum of squares	Mean-square	F-Test
Formulations	1	0.0101701	0.01017	1.558 n.s.
Periods	1	0.0001614	0.00016	0.025 n.s.
Subjects	11	1.7628252	0.16026	24.547 $p < 0.001$
Error	10	0.0652858	0.00653	
Total	23	1.8384425		

Westlake's condition:

$k_2 = -4.657 \Rightarrow \text{antilog}\,(\bar{d}_{log} + k_2 \sqrt{2\,MSE_{log}/n}) = 0.894$

$k_1 = 1.818 \Rightarrow \text{antilog}\,(\bar{d}_{log} + k_1 \sqrt{2\,MSE_{log}/n}) = 1.106$

theophylline from a sustained-release aminophylline preparation are given in Table 4. The case in which the drug was taken after fasting overnight and a standard breakfast was eaten 2 h after taking the drug serves as the reference (R). Drug intake directly after consumption of the same standard breakfast is the test situation (T). The standard breakfast consisted of 2 rolls, butter, jam or honey, cold cuts or cheese spread, orange juice 0.2 l, and herb tea (camomile, rose-hip, peppermint) on request. For each participant, the randomly allocated sequence of administration is also given in Table 4. Only Subjects 1 and 6 were smokers, but they abstained from smoking during the study. No xanthine-containing foods and beverages were allowed during the study.

The administered dose, converted to anhydrous theophylline, was 385.6 mg. Serum theophylline levels were determined before and 0.5, 1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 24, 28 and 32 h after administration using radio-immuno assay RIA-mat® (Byk-Mallinckrodt, Dietzenbach; Zech et al. 1980).

The area under the concentration/time curve, AUC, was used to characterize the extent of absorption. AUC was determined by the log-trapezoidal rule up to the last sampling point, C_{last}, and then extrapolated to infinity by adding $\hat{C}_{last}/\hat{\beta}$. In order to avoid over- or underestimation of the extrapolated area, the measured C_{last} was replaced by its estimate $\hat{C}_{last}$, which in turn was obtained from the terminal log-linear regression line. AUC-values under both conditions (fasting and standard breakfast), as well as the corresponding ratios are given in Table 5. For the statistical methods presented, point estimate and 95%-confidence limits of the bioavailability ratio are summarized in Table 6. The detailed ANOVA-tables under the assumption of normal and lognormal distributions are given in Tables 7 and 8, respectively. The Tukey procedure is illustrated in Table 9, which gives the ordered values of the 78 geometric Walsh averages, together with the corresponding pairs of indices. The analogous table for Pitman's permutation test includes 4095 geometric averages and therefore cannot be reproduced in this paper. It is available from the authors upon request.

While the extent of absorption is readily characterized by AUC, it is much more difficult to characterize the rate of absorption by a single parameter such as t_{peak}, the time to reach the peak concentration. This is particularly true in the case of a sustained-release formulation, which produces a concentration plateau rather than a distinct peak. In the case of theophylline, serum concentrations differing by less than 0.5 mg/l are practically indistinguish-

Table 9. Tukey procedure for data given in Table 5: ordered values of 78 geometric Walsh averages. The 12 observed ratios are given by the pairs 1/1, . . ., 12/12. The 95%-confidence limits have been framed

No	Pair of ordered ratios	Geometric Walsh average	No	Pair of ordered ratios	Geometric Walsh average	No	Pair of ordered ratios	Geometric Walsh average
1	1/1	0.8815	27	4/ 8	0.9970	53	3/11	1.0844
2	1/2	0.9213	28	5/ 5	0.9978	54	4/11	1.0858
3	1/3	0.9258	29	5/ 6	1.0023	55	5/10	1.0957
4	1/4	0.9269	30	6/ 6	1.0067	56	5/11	1.0986
5	1/5	0.9379	31	2/ 9	1.0077	57	6/10	1.1006
6	1/6	0.9421	32	5/ 7	1.0077	58	6/11	1.1035
7	1/7	0.9472	33	5/ 8	1.0087	59	7/10	1.1065
8	1/8	0.9481	34	6/ 7	1.0122	60	2/12	1.1074
9	2/2	0.9628	35	3/ 9	1.0126	61	8/10	1.1076
10	1/9	0.9642	36	6/ 8	1.0132	62	7/11	1.1095
11	2/3	0.9675	37	4/ 9	1.0138	63	8/11	1.1106
12	2/4	0.9688	38	7/ 7	1.0177	64	3/12	1.1128
13	3/3	0.9722	39	7/ 8	1.0187	65	4/12	1.1142
14	3/4	0.9735	40	8/ 8	1.0198	66	9/10	1.1264
15	4/4	0.9747	41	5/ 9	1.0258	67	5/12	1.1273
16	2/5	0.9802	42	1/10	1.0298	68	9/11	1.1294
17	2/6	0.9845	43	6/ 9	1.0304	69	6/12	1.1323
18	3/5	0.9849	44	1/11	1.0326	70	7/12	1.1385
19	4/5	0.9862	45	7/ 9	1.0360	71	8/12	1.1396
20	3/6	0.9893	46	8/ 9	1.0370	72	9/12	1.1589
21	2/7	0.9899	47	9/ 9	1.0545	73	10/10	1.2031
22	4/6	0.9906	48	1/12	1.0596	74	10/11	1.2063
23	2/8	0.9909	49	2/10	1.0763	75	11/11	1.2096
24	3/7	0.9947	50	2/11	1.0792	76	10/12	1.2379
25	3/8	0.9957	51	3/10	1.0815	77	11/12	1.2412
26	4/7	0.9960	52	4/10	1.0829	78	12/12	1.2736

