A Nonparametric Test for Similarity of Marginals - with Applications to the Assessment of Population Bioequivalence

Gudrun Freitag [a,1] Claudia Czado [b] Axel Munk [a]

[a] *Georg-August Universität Göttingen, Institut für Mathematische Stochastik*

[b] *Technische Universität München, Zentrum Mathematik*

Abstract

In this paper we suggest a completely nonparametric test for the assessment of similar marginals of a multivariate distribution function. This test is based on the asymptotic normality of Mallows distance between marginals. It is also shown that the n out of n bootstrap is weakly consistent, thus providing a theoretical justification to the work in Czado & Munk [12]. The test is extended to cross-over trials and is applied to the problem of population bioequivalence, where two formulations of a drug are shown to be similar up to a tolerable limit. This approach was investigated in small samples using bootstrap techniques in [12], showing that the bias corrected and accelerated bootstrap yields a very accurate and powerful finite sample correction. A data example is discussed.

Key words: Bioequivalence; Cross-over trials; Hadamard derivative; Limit law; Multivariate empirical process; Pre-post comparison

1 Introduction

In many applications the aim is to compare the marginals F and G of a bivariate distribution H by means of an *i.i.d.* sample $Z_i = (X_i, Y_i) \sim H$, $i = 1, \ldots, n$. This problem occurs when several measurements under two different experimental conditions are taken from one observation unit. For the hypothesis of equal marginals, $H_0 : F = G$, various rank tests have been

[1] Address for Correspondence: Institut für Mathematische Stochastik, Georg-August-Universität Göttingen, Maschmühlenweg 8-10, 37073 Göttingen, Germany. Tel. +49 551 3913510, Fax +49 551 3913505, Email: freitag@math.uni-goettingen.de

suggested [27,28,43]. Often, however, it is the aim to demonstrate *similarity* instead of a difference between F and G, i.e. that F and G are sufficiently 'close', in a sense to be made precise below. On the one hand, this is of practical interest in many applications. On the other hand, surprisingly it turns out that the mathematical analysis becomes often more simple (cf. Theorem 1 and Remark 10).

One example is the investigation of diagnostic procedures where a new measuring method has to be shown as comparable to a gold standard, e.g. when several sonographic methods are compared to histology for the determination of skin cancer size [22]. Another application is the assessment of population bioequivalence, where the aim is to demonstrate similar bioavailibility for different formulations of a drug (cf. [50,45,25,49]).

The aim of this paper is to provide a nonparametric test for the assessment of similar marginals F and G. Extending Munk & Czado's [42] approach, we suggest as a measure to compare F and G a trimmed version of Mallows distance,

$$\Gamma_\beta(F,G) = \left\{ \frac{1}{(1-2\beta)} \int_\beta^{1-\beta} |F^{-1}(u) - G^{-1}(u)|^2 du \right\}^{\frac{1}{2}}, \qquad \beta \in [0, \frac{1}{2}) \quad (1)$$

where we assume that F and G are in the class of cumulative distribution functions (c.d.f.s) defined by

$$\mathcal{F}_2 := \{F : F \text{ is a continuous c.d.f. and } \int x^2 dF(x) < \infty\} . \quad (2)$$

We mention that trimming has several advantages. It leads to a robustification of the suggested methods (see also [7]). Further, it simplifies the mathematical analysis significantly (cf. Section 2.2).

The paper is organized as follows. In Section 2 we propose an estimator for $\Gamma_\beta(F,G)$, and its asymptotic normality will be shown. This allows to construct tests for the hypotheses

$$H_\Delta : \Gamma_\beta(F,G) > \Delta_0 \quad \text{versus} \quad K_\Delta : \Gamma_\beta(F,G) \leq \Delta_0, \quad (3)$$

where $\Delta_0 > 0$ is a pre-specified bound the experimenter is willing to tolerate between F and G. Note that rejection of H_Δ allows the assessment of similar F and G at a controlled error rate, in contrast to the more conventional hypothesis $H : F = G$. For computational reasons, we suggest in Section 3.1 the use of the bias corrected and accelerated bootstrap of the proposed test statistic. Furthermore, applications to the assessment of population bioequivalence

are discussed in Sections 3.2 to 3.4. For this the proposed test is extended to a cross-over design as it is custom in bioequivalence trials. An example of a bioequivalence study for a vasoactive drug is presented in Section 3.5. Finally, in Section 4 we discuss various extensions, such as the similarity of more than two marginals and higher order cross-over designs. We mention that Sections 3 and 4 are understandable without previous reading of the more technical Section 2.2.

2 Asymptotic theory for Mallows distance

2.1 General properties and estimation of Mallows distance

Following the definition (1), the trimmed Mallows distance Γ_β quantifies the trimmed L^2-norm between the quantile functions F^{-1} and G^{-1} of F and G. Here $F^{-1}(t) = \inf\{u : F(u) \geq t\}$ denotes the left continuous inverse of a c.d.f. F. In the following we often write $\gamma_\beta := \Gamma_\beta^2$. This distance was previously investigated for the situation of two independent treatment groups by Munk & Czado [42] and in the context of goodness of fit testing by del Barrio et al. [13–15]. Observe that in the untrimmed case for $\beta = 0$, we obtain Mallows [40] L^2-distance which is also sometimes denoted as Wasserstein or Kantorovitch-Rubinstein metric (cf. Dobrushin [16]). This distance (for $\beta = 0$) has received some interest among probabilists, because it can be shown to be equal to $\inf_{\mu \in \mathcal{M}(F,G)} \{\int |x-y|^2 \mu(dx, dy)\}$, where $\mathcal{M}(F, G)$ denotes the set of distributions with given marginals F and G (cf. Rüschendorf & Rachev [44] for a survey). Some nice properties as a distance between distribution functions reveal Γ_β as a suitable measure for our purposes (cf. Munk & Czado [42], and Section 3.2).

In order to estimate $\Gamma_\beta(F, G)$, F and G are replaced by the empirical c.d.f.s F_n and G_n, respectively. Hence we obtain

$$\hat{\Gamma}_\beta = \Gamma_\beta(F_n, G_n) = \left\{ \frac{1}{1-2\beta} \int_\beta^{1-\beta} |F_n^{-1}(u) - G_n^{-1}(u)|^2 \, du \right\}^{1/2},$$

where

$$F_n^{-1}(t) = \inf\{u : F_n(u) \geq t\} = \sum_{i=1}^n X_{(i)} \mathbf{1}_{\{(i-1)/n < t \leq i/n\}}, \quad t \in (0,1), \tag{4}$$

denotes the left continuous inverse of the empirical c.d.f. $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}$,

and $X_{(i)}$ the i^{th} order statistic of the sample $X_1, \ldots, X_n$. Note that from (4) it follows that

$$\begin{aligned}\hat{\gamma}_\beta := \hat{\Gamma}^2_\beta &= \tfrac{1}{(1-2\beta)n} \textstyle\sum_{i=[n\cdot\beta]+1}^{n-[n\cdot\beta]} |X_{(i)} - Y_{(i)}|^2 \\ &\quad + \tfrac{[n\beta]-n\beta}{(1-2\beta)n}\big(|X_{([n\beta]+1)} - Y_{([n\beta]+1)}|^2 + |X_{(n-[n\beta])} - Y_{(n-[n\beta])}|^2\big) \\ &= \tfrac{1}{(1-2\beta)n} \textstyle\sum_{i=[n\cdot\beta]+1}^{n-[n\cdot\beta]} |X_{(i)} - Y_{(i)}|^2 + O_p(\tfrac{1}{n}),\end{aligned} \tag{5}$$

where $[x]$ denotes the largest integer smaller than or equal to x for any $x \in \mathbb{R}$. For the case of no trimming this simplifies to $\hat{\gamma}_0 = \frac{1}{n}\sum_{i=1}^n |X_{(i)} - Y_{(i)}|^2$, and for the case $\beta \nearrow 1/2$ we have $\hat{\gamma} = |\mathrm{med}_X - \mathrm{med}_Y|^2$, the squared difference of the sample medians. For the subsequent test procedures we will consider the test statistic

$$T_{\Delta_0,\beta}(F_n, G_n) := \sqrt{n}(\hat{\gamma}_\beta - \Delta_0^2) \tag{6}$$

to establish similarity of F and G for a fixed bound $\Delta_0 > 0$ (cf. (3)).

