Higher-Order Crossover Design Outlier Detection

Vesna Lužar-Stiffler[1,2]
Charles Stiffler[2]
[1]*University of Zagreb, University Computing Centre*
[2]*CAIR Research Centre, Zagreb, Croatia*
vluzar@srce.hr, charles.stiffler@cair-center.hr

Abstract. *The purpose of this article is to introduce a new procedure for detecting subject outlier(s) in data from a fully replicated crossover design. The suggested procedure is an adjustment to a test based on the order statistics of the two-sample Hotelling T^2, introduced by Liu and Weng. Results from an empirical power study indicate that the proposed procedure is, on average, as powerful as, for example, the equivalent procedure for detecting outliers in data from a standard crossover design.*

Keywords. Crossover design, higher-order, bioequivalence, bioavailability, outlier detection, Hotelling T^2, Monte Carlo distribution, simulation, power study.

1. Introduction

In clinical trials there are two different kinds of study objectives: a superiority hypothesis that the new treatment is more effective than the active control treatment and a non-inferiority (or equivalence) hypothesis that the new (experimental) treatment is clinically no worse than (or equivalent to) the active control within a defined margin or range. The non-inferiority (equivalence) objective can be very attractive because of the intent to develop an alternative treatment that is less toxic, less costly and/or easier to administer, yet has a similar effect.

Bioequivalence studies are based on bioequivalence of different formulations of the same treatment, usually taken to mean equivalence with respect to rate and extent of drug absorption. The Food and Drug Administration (FDA) pre-specified the rule that bioequivalence studies must have at least 80% power of detecting a 10% difference between the parameters of interest. It was also suggested that 90% confidence intervals (CI) be used in such circumstances. Two sided tests are appropriate, since the objective is to find if the new treatment is better or worse.

Equivalence studies are different from other clinical studies in that the desired inference in equivalence studies, instead of the usual "significant difference", is "practical difference". Therefore, the null hypothesis tested in equivalence studies is "The test formulation is NOT equivalent to the reference formulation" whereas the alternative hypothesis is "The test treatment IS equivalent to the reference formulation"([1], [10], [11]).

In bioequivalence studies, formulations are compared with respect to pharmacokinetic characteristics (e.g., AUC, CMAX, TMAX). Typically, lognormality is assumed, i.e, prior to the analysis, logarithmic transformation is applied to the pharmacokinetic responses.

Although designs for more than 2 formulations have been studied and applied ([3], [10]), the presentation here is limited to the more common situation of 2 formulations. The usual approach to testing average bioequivalence for 2 formulations is the "standard" 2x2 crossover design i.e., each subject is randomly assigned to one of the two sequences . Subjects in sequence 1 receive (in the first period) the test formulation, and after a so-called "wash-out" period, the reference formulation, whereas for subjects in sequence 2, it is presented vice-versa. Besides a standard 2x2 crossover design, the FDA recommends (specifically for testing individual bioequivalence) higher-order crossover designs ([2],[3],[13],[14]). In this paper we will address a frequently seen higher-order design (presented in Table 1) which here will be called a (fully) replicated crossover design. As indicated by the name, this design simply duplicates the original crossover study using the same subjects, response measures, treatments, sequences, and procedures. A "T" and an "R" are commonly used to denote the test and reference formulation, respectively.

Table 1. A fully replicated crossover design (TRTR, RTRT)

Period	Replication	Sequence 1	Sequence 2
1	1	T	R
2		R	T
3	2	T	R
4		R	T

A common problem in bioavailability studies is the occurrence of extremely large or small (i.e., outlying) observations. These observations may influence the conclusion drawn regarding bioequivalence. A confidence interval estimated excluding an outlying subject may lead to a different conclusion regarding bioequivalence (according to FDA recommended criterion for average bioequivalence) than a confidence interval estimated with all subjects included. Therefore, several tests have been proposed for detecting outliers in studies based on the standard crossover design ([4],[8],[15]). However, no statistical test has as yet been specifically proposed for data from the replicated crossover design.

Therefore, the aim of this research was to develop a test that could be used (in a fully replicated crossover design) to detect whether subjects (one or more) are probably an outliers.