able. Within these limits, the observed plateaus lasted for 4–8 h, and so did not permit reliable estimation of t_{peak}.

In order to characterize the rate of absorption, we used classical deconvolution methods (Wagner and Nelson 1963; Loo and Riegelman 1968) to obtain in vivo absorption profiles. The individual theophylline disposition kinetics were obtained from an intravenous study with short-term infusion over 20 min of aminophylline 480 mg. As expected, absorption of the drug given after the standard breakfast was initially delayed.

Discussion and Conclusions

A review of statistical procedures, both parametric and nonparametric, to obtain 95%-confidence limits of expected bioavailability has been presented. It has been pointed out that the classical analysis of variance depends on certain assumptions, which are necessary to handle the mathematics, but have a limited bearing on clinical reality. In consequence, nonparametric procedures based on plausible assumptions are preferable. If the bioavailability characteristic has a continuous distribution, as is the case with AUC, then the Tukey procedure is the favoured choice.

With regard to the nominal confidence level of 95%, the improvement gained by Pitman's permutation test in comparison with the Tukey procedure is virtually negligible. This is also true for the effect of the interpolation formula. In practice, these refinements must be seen in relation to the error in AUC-values, which can easily amount to 5%, due to different methods of calculation alone (Yeh and Kwan 1978). In the example presented, AUC's obtained by curve-fitting and by the log-trapezoidal rule differed by 2% (−1%, 4%) under fasting conditions, and 0.5% (−5%, 4%) when administering the drug after the standard breakfast. The percentages given represent the median, and in brackets the minimum and maximum. On the other hand, interpolation affects the confidence limits insignificantly, namely from (0.973, 1.114) to (0.974, 1.113) when using the Tukey procedure. In the case of Pitman's permutation test, this effect is even smaller.

Another point of discussion is extrapolation of the AUC beyond the last sampling point, which is subject to debate if the extrapolated area exceeds a

certain fraction of the sampled area, say 20%. In an extreme case (Anttila et al. 1979), the decision on bioequivalence may even depend on the augmented AUC. Generally, if the AUC cannot be estimated correctly after a single dose (no samples taken at night; analytical limit of detection), bioavailability must be determined during steady-state conditions.

Regarding the representation of bioavailability data, demographic data, sequence of administration, individual values of the bioavailability characteristic (e.g. AUC), and individual bioavailability ratios should be given (cf. Tables 4 and 5). The point estimate and the 95%-confidence limits of bioavailability should also be given (cf. Table 6), and the statistical method used should be justified (e.g. residual plots).

In a series of about 30 examples, the confidence limits obtained by the procedures presented differed by less than 10% if 12 subjects were included in the study. In particular, results based on the ANOVA and the paired *t*-test were very similar, due to a period effect that usually was not significant. In studies with 6 subjects, the nonparametric confidence limits are directly based on the smallest and the largest bioavailability ratios and hence may be substantially wider than the corresponding parametric values. As far as the latter are concerned, the residual plots did not permit a distinction to be made in most cases between a normal and a log-normal distribution.

The largest deviations occured with Westlake's method, because the confidence interval is shifted away from the direction in which the sample difference has been found. In a theoretical example with a 3% coefficient of variation ($100 \times \sqrt{\text{MSE}}/\bar{y}_{\text{Reference}}$), Westlake's confidence limits ranged from 80 to 120% bioavailability ("bioequivalence"), although the bioavailability for all 12 subjects was below 85%. Generally, if the expected mean difference $|\delta|$ increases, or the error variance decreases, Westlake's method progressively changes – in favour of accepting bioequivalence from a two-sided to a one-sided approach (Kirkwood 1981). This effect and the loss of information on the direction of change when switching from reference to test formulation are major criticisms of Westlake's symmetrization (Shirley 1976; Mantel 1977; Kirkwood 1981; Mandallaz and Mau 1981; Mau 1981). Particularly in the case of a log-normal distribution and the corresponding ratio analysis, confidence limits symmetrical about unity are debatable. As they are frequently justified by symmetrical regulatory requirements (Westlake 1979, 1981), the latter should be modified so that they account for the multiplicative character of the log-normal distribution; for example, the bioequivalence range of [0.80, 1.20] should be replaced by [0.80, 1.25] (Mantel 1977; Steinijans 1981; Kirkwood 1981).