2.2 Asymptotic theory

For the case of two *independent* samples $X_1, \ldots, X_m \sim F$ and $Y_1, \ldots, Y_n \sim G$ and under some smoothness and growth conditions on F and G, Munk & Czado [42] derived the asymptotic normality of the test statistic

$$(\tfrac{nm}{n+m})^{1/2}\{\gamma_{\beta_{m\wedge n}}(F_m, G_n) - \gamma_\beta(F, G)\},$$

provided that $\gamma_\beta > 0$. The proof utilizes strong approximation results of [11] of the weighted quantile process. Here $m \wedge n$ denotes the minimum of m and n, and $\beta_n \searrow 0$ is a sequence of trimming bounds which does not converge too fast to zero, i.e.

$$\beta_n \geq c\,\frac{\log\log n}{n} \tag{7}$$

for some $c > 0$. The purpose of this section is to extend this result to the situation where dependencies between X and Y may occur. We mention that this setup occurs in many experimental designs, such as pre-post comparisons or cross-over trials. This will be discussed in detail in the next section. In the following we will show the asymptotic normality of

$$T_n := \sqrt{n}(\hat{\gamma}_{\beta_n} - \gamma_\beta),$$

as long as $\beta_n \to \beta > 0$, $\gamma_\beta > 0$, and H satisfies some smoothness conditions. However, the asymptotic variance of T_n turns out to be rather complicated, and its estimation would be very cumbersome. Note that for the case of independent X and Y, as treated in [42], this leads already to a rather difficult problem. Therefore, we suggest in the following bootstrap tests for the hypotheses in (3) which offers an appealing alternative. The second result of Theorem 1 shows that indeed the n out of n bootstrap distribution $\mathbf{P}^*$ mimics asymptotically the law $\mathbf{P}_n$ of T_n. Let H_n denote the bivariate empirical c.d.f. of the sample $\mathbf{Z}_n = (Z_i)_{i=1,\cdots,n}$ and H_n^* the corresponding bootstrap c.d.f. obtained by drawing a sample of size n from $\mathbf{Z}_n$ with replacement.

Theorem 1 *Let $Z_1, \cdots, Z_n \sim H$ i.i.d., where H denotes a continuously differentiable bivariate distribution function with marginals $F(\cdot) = H(\cdot, \infty)$ and $G(\cdot) = H(\infty, \cdot)$, such that F and G are continuously differentiable with positive densities $f > 0$ and $g > 0$ on the real line, and let $0 < \beta < 1/2$ and $\beta_n \to \beta$.*

(1) Then, if $0 < \gamma_\beta(F, G) < \infty$, we have that

$$T_n \xRightarrow{\mathcal{D}} \mathcal{N}(0, \tau_\beta^2(H)), \tag{8}$$

where the limiting variance is calculated as

$$\begin{aligned}\tau_\beta^2(H) = \frac{4}{(1-2\beta)^2} \Bigg\{ & \int_0^{1-\beta} \xi_{(F|G)}^2(s)ds - \Big(\int_0^{1-\beta} \xi_{(F|G)}(s)ds\Big)^2 \\ & + \int_0^{1-\beta} \xi_{(G|F)}^2(s)ds - \Big(\int_0^{1-\beta} \xi_{(G|F)}(s)ds\Big)^2 \\ & + 2\int_0^{1-\beta}\int_0^{1-\beta} \xi_{(F|G)}(r)\xi_{(G|F)}(s)\frac{\partial^2 H(F^{-1}(r), G^{-1}(s))}{\partial r \partial s}drds \\ & - 2\int_0^{1-\beta} \xi_{(F|G)}(r)dr \int_0^{1-\beta} \xi_{(G|F)}(s)ds \Bigg\}, \end{aligned} \tag{9}$$

which is finite as $0 < \beta < 1/2$. Here we have used the abbreviations

$$\xi_{(F|G)}(s) = \textstyle\int_{\beta\vee s}^{1-\beta} \frac{F^{-1}(t)-G^{-1}(t)}{f\circ F^{-1}(t)}dt,\ \xi_{(G|F)}(s) = \int_{\beta\vee s}^{1-\beta} \frac{G^{-1}(t)-F^{-1}(t)}{g\circ G^{-1}(t)}dt\ ,$$

where $x \vee y := \max\{x, y\}$.

(2) Furthermore, we have with $\psi(H) := \gamma_\beta(F, G)$ that

$$\sup_{t\in\mathbb{R}} \Big|\mathbf{P}^*[n^{\frac{1}{2}}(\psi(H_n^*) - \psi(H_n)) \le t] - \mathbf{P}[\dot{\psi}_H(\mathbb{B}_H) \le t]\Big| \xrightarrow[n\to\infty]{P} 0,$$

where $\mathbb{B}_H(\cdot,\cdot)$ denotes a two-dimensional Brownian sheet with

$$Cov[\mathbb{B}_H(\mathbf{s}), \mathbb{B}_H(\mathbf{t})] = H(\mathbf{t} \wedge \mathbf{s}) - H(\mathbf{t})H(\mathbf{s}),$$

$\mathbf{s}, \mathbf{t} \in \mathbb{R}^2$, and the minimum is to be understood componentwise. Here $\dot{\psi}_H(\cdot)$ denotes the Hadamard derivative of ψ at H. Hence, the n out of n bootstrap is uniformly weakly consistent.

Sketch of the proof: The proof of the first assertion of Theorem 1 follows from the proof of Theorem 2.5 in Freitag et al. [23]. There the Hadamard differentiability of the functional ψ tangentially to a subspace is shown and the Hadamard derivative $\dot{\psi}_H$ is given. Further, we have the invariance principle for the multivariate empirical process, $n^{1/2}(H_n - H) \stackrel{\mathcal{D}}{\Longrightarrow} \mathbb{B}_H$, on the space $D[0,1]^2$ equipped with the supremum norm and open ball σ-field (cf. Bickel & Wichura [4]). Thus, the functional delta method can be applied, which can then be transferred to the bootstrapped functional (cf. Section 3 in [23]), yielding the second assertion of the theorem. The underlying theory can be found in Gill [24] and van der Vaart & Wellner [52]. Note that the limiting random element in (8) is equal in distribution to $\dot{\psi}_H(\mathbb{B}_H)$, i.e. the functional derivative applied to the limiting Gaussian process $\mathbb{B}_H$. Lemma 7.1 in [23] can then be applied to calculate the variance of this normal random variable, which is equal to $\tau_\beta^2(H)$ in (9).

Remark 2 *We mention that for the case of independent X and Y the asymptotic variance in (9) reduces to*

$$\tau_\beta^2 = \frac{4}{(1-2\beta)^2} \left\{ \int_0^{1-\beta} \xi_{(F|G)}^2(s)ds - \Big(\int_0^{1-\beta} \xi_{(F|G)}(s)ds \Big)^2 \right.$$
$$\left. + \int_0^{1-\beta} \xi_{(G|F)}^2(s)ds - \Big(\int_0^{1-\beta} \xi_{(G|F)}(s)ds \Big)^2 \right\},$$

because in this case (for $r, s \in [\beta, 1-\beta]$), $\frac{\partial^2 H(F^{-1}(r), G^{-1}(s))}{\partial r \partial s} \equiv 1$, which gives Theorem 1 in [42] for $\beta > 0$. Observe that the scaling factor $4/(1-2\beta)^2$ of τ_β^2 is slightly different from Munk & Czado [42], because the trimmed Mallows distance in this paper has a different scaling. Interestingly, we did not succeed in proving a similar result to Munk & Czado's Theorem 1 for trimming bounds $\beta_n \searrow 0$ with (7). This is related to the difficult problem of finding jointly *a strong approximation of the bivariate weighted quantile process*

$$q_n(\cdot,\cdot) := n^{1/2} \Big\{ f \circ F^{-1}(\cdot)(F_n^{-1}(\cdot) - F^{-1}(\cdot)), g \circ G^{-1}(G_n^{-1}(\cdot) - G^{-1}(\cdot)) \Big\}$$

by Brownian bridges $(\mathbb{B}_F^{(n)}(\cdot), \mathbb{B}_G^{(n)}(\cdot))$ *with copula* $H(F^{-1}(s), G^{-1}(t)) - st$. *Nevertheless, observe that in most applications the trimming bound can still be chosen as* $\beta = 0$, *because the last theorem remains valid for this case when the support of* F *and* G *is compact. This is, e.g., the case for all pharmacokinetic quantities under consideration in bioequivalence trials.*