In the next section we first introduce a general statistical model for a fully replicated 2x2 crossover design. In Section 3 we provide a brief definition and classification of outliers that typically appear in bioequivalence studies. Existing procedures for outlier detection in data from a standard crossover design are described in Sections 4. The proposed procedure for identifying any number of outlying subjects in data coming from a replicated crossover design is introduced in Section 5. Results of an empirical power study, based on a simulation experiment designed to compare the power of the newly proposed procedure with the equivalent procedure for standard crossover design are presented in Section 6.

2. Statistical Model

The statistical model for a fully replicated 2 sequence and 2 period crossover design (2x2 design) comparing 2 formulations (T versus R) is a mixed model that can be expressed as follows:

The response Y_{ijkl} of the ith subject in the lth period of the kth sequence, treated by jth formulation, can be expressed in the matrix form as:

$$Y = X\beta + Zu + \varepsilon, \qquad (1)$$

where:

Y is the vector of the log-transformed bioavailability measures,

X is fixed effects (formulation, sequence, period) design matrix,

β are fixed effect parameters,

Z is the random effects design matrix,

u are random effect parameters, distributed MVN(0,G),

ε is the (within subject) random error in observing Y, distributed MVN(0,R),

n_k is the number of subjects in the kth sequence;

μ= overall mean; μ_A and μ_B are treatment means.

The model assumes that the subject-specific means μ_{Ti} and μ_{Ri} come from a distribution with population means μ_T and μ_R, and between-subject variances σ_{bT}^2 and σ_{bR}^2, respectively. The model allows for a correlation, ρ, between μ_{Ti} and μ_{Ri}.

For a given subject, the observed data for the log-transformed bioavailability measure are assumed to be independent observations from distributions with means μ_{Ti} and μ_{Ri}, and within-subject variances σ_{WT}^2 and σ_{WR}^2. The total variances for each formulation are defined as the sum of the within- and between-subject components (i.e., $\sigma_{TT}^2 = \sigma_{WT}^2 + \sigma_{bT}^2$ and $\sigma_{TR}^2 = \sigma_{WR}^2 + \sigma_{bR}^2$).

3. Outlier definition

There are three different types of outliers (Chow, Liu [3]) in bioequivalence studies:

1. unexpected observations in the blood or plasma concentration – time curve,
2. extremely large or small observations within a given formulation, and
3. unusual subjects who exhibit extremely high or low bioavailability relative to the reference formulation.

In this paper we are concerned only with the last type of outlier: outlying subjects. (Also, FDA Guidance [13] points out that … “the most important type of outlier is the within-subject outlier, where one subject or a few subjects differ notably from the rest of the subjects with respect to a within-subject T-R comparison.”). Each outlier type can be visualized using different types of plots, and may be detected using different statistical testing procedures, under the standard or replicated crossover design.

Although a lot of literature is available for the identification of outliers and influential observations in the context of the regression

model, little has appeared with regard to detection of outlying subjects in bioequivalence studies. It should be noted that the standard tests (such as Lund's method) for outliers in linear regression setting (with independent observations) is not appropriate for data from the crossover design where the responses from the same subject are correlated.

4. Statistical Tests for Outlying Subjects in a Standard Crossover Design

The procedure proposed by Chow and Tse [4] is based on the asymptotic distribution of likelihood distance (LD) and estimated distance (ED). The disadvantage of this procedure is the fact that the sample size for a bioavailability study is often too small to apply asymptotic distributions of LD and ED. Therefore, Liu and Weng [8] proposed a procedure to detect subject outliers based on the order statistics of the two-sample Hotelling T^2, computed from each subject. Recently, Wang and Chow [15] developed a test procedure based on the likelihood function. All three procedures are proposed in the context of a standard crossover design and some adjustments are required if we want to apply them to the replicated design and the statistical model described above. Since the Chow and Tse [4] procedure is asymptotic and therefore of questionable applicability for relatively small sample sizes, we do not consider it here. However, one of the other two procedures can be applied/ adapted for the analysis of the data coming from the replicated crossover design, as will be described in the Section 5.