Recently, there have been several papers adopting a Bayesian viewpoint in bioequivalence assessment (Rodda and Davis 1980; Selwyn et al. 1981; Mandallaz and Mau 1981; Fluehler et al. 1981). The Bayesian approach not only provides a convenient way to review the difference between the conventional and Westlake's confidence-interval procedures (Armitage 1981), but it also allows generalization of the latter. For normal distributions and fixed subjects effects, Mandallaz and Mau (1981) derived an exact version of Westlake's procedure, i.e. without the approximation of $\mu_1 = \mu_{\text{Reference}}$ by $\bar{y}_{\text{Reference}}$. Using the conventional (improper, vague) prior distributions for μ_1, μ_2, and the error variance, they showed that the posterior probability of μ_2/μ_1 lying in $[r_1, r_2]$ is greater than or equal to $1\text{-}\alpha$, if and only if the exact, symmetrical $(1\text{-}\alpha)$-confidence interval lies in $[r_1, r_2]$. As before, r_1 and r_2, $0 < r_1 < 1 < r_2 < 2$, are bounds on the ratio of the expected formulation means μ_2 and μ_1, such that for $r_1 < \mu_2/\mu_1 < r_2$ the test and reference formulations are considered bioequivalent. In other words, the characteristic feature of Westlake's decision rule for bioequivalence is not symmetry, but rather its Bayesian interpretation. This is more readily seen from the approach taken by Rodda and Davis (1980). In their terminology, the odds are 19:1 (95:5) against a bioavailability difference of Δ percent, which in Westlake's terminology means the following: with 95% confidence the mean AUC for the test formulation is within Δ percent of that for the reference formulation.

The statistical methods presented in this paper are based on the simple two-way cross-over design. If more than 2 formulations are to be compared in a multiple cross-over study, more complex statistical procedures must be employed. The commonest comparative bioavailability trial is one in which 1 formulation (possibly replicated in each subject) serves as reference and in which the others are new formulations to be tested against it. A variety of experimental designs such as Latin squares, incomplete blocks and split-plot designs, and the corresponding analyses of variance have been described in the literature (Cochran and Cox 1957; Westlake 1974; Shirley and Unwin 1978). However, all these publications deal with hypothesis testing only and do not provide confidence intervals. In this area some valuable biometrical research could be undertaken, particularly with respect to nonparametric confidence intervals.

Acknowledgements. We wish to thank Drs. D. Hartmann, J. Mau, Th. Royen, and H. P. Wijnand for their valuable remarks and suggestions.

References

Abt K (1977) Cross-over Versuchspläne: Grenzen der Anwendung und der parametrischen Auswertung. Paper presented at the Annual Meeting of the Austro-Swiss Region of the International Biometric Society, Krems/Austria, September 1977

Anttila M, Kahela P, Panelius M, Yrjänä T, Tikkanen R, Aaltonen R (1979) Comparative bioavailability of two commercial preparations of carbamazepine tablets. Eur J Clin Pharmacol 15: 421–425

Armitage P (1981) Bioequivalence testing – a need to rethink: note by editor. Biometrics 37: 593–594

Cochran WG, Cox GM (1957) Experimental design; 2nd ed. Wiley, New York, pp 117–147

Fluehler H, Hirtz J, Moser HA (1981) An aid to decision-making in bioequivalence assessment. J Pharmacokinet Biopharm 9: 235–243

Geigy Wissenschaftliche Tabellen (1980), Vol. 3; 8th ed. Thieme, Stuttgart, p 163

Grieve AP (1982) The two-period changeover design in clinical trials. Biometrics 38: 517

Grizzle JE (1965) The two-period change-over design and its use in clinical trials. Biometrics 21: 467–480. Correction note (1974) Biometrics 30: 727

Hollander M, Wolfe DA (1973) Nonparametric statistical methods. Wiley, New York

Kirkwood TBL (1981) Bioequivalence testing – a need to rethink. Biometrics 37: 589–591

Koch GG (1972) The use of non-parametric methods in the statistical analysis of the two-period change-over design. Biometrics 28: 577–584

Kramer WG, Kolibash AJ, Bathala MS, Visconti JA, Lewis RP, Reuning RH (1977) Digoxin bioavailability: evaluation of a generic tablet and proposed FDA guidelines. J Pharm Sci 66: 1720–1722