Remark 3 *From the representation of the limiting random element in Theorem 1 as* $\dot{\psi}_H(\mathbb{B}_H)$, *an alternative expression for the asymptotic variance* $\tau_\beta^2(H)$ *can be obtained as*

$$\tau_\beta^2(H) = \frac{4}{(1-2\beta)^2} \int_\beta^{1-\beta} \int_\beta^{1-\beta} K_H(s,t) ds dt,$$

where

$$\begin{aligned} K_H(s,t) = &(F^{-1}(s) - G^{-1}(s))(F^{-1}(t) - G^{-1}(t)) \\ &\left(\frac{s \wedge t - st}{f(F^{-1}(s)) f(F^{-1}(t))} + \frac{s \wedge t - st}{g(G^{-1}(s)) g(G^{-1}(t))} \right. \\ &\left. + \frac{H(F^{-1}(s), G^{-1}(t)) - st}{f(F^{-1}(s)) g(G^{-1}(t))} + \frac{H(F^{-1}(t) G^{-1}(s)) - st}{f(F^{-1}(t)) g(G^{-1}(s))} \right). \end{aligned}$$

This might be more convenient for the estimation of τ_β^2, *since it does not involve second derivatives of* H.

Remark 4 *We would like to stress that the above method of proof of the consistency of the bootstrap transfers to other metrics, such as* L^p*-norms* ($p > 1$), *as long as they are Hadamard differentiable with non-vanishing derivative. Furthermore, for the case of independent observations* X *and* Y, *the Wilcoxon-functional* $P(X > Y)$ *is also known to be Hadamard differentiable (this can even be shown uniformly over* $(D[-\infty, \infty])^2$, *c.f. Steland [51]), which yields the strong consistency of the bootstrap for the tests suggested by Wellek [56] or by Schall [45] as a by-product).*

3 Applications

3.1 Bootstrapping Mallows distance

In order to perform the bootstrap test, the percentile (PC) method or Efron & Tibshirani's [17] bias corrected and accelerated (BC_a) method for constructing bootstrap confidence intervals can be used (see Efron & Tibshirani [17],

Chapters 13-14). For this let $T_{\Delta_0,B,1-\alpha_{sig}}$ be the $(1-\alpha_{sig})$-th empirical quantile based on $(T^1_{\Delta_0},\cdots,T^B_{\Delta_0})$. Here, α_{sig} is the significance level, and $T^b_{\Delta_0}$ denotes the b^{th} bootstrapped test statistic T_{Δ_0} from (6), where we have suppressed the dependency on the trimming constant β $(b = 1, ..., B)$. Details on how to actually obtain the bootstrap samples in particular cases will be given in Section 3.4.

The PC method consists in rejecting H_Δ from (3) at significance level α_{sig} if $T_{\Delta_0,B,1-\alpha_{sig}} \leq 0$. For the BC_a method the $(1-\alpha_{sig})$-th percentile of the bootstrap sample from the PC method is replaced by the α_{up}-th percentile, where α_{up} is defined as

$$\alpha_{up} = \alpha_{up}(\alpha_{sig}) = \Phi\left(\hat{Z}_0 + \frac{\hat{Z}_0 + z_{1-\alpha_{sig}}}{1-\hat{a}(\hat{Z}_0 + z_{1-\alpha_{sig}})}\right).$$

Here

$$\hat{Z}_0 = \Phi^{-1}\Big(\frac{\#\{T^b_{\Delta_0} < T_{\Delta_0}, b = 1,\cdots,B\}}{B}\Big), \quad \hat{a} = \frac{\sum_{i=1}^n (T_{(\cdot)} - T_{(i)})^3}{6(\sum_{i=1}^n (T_{(\cdot)} - T_{(i)})^2)^{\frac{3}{2}}},$$

z_p and Φ are the p-quantile and the c.d.f. of the standard normal distribution, respectively, and $\#A$ is the number of elements of the set A. Finally, $T_{(i)}$ is the resulting test statistic when the i-th observation is removed, and $T_{(\cdot)}$ is the mean of $T_{(i)}$, i.e. $T_{(\cdot)} = \frac{1}{n}\sum_{i=1}^n T_{(i)}$. Note that if $\hat{a} = \hat{Z}_0 = 0$, we have $\alpha_{up} = 1 - \alpha_{sig}$, i.e. the BC_a method coincides with the PC method. This correction allows the bootstrap confidence interval to be second-order accurate (Efron & Tibshirani [17], p. 325). Now it is possible to calculate from this the corresponding p-values for the PC test as

$$\text{p-value}(PC) = \min\{p : T_{\Delta_0,B,1-p} \leq 0\} = 1 - \frac{\sum_{b=1}^B 1_{\{T^b_{\Delta_0} \leq 0\}}}{B},$$

and for the BC_a test as

$$\text{p-value}(BC_a) = \min\{p : T_{\Delta_0,B,\alpha_{up}(p)} \leq 0\} = \alpha_{up}^{-1}\left(\frac{\sum_{b=1}^B 1_{\{T^b_{\Delta_0} \leq 0\}}}{B}\right).$$

Czado & Munk [12] have investigated in a comprehensive Monte Carlo study the PC and the BC_a method for testing the similarity of marginals. In summary, we found that the BC_a bootstrap outperforms the naive PC bootstrap (and also the test based on the asymptotic normality with estimated variance; cf. Theorem 2.1) over a broad range of possible underlying distributions, trimming bounds and sample sizes. The BC_a test performs particularly well when

distributions are skewed, a situation which occurs rather often in bioequivalence trials.

3.2 Nonparametric population bioequivalence

In this section we will apply the proposed tests to the problem of population bioequivalence. Bioequivalence studies are conducted in order to show similar bioavailability for different formulations of a drug, typically a reference formulation and a generic one. In this case it is accepted that the formulations are therapeutically similar, which implies that the generic one is allowed to replace the standard drug. Until now, three different types of bioequivalence are suggested by regulatory guidelines: average, individual and population bioequivalence [19]. During the past the most common approach was average bioequivalence, which means a similar absorption of the active ingredient in mean [18,8,9,20]. More recently, there has begun a controversial discussion on the use of average bioequivalence in order to guarantee so-called *prescribability* and *switchability* of two formulations. Hauck & Anderson [30] argued forcefully that similar prescribability requires the assessment of population bioequivalence, which means equivalence with respect to the underlying distribution functions. This is founded in the fact that prescribability refers to the situation where the patient starts on a new drug, and no individual characteristics are taken into account (cf. [50,46,45,25]). Although there is an agreement that the *entire* distribution functions of the test and reference formulation should be taken into account for the assessment of population bioequivalence, the suggested methodology of testing is restricted in most cases to moment-based criteria (e.g. [1,29,35,31,54,26,58]). This is also reflected in the corresponding guidelines of the FDA [19] and of the CPMP [9].

In a parametric setup (typically it is assumed that the data are log-normally or normally distributed), similarity of the first two moments implies also similarity of distribution functions, however, in a nonparametric framework this is not sufficient. Nevertheless, it has been recognized during the past by various authors that the normality assumption is often questionable (such an example will be discussed in the following). Further, outliers may have drastic consequences on parametric analyses of bioequivalence studies, as pointed out by Chow & Tse [7] or Liu & Weng [39].

Hence, distribution-free methods were proposed in [32,45,41,56,57,36], where most often the Mann-Whitney measure $p := P(X_R > X_T)$ was used. Bioequivalence is then claimed if $|p - 1/2| \leq \Delta$ for some limit Δ. We would like to stress, however, that a serious lack of the functional $P(X_R > X_T)$ (and related ones) consists in the fact that it does not allow for a rigid assessment of similar marginals F and G ($X_R \sim F, X_T \sim G$) in a completely nonparametric

framework. Typically, a valid interpretation of p for population bioequivalence can only be obtained in semiparametric models, such as location (see Wellek [56]) or Lehmann (see Munk & Czado [41]) families of alternatives. In order to illustrate the difficulties encountered with the Mann-Whitney functional p we show in the following that 'perfect equivalence' of F and G may hold, i.e. $p = 1/2$, although the shapes of the distributions are completely different. The following result gives sufficient conditions under which $p = 1/2$. For this we assume that X_R and X_T are independent, i.e.

$$P(X_R > X_T) = \int F(x)G(dx) =: \omega(F, G).$$

Theorem 5 *Assume that g is a transformation such that $-g(x) = g(-x)$ and let F a continuous c.d.f. symmetric around zero, i.e. $F(x) = 1-F(-x)$. Then we have $\omega(F \circ g, F) = 1/2$.*

Proof: We have

$$\omega(F \circ g, F) = \int_{-\infty}^{\infty} F(g(x))dF(x) = 1 - \int_{-\infty}^{\infty} F(g(-x))dF(x),$$

because $1 - F(g(-x)) = 1 - F(-g(x)) = F(g(x))$. By symmetry of F,

$$\omega(F \circ g, F) = \int F(g(-x))dF(x),$$

which proves the assertion. □

As an example consider the case when F and G are in the scale family of a c.d.f. F,

$$\mathcal{F}_S := \left\{ G(\cdot) = F\left(\frac{\cdot}{\sigma}\right), F \text{ is a symmetric c.d.f. } \right\}.$$

Here $g(X) = X/\sigma$, which obviously satisfies the condition $-g(x) = g(-x)$. Hence semiparametric distances will be in general not suitable for the assessment of population bioequivalence because in particular changes in scale and hence in the second moment cannot be detected. Roughly speaking, the last theorem shows that in general the quantity $P(X_T > X_R)$ measures how *often* X_T exceeds X_R, but not how *much*. Similar arguments apply, of course, to related measures, such as $P(|X_T - X_R| \leq \epsilon)$ or $P(1/(1+\epsilon) \leq X_T/X_R \leq 1+\epsilon)$ and individualized versions thereof (cf. e.g. Schall's [45] definition (2.2)). This has never been recognized in the literature, because in a location scale model with homoscedastic error $\omega(F, G)$ is proportional to the mean difference standardized by the scale parameter (cf. [45] for the case of normal distributions). Homoscedasticity, however, is only a valid assumption in crossover experiments

when the analysis is based on one-dimensional independent observations (such as individual differences or ratios), provided the variance does not depend on the sequence (which can be excluded in general (Hauschke et al. [33]). In a nonparametric framework, however, a reduction of the array of observations associated with each individual under study to a *one-dimensional* quantity can be completely misleading for the nonparametric assessment of population bioequivalence. This becomes transparent in Section 3.4, where it is shown that the suggested test statistic *cannot* be reduced to a statistic based on individualized quantities. Hence, Theorem 5 explains the somewhat curious phenomenon observed in various (average) bioequivalence studies during the past that Mann-Whitney type tests such as Hauschke et al.'s [32] procedure have led to a decision in favor of equivalence for a given data set, although the parametric standard test 'TOST' (two one-sided t-tests; cf. Schuirmann [47], Berger & Hsu [2]) does not. This seems to be a main reason why regulatory agencies, such as the FDA, did prescribe parametric methods in the past (cf. e.g. the FDA guidance [19], or the CPMP note for guidance [9]).

In contrast to the Mann-Whitney measure, the suggested Mallows distance Γ_β as a measure of similarity combines various useful aspects which are demanded for a measure of population bioequivalence.

(1) Trimming allows for robustification against outliers.
(2) In location-scale families the Mallows distance Γ_β leads to an aggregate bioequivalence criterion which controls simultaneously differences in the means, μ, η, and in the variances, σ^2, τ^2 (cf. (10); [42]). Further, in a pure location model we find for any $\beta \in [0, 1/2)$ that $\Gamma_\beta = |\mu - \eta|$, i.e. it coincides with the classical criterion of average bioequivalence. This has been recently demanded by Hauschke & Steinijans [34], who proposed for any criterion of population bioequivalence that it should always contain average bioequivalence as a special case.
(3) Finally, any bioequivalence criterion based on Mallows distance is defined on the original scale (there is even no need to use logarithmic transforms of the data), which allows a direct comparison of observed drug effects in terms of the relevant pharmacokinetic quantities such as AUC, T_{max}, etc., or transformations of them.

Observe that in a normal setup, the squared Mallows distance reduces to

$$\gamma_0 = (\mu - \eta)^2 + (\sigma - \tau)^2. \tag{10}$$

This is quite similar to the unscaled population bioequivalence measure for standard cross-over designs proposed by Schall & Luus [46], which is defined by

$$D_p = (\mu - \eta)^2 + (\sigma^2 - \tau^2). \tag{11}$$

However, (11) can take on negative values while (10) is strictly positive. The recent FDA guidance [19] proposes a (mixed) scaled version of D_p as a criterion for population bioequivalence (cf. (16)). However, this was criticized by various authors, cf. e.g. [5,26,34,58]. The main issue is the possibility to mask a large mean difference by a comparatively large negative value of $\sigma^2 - \tau^2$. Thus, the FDA guidance [19] demands to accompany the test on the criterion for population bioequivalence by checking whether the point estimate of mean difference falls within the usual bioequivalence region. Still, this does not guarantee that average bioequivalence follows from population bioequivalence, especially for highly variable drugs.

In general, it is required to conduct bioequivalence trials in a cross-over design, hence we will extend in the following our testing procedure to this design. For this it is necessary to discuss cross-over designs in a completely nonparametric framework, including the definition of corresponding effects, such as the main effect and period effects. In order to make ideas more transparent, we consider in the following only a 2×2 cross-over design involving two periods and two sequences, whereas extensions to higher order designs are outlined briefly in Section 4.

3.3 Population bioequivalence for cross-over designs

The need to allow for nonparametric assumptions in cross-over designs has been recognized by many authors (see for example [37,48,53]). However, these authors consider only the problem of testing the presence of a treatment effect, adjusting for possible period effects by using modifications of standard nonparametric point hypothesis tests based on signs or ranks. These can be used to assess average bioequivalence using corresponding confidence intervals. The nonparametric assessment of population bioequivalence using interval hypotheses remains, however, an interesting open problem, which we consider now for the case of a 2×2 cross-over trial.

Let in the following Y_{ijk} denote the response of the i^{th} subject in the k^{th} sequence ($k = 1, 2$) at the j^{th} period ($j = 1, 2$). Hence, in the first sequence we observe n_1 bivariate i.i.d. observations

$$(Y_{111}, Y_{121}), \cdots, (Y_{n_1 11}, Y_{n_1 21}) \sim H_1(\cdot, \cdot),$$

and independently in the second sequence n_2 independent observations

$$(Y_{122}, Y_{112}), \cdots, (Y_{n_2 22}, Y_{n_2 12}) \sim H_2(\cdot, \cdot),$$

where we assume that $H_1(H_2)$ has marginals $F_1(F_2)$ and $G_1(G_2)$. In order to

simplify notation, in the bivariate sample for H_2 the order of the observations from Table 1 is interchanged. This allows to interpret the first (second) component of all samples as subjects with treatment T (R). Further Y_{i1k} and Y_{i2k} ($k = 1, 2$) are dependent, since observations are drawn from the same subject.

Table 1
The 2×2 cross-over design.