Loo JCK, Riegelman S (1968) New method for calculating the intrinsic absorption rate of drugs. J Pharm Sci 57: 918–928

Mandallaz D, Mau J (1981) Comparison of different methods for decision-making in bioequivalence assessment. Biometrics 37: 213–222

Mantel N (1977) Do we want confidence intervals symmetrical about the null value? Biometrics 33: 759–760

Mau J (1981) Die Verwendung von Fiduzialwahrscheinlichkeiten zur Beurteilung der Bioäquivalenz. To be published in the Proceedings of the 27th meeting of the German region of the International Biometric Society, Bad Nauheim, March 1981

Metzler CM (1974) Bioavailability – a problem in equivalence. Biometrics 30: 309–317

Ostle B (1966) Statistics in research. Iowa State University Press, Iowa

Rietbrock N, Alken RG, Ebert W (1979) Vergleichende Untersuchung der absoluten Bioverfügbarkeit von vier oralen Digoxin-Präparaten. Arzneim Forsch (Drug Res) 29: 1742–1745

Rodda BE, Davis RL (1980) Determining the probability of an important difference in bioavailability. Clin Pharmacol Ther 28: 247–252

Royen T (1978) Randomisierungstests zum Vergleich verbundener Stichproben. EDV in Medizin und Biologie 9: 104–106

Selwyn MR, Dempster AP, Hall NR (1981) A Bayesian approach to bioequivalence for the 2×2 changeover design. Biometrics 37: 11–21

Shirley E (1976) The use of confidence intervals in biopharmaceutics. J Pharm Pharmacol 28: 312–313

Shirley EAC, Unwin PF (1978) The analysis of data from comparative bioavailability studies. Eur J Drug Metab Pharmacokinet 3: 165–170

Spriet A, Beiler D (1978) Table to facilitate determination of symmetrical confidence intervals in bioavailability trials with Westlake's method. Eur J Drug Metab Pharmacokinet 3: 129–132

Steinijans VW (1981) Verteilungsfreier Punktschatzer und Vertrauensgrenzen für Bioverfügbarkeitsquotienten. To be published in the Proceedings of the 27th meeting of the German Region of the International Biometric Society, Bad Nauheim, March 1981

Steinijans VW, Eicke R, Ahrens J (1982) Pharmacokinetics of theophylline in patients following short-term intravenous infusion. Eur J Clin Pharmacol 22: 417–422

Upton RA, Sansom L, Guentert TW, Powell JR, Thiercelin JF, Shah VP, Coates PE, Riegelman S (1980) Evaluation of the absorption from 15 commercial theophylline products indicating deficiencies in currently applied bioavailability criteria. J Pharmacokinet Biopharm 8: 229–242

Vila JL, Martinez R, Giménez J, Llabrés M (1980) MANOVA of statistical moments in biopharmaceutical studies: a numerical example with three equally spaced doses of amoxicillin. J Pharmakokinet Biopharm 8: 411–420

Wagner JG, Nelson E (1963) Per cent absorbed time plots derived from blood level and/or urinary excretion data. J Pharm Sci 52: 610–611

Wallenstein S, Fisher AC (1977) The analysis of the two-period repeated measurements crossover design with applications to clinical trials. Biometrics 33: 261–269

Westlake WJ (1972) Use of confidence intervals in analysis of comparative bioavailability trials. J Pharm Sci 61: 1340–1341

Westlake WJ (1973) Use of statistical methods in evaluation of in vivo performance of dosage forms. J Pharm Sci 62: 1579–1589

Westlake WJ (1974) The use of balanced incomplete block designs in comparative bioavailability trials. Biometrics 30: 319–327

Westlake WJ (1976) Symmetrical confidence intervals for bioequivalence trials. Biometrics 32: 741–744

Westlake WJ (1979) Statistical aspects of comparative bioavailability trials. Biometrics 35: 273–280

Westlake WJ (1981) Bioequivalence testing – a need to rethink: response. Biometrics 37: 591–593

Wilcoxon F, Katti SK, Wilcox RA (1973) Probability levels for the Wilcoxon signed rank test. In: Harter HL, Owen DB (eds) Selected Tables in Mathematical Statistics, Vol. I. American Mathematical Society, Providence, Rhode Island

Yeh KC, Kwan KC (1978) A comparison of numerical integrating algorithms by trapezoidal, Lagrange, and spline approximation. J Pharmacokinet Biopharm 6: 79–81

Zech K, Borner K, v. Stetten O (1980) Monitoring of serum or saliva theophylline concentrations by a new radioimmunoassay. Anal Chem 301: 114

Received: February 8, 1982
in revised form: July 22, 1982
accepted: July 23, 1982

Dr. Volker W. Steinijans
Biometry Department
Byk Gulden Research Centre
Byk-Gulden-Str. 2
D 7750 Konstanz
Federal Republic of Germany