	Period 1	Period 2
Sequence 1	Treatment T (F_1)	Treatment R (G_1)
	Y_{111}	Y_{121}
	$\vdots$	$\vdots$
	$Y_{n_1 11}$	$Y_{n_1 21}$
Sequence 2	Treatment R (G_2)	Treatment T (F_2)
	Y_{112}	Y_{122}
	$\vdots$	$\vdots$
	$Y_{n_2 12}$	$Y_{n_2 22}$

Similar to the parametric situation we have to exclude a carry-over effect, which in practice has to be guaranteed by a sufficient wash-out period. Note that in this nonparametric framework we cannot identify carry-over effects as fixed *linear* effects as it is custom in a parametric model (cf. [6]). Therefore, in order to identify the main effects in a nonparametric 2×2 cross-over trial, we make the following basic assumption.

Assumption A: *We assume that the marginals occurring in the second period (G_1 and F_2) are solely generated by the direct drug effect, possibly together with an effect which is solely generated by the period.*

Hence, we admit that the period in which the drug is administered may have an effect on the outcome. This will be denoted as a *period effect*.

Definition 6 *Assume that the basic Asumption A holds. In the nonparametric 2×2 cross-over design we say that a period effect is present, if and only if $H_1 \neq H_2$.*

We will briefly explain the connections and differences to the parametric model under normality which is used by the FDA [19]. Here the assumption of no or equal carry-over effects is used, where equal carry-over effects are supposed to enter into the period effects. The model for the case of a 2×2 cross-over trial

is given by

$$Y_{ijk} = \mu + F_l + W_{jk} + S_{ikl} + \epsilon_{ijk},$$

where μ is the overall mean, F_l is the fixed effect of the lth formulation ($l = T, R$, with $F_T + F_R = 0$), W_{jk} are fixed period, sequence, and interaction effects ($\sum_k W_{jk} = 0$), S_{ikl} is the random effect of the ith subject in the kth sequence under formulation l and (S_{ikT}, S_{ikR}), $i = 1, ..., n_k, k = 1, 2$, are independent and identically distributed bivariate normal random vectors with mean 0 and covariance matrix

$$\begin{pmatrix} \sigma_{BT}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{pmatrix}, \tag{12}$$

the ϵ_{ijk} are independent random errors with distribution $\mathcal{N}(0, \sigma_{Wl}^2)$, and the S_{ikl} and the ϵ_{ijk} are mutually independent.

Under parametric assumptions it would be possible to determine whether a carry-over effect is present. However, in this case no further analysis can be performed. In the nonparametric setup a mathematical separation between carry-over and period effects is not possible. Therefore, we have excluded carry-over effects a priori in Assumption A.

A second important distinction between the parametric and the nonparametric case is that a reduction by sufficiency allows in the parametric setup to base the analysis on individual log ratios with variance $\tau_T^2 + \tau_R^2$. This simplifies the model substantially, because we end up with a homoscedastic error structure (cf. Hauschke et al. [33] for a discussion).

Note, however, that it is not possible in a nonparametric framework to base a priori the analysis on one-dimensional within-subject quantities, such as the individual log ratios. Hence, the assumption of homogeneous variances is not valid in general. This is highlighted in the next section, where tests for the nonparametric bioequivalence problem are given which require the information of the full two dimensional data vector without the possibility of a one-dimensional reduction of the observations drawn from each individual. In particular, this illustrates that violation of a normal model may have particularly drastic consequences in bioequivalence studies as pointed out by Chow & Tse [7].

Now we are ready to define nonparametric bioequivalence hypotheses in a 2×2 cross-over design.

Definition 7 *Assume the nonparametric* 2×2 *cross-over model under As-*

sumption A. Then, population bioequivalence is defined as similar marginals under T and R in each sequence, respectively, i.e.

$$\gamma_{\beta,p} := \frac{1}{2}\{\gamma_\beta(F_1, G_2) + \gamma_\beta(F_2, G_1)\} \leq \Delta_0^2, \tag{13}$$

for a fixed bound Δ_0.

We would like to comment briefly on this definition. Because period effects (in the sense of Definition 6) cannot be excluded a priori, bioequivalence has to be defined for each sequence separately, i.e. we cannot assume in general that $F_2 = F_1$ and $G_1 = G_2$. Observe, however, that in the case of no period effect we have

$$\gamma_{\beta,p} = \gamma_\beta(F, G), \tag{14}$$

where $F_1 = F_2 = F$ and $G_1 = G_2 = G$. This allows to perform a different and more efficient analysis for the case of no period effect, as we will see in the next section.

Remark 8 *(*Determining the equivalence bound*). In the example in Section 3.5 we will define the equivalence limit in accordance with the average bioequivalence criterion in a normal homoscedastic setup with log-transformed data, i.e. $\Delta_0 = \log(1.25)$. Note that this is based on the fact that in a location model Mallows distance reduces to the difference of the means in both treatment groups (cf. Section 2.1). This is certainly a conservative choice, because no additional tolerance limit for deviations of the variances, say, is admitted.*

3.4 Test statistics for population bioequivalence

We present now two different tests for population bioequivalence. One is appropriate when no period effects are present, while the other adjusts for period effects.

No period effects. We will first discuss the case of no period effect, i.e. we assume $H_1 = H_2$. This allows to reduce to the case of $n = n_1 + n_2$ i.i.d. observations $Z_i = (X_i, Y_i), i = 1, \ldots, n_1 + n_2$, where

$$Z_i = (X_i, Y_i) = \begin{cases} (Y_{i11}, Y_{i21}) & i = 1, \ldots, n_1 \\ (Y_{i22}, Y_{i12}) & i = n_1 + 1, \ldots, n_1 + n_2 \end{cases},$$

such that $Z_i \sim H(\cdot,\cdot)$ with marginals F and G. In this case the bioequivalence measure $\gamma_{\beta,p}$ in (13) reduces to $\gamma_\beta(F,G)$ (cf. (14)), which can simply be estimated by $\gamma_\beta(F_n,G_n)$. Now we draw B, say, bootstrap samples of size n from the bivariate observed data $\{Z_i, i=1,\ldots,n\}$ and calculate the corresponding empirical marginal distributions $F^*_{n,b}$ and $G^*_{n,b}$, $b=1,...,B$. Thus, we obtain B bootstrapped test statistics $T^b_{\Delta_0} := T_{\Delta_0,\beta}(F^*_{n,b},G^*_{n,b})$, $b=1,...,B$ (see (6)).

Period effects. If a period effect cannot be excluded, separate estimation of $\gamma_\beta(F_1,G_2)$ and $\gamma_\beta(F_2,G_1)$ becomes necessary. The appropriate test statistic is therefore

$$\begin{aligned} &T_{\Delta_0,\beta}(F_{1,n_1},F_{2,n_2},G_{1,n_1},G_{2,n_2}) \\ &= \sqrt{\tfrac{n_1 n_2}{n_1+n_2}}\Big[\tfrac{1}{2}(\gamma_\beta(F_{1,n_1},G_{2,n_2})+\gamma_\beta(F_{2,n_2},G_{1,n_1}))-\Delta_0^2\Big]. \end{aligned} \tag{15}$$

We now bootstrap from each sequence separately, i.e. for the b^{th} bootstrap test statistic we draw n_1 times from the bivariate data $\{(Y_{i11},Y_{i21}), i=1,\ldots,n_1\}$ and n_2 times from $\{(Y_{i22},Y_{i12}), i=1,\ldots,n_2\}$, yielding the corresponding estimators $F^*_{k,n_k,b}$ and $G^*_{k,n_k,b}$, $k=1,2$, which can be plugged in to obtain the bootstrapped test statistics. We mention finally that, of course, similar results to Theorem 1 can be proved for the case of period effects in exactly the same way using the above stratified bootstrap procedures.

It remains to decide which test is appropriate for data at hand. For the data analyst it is difficult to know in advance if there are period effects present in the data. One simplistic approach is to first test for period effects and then proceed testing for population bioequivalence depending on the outcome. However this two-stage approach might lead to unsatisfactory answers if the two tests are correlated (see Senn [48], p. 52-54, and Freeman [21] for the problem of testing for carry-over effects in a standard normal cross-over analysis). Without further investigations it is too early to advise the use of such a two-stage approach. However, if one is interested in conducting such a two-stage approach, we suggest using the methodology adopted in this article, i.e. to test the hypothesis $H : \frac{1}{2}[\gamma_\beta(F_1,F_2)+\gamma_\beta(G_1,G_2)] > \Delta_0^2$, but where possibly a different value for the hypothesis boundary is used compared to the second test.

Finally, we would like to give a guide for the choice of the trimming bound β. As the sample size increases, trimming up to 10% of the total sample size was always found to improve the accuracy of the bootstrap. For the most usual setting in a standard bioequivalence trial we have $n_1 = n_2 = 12$. Here we suggest trimming on both tails by 1 observation, i.e. $\beta = \frac{1}{12}$. This is also in accordance with Chow & Tse's (1990) finding [7], who report on drastic consequences of an outlier for the evaluation of bioequivalence.

In summary, the nonparametric approach allows to assess bioequivalence in a cross-over design without making the normal error assumption or the additivity assumption of the period and treatment effects.

3.5 Example: Vasoactive bioavailability study

The following data on a vasoactive drug were kindly supplied by a pharmaceutical company, and the log-transformed data are given in Table 2.

Table 2
The log-transformed data from the bioavailability study on a vasoactive drug

Sequence 1		Sequence 2		Sequence 1		Sequence 2	
Per. 1	Per. 2	Per. 1	Per. 2	Per. 1	Per. 2	Per. 1	Per. 2
Ref.	Test	Test	Ref.	Ref.	Test	Test	Ref.
3.6188	3.3569	3.6836	3.5172	3.4770	3.6641	3.9310	3.5149
3.3540	3.3572	4.2773	3.9694	3.1708	3.6026	3.2648	3.1881
3.2128	3.6398	3.2856	3.1885	3.0632	3.2864	4.0409	3.5149
2.9158	3.2414	3.4421	3.1629	3.1679	2.0248	3.3186	3.2482
3.3306	3.7818	3.0198	2.9132	3.3996	3.7490	3.1716	3.5189
2.8543	3.7632	3.1430	2.4570	3.1554	3.1062	3.9867	3.2993
2.1964	2.7847	2.9153	2.6331	3.5435	3.5204	3.4659	3.5815

In the underlying study a generic and a reference formulation of a vasoactive ingredient were compared. Figure 1 gives the histograms and kernel density estimators for the period differences and the empirical inverse distribution functions. An outlier in Sequence 2 can be seen, which is mostly responsible for the differences in the empirical distributions of the period differences.

The results of tests on normality of the period differences are presented in Table 3. They show some evidence against the normal assumption of the period differences for Sequence 2.

Table 3
Tests of normality for the period differences of the (log-)vasoactive drug data

Period	Shapiro-Wilks		Cramer-von Mises	
Differences	Statistic	p-value	Statistic	p-value
Sequence 1	.9630	.7294	.0377	.7173
Sequence 2	.8791	.0561	.0883	.1680

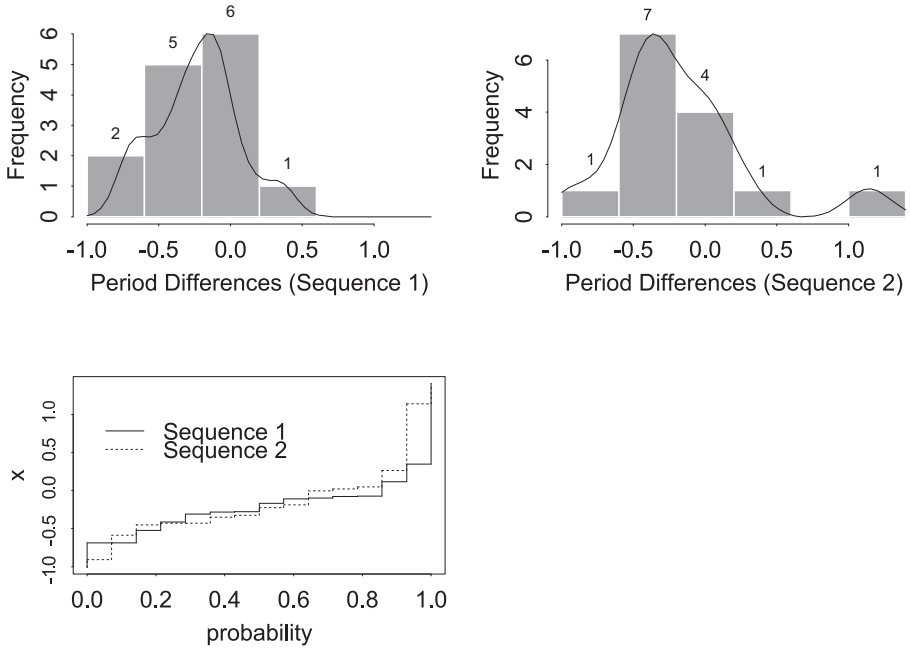

Fig. 1. Histograms of period differences (− nonparametric kernel approximation) and empirical quantile functions of the log-transformed vasoactive drug data

Figure 2 gives histograms for the treatment and reference measurements assuming no period effects, showing skewness and hence non-normality of the test and reference measurements. Population equivalence cannot be expected, since the empirical distributions differ, as seen in the empirical quantile function plot. Again, this indicates that the results of a classical approach might be misleading, since the underlying assumptions are violated in this example.

For comparison, we present now the results of the parametric approach and of the nonparametric approach in Table 4. The parametric analysis shows that there is strong evidence for a treatment effect, while there is no evidence for a period or a carry-over effect. Average bioequivalence cannot be shown with Schuirmann's [47] TOST procedure. However, if the FDA [19] population bioequivalence criterion PBC is used, i.e.

$$H : PBC = \frac{(F_T - F_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2}{\max(\sigma_{TR}^2, \sigma_0^2)} > \theta_p, \tag{16}$$

with the commonly used values $\theta_p = ((\ln 1.25)^2 + \varepsilon_p)/\sigma_0^2$, $\varepsilon_p = 0.02$, $\sigma_0 = 0.2$, then population bioequivalence can be concluded in this example (using the procedure suggested by Lee et al. [38]). Here $\sigma_{Tl}^2 = \sigma_{Bl}^2 + \sigma_{Wl}^2$, $l = R, T$ (cf. (12)). This is mainly due to the fact that we have highly variable drugs in this example, for which the use of the PBC criterion is questionable.

For the nonparametric analysis, we first investigate whether there is a period effect. The effect of the outlier in the period differences for Sequence 2 (see

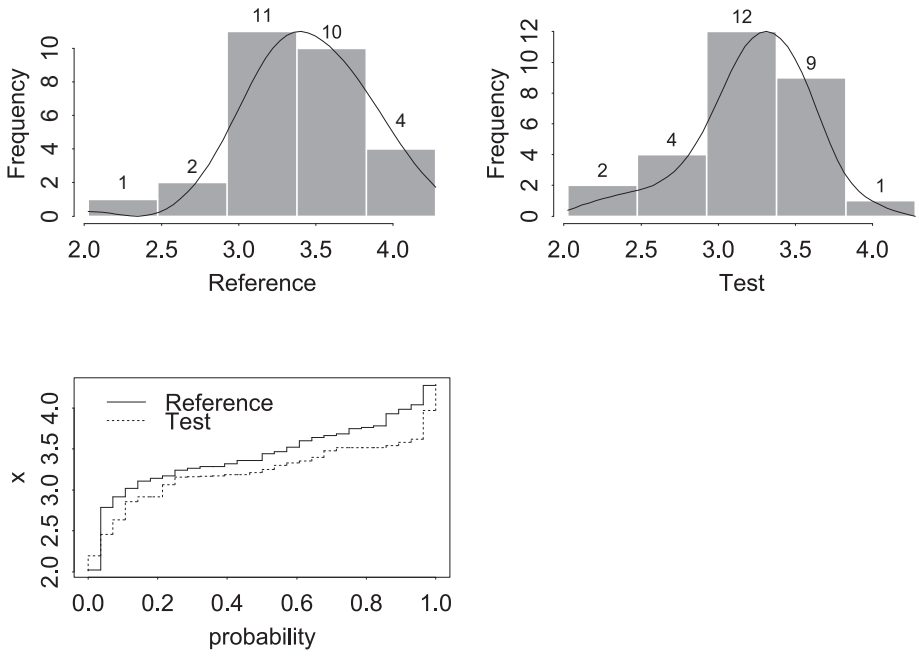

Fig. 2. Histograms of reference and test measurements ignoring period effects (– kernel approximation) and empirical quantile functions of the log-transformed vasoactive drug data

Figure 1) can be clearly seen when no trimming is assumed. In this case evidence for a period effect is present (p-value$>$ .33). Regarding the treatment under the assumption of no period effect, the BC_a method cannot decide in favor of bioequivalence, thus indicting the presence of a treatment effect (as in the parametric analysis of ABE). In addition, trimming has little effect on the assessment of a treatment effect in this case. If we are allowing for the presence of a period effect, we still cannot reject the bioequivalence hypothesis, regardless whether trimming or no trimming is used. However, in case of trimming the p-value is considerably lower than without trimming, reflecting the effect of the outlier in the period differences for Sequence 2.

Finally, the estimated Mallows distance $\Gamma_\beta(F_n, G_n)$ for the treatment effect (allowing for no period effect) is estimated as .2044, with estimated bootstrapped standard error .0065, using $B = 2000$ and $\beta = \frac{1}{14}$. To interpret an estimated Mallows distance of 0.2044, recall that in a location-scale family with equal variances this would correspond to an absolute mean difference of .2044 (cf. [42]). In comparison, for the parametric analysis the absolute treatment effect $|F_T - F_R|$ is estimated as 0.202, with estimated standard error 0.07.

Table 4
Classical and nonparametric analyses for the vasoactive drug data ($\Delta_0 = \ln(1.25)$)

Effect	Parametric Analysis	Nonparametric Analysis ($B = 2000$)	
		$\beta = \frac{1}{14}$	$\beta = 0$
Carry-Over	$H : C_T = C_R = 0$ p-value = .40		
Period	$H : P_1 = P_2 = 0$ p-value = .70	$H : \frac{1}{2}[\gamma_\beta(F_1, F_2) + \gamma_\beta(G_1, G_2)] > \Delta_0^2$ BC_a: p-value $< .01$	BC_a: p-value = .34
Treatment (no carry-over)	$H : F_T = F_R = 0$ p-value = .01		
	TOST $H : \lvert F_T - F_R \rvert > \Delta_0$ p-value = .41	$H : \Gamma_\beta(F, G) > \Delta_0$ (no period effects) BC_a: p-value = .26	BC_a: p-value = .30
	FDA PBE $H : PBC > \frac{\Delta_0^2 + 0.02}{0.04}$ p-value $< .01$	$H : \frac{1}{2}[\gamma_\beta(F_1, G_2) + \gamma_\beta(F_2, G_1)] > \Delta_0^2$ (allows for period effects) BC_a: p-value = .19	BC_a: p-value = .73

4 Further remarks and extensions

Remark 9 (Higher order cross-over designs). *The proposed method can also be applied in higher order cross-over designs, i.e. when either the number of periods or sequences is larger than two. To this end period effects have to be defined similarly as in Definition 6. The bootstrap algorithm and the statistic $\hat{\gamma}_\beta$ have to be modified due to the constraint induced by the periods. As an example, consider a dual two-sequence three-period cross-over design, where in Sequence I n observations with $T - R - R$ are available and in Sequence II n observations with $R - T - T$. Here we have in the first sequence a three-dimensional distribution H^1 with marginals F_1^1, G_1^1, G_2^1, and in the second sequence H^2 with marginals G_1^2, F_1^2, F_2^2 (cf. Table 5).*

We will illustrate the analysis in the simplest case, if we exclude period effects. This would mean that all two-dimensional marginals from sequence I and II, respectively, are equal. The test statistic is

$$\hat{\gamma}_\beta = \frac{1}{(1-2\beta)3n} \sum_{i=[3n\cdot\beta]+1}^{3(n-[n\cdot\beta])} |X_{(i)} - Y_{(i)}|^2 + O_p(\tfrac{1}{n}),$$

where $X_{(1)}, \ldots, X_{(3n)}$ and $Y_{(1)}, \ldots, Y_{(3n)}$ refer to the ordered samples of the combined $3n$ observations under T and under R, respectively (cf. also (5)). Bootstrap samples have to be drawn from each sequence separately. Note that under presence of period effects, the test statistic has to be split into the sum of three parts, i.e. we have to estimate the Mallows distance in each period and define a bioequivalence measure as

$$\gamma_{\beta,p} = 1/3 \left\{ \gamma_\beta(F_1^1, G_1^2) + \gamma_\beta(G_1^1, F_1^2) + \gamma_\beta(G_2^1, F_2^2) \right\}.$$

Table 5
The nonparametric 2×3 cross-over design.

	Period I	Period II	Period III
Sequence I (H^1)	F_1^1	G_1^1	G_2^1
Sequence II (H^2)	G_1^2	F_1^2	F_2^2

Remark 10 (Calculation of power and sample size). *Most important for the calculation of the sample size is the power under equality of the c.d.f.s F and G, i.e. $\gamma_\beta(F, G) = 0$. Note that in this case asymptotic normality as in Theorem 1 fails to hold. Here it can be shown that*

$$n\hat{\gamma}_\beta \stackrel{\mathcal{D}}{\Longrightarrow} \int_\beta^{1-\beta} \left(\frac{\mathbb{B}_1(t) - \mathbb{B}_2(t)}{f \circ F^{-1}(t)} \right)^2 dt, \tag{17}$$

where $\mathbb{B}_1(t)$ and $B_2(t)$ are Brownian bridges with $Cov[\mathbb{B}_1(s), \mathbb{B}_2(t)] = (F^{-1}(s), G^{-1}(t)) - st$. Observe that this holds also true for $\beta = 0$, provided $\int x^2 F(dx) < \infty$ and $\int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))} dt < \infty$ (cf. [10] Lemma 5.3.2). Furthermore, note that the scaling factor in (17) is n and not $n^{1/2}$ as in Theorem 1. For sample size calculations, F has to be specified in advance to allow the use of the right-hand side of (17).

Acknowledgements

The authors would like to thank D. Hauschke and A. Steland for pointing out some references. Helpful comments of W. Stute and M. Vogt are gratefully

acknowledged. We are thankful to S. Chow for providing the data of Example 2. Finally, the authors would like to thank a referee for valuable comments which lead to an improvement of the article. The work of A. Munk and G. Freitag was supported by the Deutsche Forschungsgemeinschaft (TR 471/1) and C. Czado was supported by the Sonderforschungsbereich 386 *"Statistische Analyse Diskreter Strukturen"*.

References

[1] P. Bauer and M. M. Bauer, Testing equivalence simultaneously for location and dispersion of two normally distributed populations, *Biom. J.* **36** (1994) 643–660.

[2] R. L. Berger and J. C. Hsu, Bioequivalence trials, intersection-union tests and equivalence confidence sets (with discussion), *Statist. Sci.* **11** (1996) 283–319.

[3] P. J. Bickel and D. A. Freedman, Some asymptotic theory for the bootstrap, *Ann. Statist.* **9** (1981) 1196–1217.

[4] P. Bickel and M. J. Wichura, Convergence criteria for multiparameter stochastic processes and some applications, *Ann. Math. Statist.* **42** (1971) 1656–1670.

[5] S.-C. Chow, Individual Bioequivalence - a review of th FDA draft guidance, *Drug inf. J.* **33** (1999) 435–444.

[6] S.-C. Chow and J.-P. Liu, *Design and Analysis of Bioavailability and Bioequivalence Studies* (Marcel Dekker, New York, 1992).

[7] S.-C. Chow and S. K. Tse, Outlier detection in bioavailability/bioequivalence studies, *Statist. Med.* **9** (1990) 549–558.

[8] CPMP (Committee for Proprietary Medicinal Products, Working Party on Efficacy of Medicinal Products), *Draft Guideline: Biostatistical methodology in clinical trials in applications for marketing authorization for medical products* (1993).

[9] CPMP (Committee for Proprietary Medicinal Products, Working Party on Efficacy of Medicinal Products), *Note for Guidance on the Investigation of Bioavailability and Bioequivalence* (July 2001).

[10] M. Csörgö and L. Horvath, *Weighted Approximations in Probability and Statistics* (Wiley & Sons, Chichester, 1993).

[11] M. Csörgö and P. Révész, *Strong Approximations in Probability and Statistics* (Academic Press, New York, 1981).

[12] C. Czado and A. Munk, Bootstrap methods for the nonparametric assessment of population bioequivalence and similarity of distributions, *J. Statist. Comp. Simul.* **68** (2001) 243–280.

[13] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán and J. M. Rodríguez-Rodríguez, Tests of goodness of fit based on the L_2-Wasserstein distance, *Ann. Statist.* **27** (1999) 1230–1239.

[14] E. del Barrio, E. Giné and C. Matrán, Central limit theorems for the Wasserstein distance between the empirical and the true distributions, *Ann. Probab.* **27** (1999) 1009–1071.

[15] E. del Barrio, J. A. Cuesta-Albertos and C. Matrán, Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. (With comments), *Test* **9** (2000) 1–96.

[16] R. L. Dobrushin, Describing a system of random variables by conditional distributions, *Theory Probab. Appl.* **15** (1970) 458–486.

[17] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, London, 1993).

[18] FDA (Food and Drug Administration) *Guidance on Statistical Procedures for Bioequivalence Studies using a Standard Two-Treatment Crossover Design* (Office of Generic Drugs, Rockville, MD, July 1992).

[19] FDA (Food and Drug Administration) *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence* (U.S. Department of Health and Human Services, Center for Drug Evaluation and Research, January 2001).

[20] FDA (Food and Drug Administration) *Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products - General Considerations* (U.S. Department of Health and Human Services, Center for Drug Evaluation and Research, March 2003).

[21] P. R. Freeman, The performance of the two-stage analysis of two-treatment, two-period crossover trials, *Statist. Med.* **8** (1989) 1421–1432.

[22] G. Freitag, A. Munk and K. Hoffmann, Vergleich zweier Messmethoden mit einem Goldstandard am Beispiel der 20MHz-Sonographie und der klinischen Palpation zur Dickenbestimmung von pigmentierten Tumoren der Haut, *Ultrasch. Med.* **24** (2003) 184–189.

[23] G. Freitag, A. Munk and M. Vogt, Assessing structural relationships between distributions - a quantile process approach based on Mallows distance, in: *Recent Advances and Trends in Nonparametric Statistics* Editors: M. G. Akritas, D. N. Politis (Elsevier Science B. V., Amsterdam, 2003).

[24] R. Gill, Non- and semi-parametric maximum likelihood estimators and the von Mises method (part 1), *Scand. J. Statist.* **16** (1989) 97-128.

[25] A. L. Gould, Discussion of individual bioequivalence by M.L. Chen, *J. Biopharm. Statist.* **7** (1997) 23–31.

[26] A. L. Gould, A practical approach for evaluating population and individual bioequivalence, *Statist. Med.* **19** (2000) 2721–2740.

[27] Z. Govindarajulu, A class of asymptotically distribtution free test procedures for equality of marginals under multivariate dependence, *Amer. J. Math. Manag. Sci.* **15** (1995) 375–394.

[28] Z. Govindarajulu, A class of asymptotically distribution free tests for equality of marginals in multivariate populations, *Math. Meth. Statist.* **6** (1997) 92–111.

[29] O. Guilbauld, Exact inferences about the within subject variability in 2×2 crossover trials, *J. Amer. Statist. Assoc.* **88** (1993) 939–946.

[30] W. W. Hauck and S. Anderson, Types of bioequivalence and related statistical considerations, *Int. J. Clin. Pharmacol., Ther. Toxic.* **30** (1992) 181–187.

[31] W. W. Hauck, F. Y. Bois, T. Hyslop, L. Gee and S. Anderson, A parametric approach to population bioequivalence, *Statist. Med.* **16** (1997) 441–454.

[32] D. Hauschke, V. W. Steinijans and E. Diletti,

A distribution-free procedure for the statistical analysis of bioequivalence studies, *Int. J. Clin. Pharmacol., Ther. Toxic.* **28** (1990) 72–78.

[33] D. Hauschke, V. Steinijans and L. A. Hothorn, A note on Welch's approximate t-solution to bioequivalence assessment, *Biometrika* **83** (1997) 236–237.

[34] D. Hauschke, and V. W. Steinijans, The U.S. draft guidance regarding population and individual bioequivalence approaches: comments by a research-based pharmaceutical company, *Statist. Med.* **19** (2000) 2769–2774.

[35] D. J. Holder and F. Hsuan, Moment-based criteria for determining bioequivalence, *Biometrika* **80** (1993) 835–46.

[36] A. Janssen, Nonparametric bioequivalence tests for statistical functionals and their efficient power functions, *Statist. Decis.* **18** (2000) 49–78.

[37] B. Jones and M. G. Kenward, *Design and analysis of cross-over trials* (Chapman & Hall, London, 1988).

[38] Y. Lee, J. Shao and S.-C. Chow, Modified large-sample confidence intervals for linear combinations of variance components: estimation, theory, and application, *J. Amer. Statist. Assoc.* **99** (2004) 467–478.

[39] J. P. Liu and C. S. Weng, Detection of outlying data in bioavailability / bioequivalence studies, *Statist. Med.* **10** (1991) 1375–1389.

[40] C. L. Mallows, A note on asymptotic joint normality, *Ann. Math. Statist.* **43** (1972) 508–515.

[41] A. Munk, Equivalence and interval testing for Lehmann's alternative, *J. Amer. Statist. Assoc.* **91** (1996) 1187–1197.

[42] A. Munk and C. Czado, Nonparametric validation of similar distributions and assessment of goodness of fit, *J. Roy. Statist. Soc., Ser. B* **60** (1998) 223–241.

[43] M. J. Podgor and J. L. Gastwirth, Efficiency robust rank tests for stratified data, in: *Research Developments in Probability and Statistics. Festschrift in honor of Madan L. Puri* Editors: E. Brunner & M. Denker (VSP International Science Publishers, 1996).

[44] L. Rüschendorf and S. T. Rachev, A characterization of random variables with minimum L^2-distance, *J. Mult. Anal.* **32** (1990) 48–54.

[45] R. Schall, Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar, *Biometrics* **51** (1995) 615–626.

[46] R. Schall and H. G. Luus, On population and individual bioequivalence, *Statist. Med.* **12** (1993) 1109–1124.

[47] D. J. Schuirmann, A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability, *J. Pharmacok. Biopharm.* **15** (1987) 657–680.

[48] S. Senn, *Cross-Over Trials in Clinical Research* (John Wiley & Sons, New York, 1993).

[49] J. Shao, J. Kübler and I. Pigeot, Consistency of the bootstrap procedure in individual bioequivalence, *Biometrika* **87** (2000) 573–585.

[50] L. B. Sheiner, Bioequivalence revisited, *Statist. Med.* **11** (1992) 1777–1788.

[51] A. Steland, Bootstrapping rank statistics, *Metrika* **47** (1998) 251–297.

[52] A. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes - With Applications to Statistics* (Springer, New York, 1996).

[53] E. F. Vonesh and V. M. Chinchilli, *Linear and Nonlinear Models for the Analysis of Repeated Measurements* (Marcel Dekker Inc., New York, 1997).

[54] W. Wang, Optimal unbiased tests for equivalence in intrasubject variability, *J. Amer. Statist. Assoc.* **92** (1997) 163–1170.

[55] L. N. Wasserstein, Marcov processes with countable state space describing a large system of automata, *Probl. Inform. Transm.* **5** (1969) 47–52.

[56] S. Wellek, A new approach to equivalence assessment in standard comparative bioequivalence trials by means of the Mann-Whitney statistics, *Biom. J.* **38** (1996) 695–710.

[57] S. Wellek, Bayesian construction of an improved parametric test for probability-based individual bioequivalence, *Biom. J.* **42** (2000) 1039–1052.

[58] S. Wellek, On a reasonable disaggregate criterion of population bioequivalence admitting of resampling-free testing procedures, *Statist. Med.* **19** (2000) 2755–2767.