4.1 Liu and Weng Procedure for a Standard 2x2 Crossover Design

Under the assumption of no period effect and compound symmetry covariance structure of 2 responses observed on subject i, the model can be expressed as follows (see eg., Liu and Weng [8]) :

$Y_{ij} = \mu + F_j + \varepsilon_{ij} = \alpha_j + \varepsilon_{ij}$, i=1,2,…N, j = 1, 2,
where $\alpha_j = \mu + F_j$. (2)

We denote by $Y_i = (Y_{i1}, Y_{i2})$ a 2x1 vector of the responses to 2 formulations observed on subject i. Thus, Y_i are 2-dimensional multivariate normal (MVN) random vectors with mean vector $\boldsymbol{\alpha}$ and covariance matrix Λ, where

$\boldsymbol{\alpha} = (\alpha_1,\alpha_2)$, and

$$\Lambda=\mathrm{Cov}(Y_{ij},Y_{i'j'}) = \begin{cases} \sigma_j^2, & \text{if } i = i' \text{ and } j = j' \\ \sigma_{jj'}^2, & \text{if } i = i' \text{ and } j \neq j' \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

The hypotheses for outlying subjects caused by a location shift can be formulated as follows:
H_0: $Y_i \sim$ MVN (α, Λ), for all i = 1,2, …N,
vs.
H_a: $Y_i \sim$ MVN ($\alpha + \delta_i$, Λ), for at least one i.

The above hypothesis can be further decomposed into the following N subhypotheses:
H_{0i}: $Y_i \sim$ MVN ($\boldsymbol{\alpha}$, Λ),
vs.
H_{ai}: $Y_i \sim$ MVN ($\boldsymbol{\alpha} + \delta_i$, Λ), where i = 1,2, …N.

Since $H_0 = \cap_i H_{0i}$ and $H_a = \cup_i H_{ai}$ the sequential stepdown closed testing procedure can be applied to the above subhypotheses for detection of possible multiple outliers. The above hypotheses can be tested using two-sample Hotelling's T^2 statistics by comparing a sample that consists only of the ith subject with a sample of the remaining N-1 subjects. The two-sample Hotelling T^2 statistic for the ith subject can be computed as

$$T_i^2 = \frac{(N-2)D_i^2}{\left(\frac{N-1}{N} - D_i^2\right)}, \quad (4)$$

where

$$D_i^2 = (Y_i - \bar{Y})' A^{-1} (Y_i - \bar{Y}), \quad i = 1,2...N. \quad (5)$$

$\bar{Y}$ and A are the sample mean and matrix of the sum of squares and cross products computed from Y_1, Y_2, …Y_N, respectively.

Hotelling T^2 is invariant under any full-rank linear transformation. Therefore, the joint distribution of $\{T_i^2, i=1,2,\ldots N\}$ is independent of the unknown parameters $\boldsymbol{\alpha}$ and Λ.

The testing procedure proposed by Liu and Weng [8] is then (in case of 2 treatments) as follows:

Let $T_{(1)}^2,\ldots T_{(N)}^2$ be the order statistics of $T_1^2,\ldots T_N^2$, and $H_{0(i)}$ be the corresponding subhypothesis based on $T_{(i)}^2$. Also, let ($W_1^2,\ldots W_N^2$) be a vector of N Hotelling T^2 statistics computed from a sample of size N from a 2-dimensional multivariate normal distribution with mean 0 and covariance matrix I_2. We start with the order subhypothesis $H_{0(N)}$. Hypothesis $H_{0(i)}$ is rejected if

$$P\left\{\max_{1\le j\le i} W_j^2 > T_{(i)}^2\right\} < \alpha, \qquad (6)$$

provided that $H_{0(N)},\ldots H_{0(i+1)}$ are rejected at the α level of significance.

Since the joint distribution of order statistics of $\{T_i^2,\ i=1,2,\ldots N\}$ is rather complicated, the sampling distribution of $T_{(i)}^2$ under $H_{0(i)}$ must be evaluated empirically by Monte Carlo simulation using standard multivariate normal vectors (because it is independent of $\boldsymbol{\alpha}$ and Λ).

4.2 Wang and Chow Procedure for Standard Crossover Design

Wang and Chow [15] developed a test for influential subjects based on the following test statistic:

$$D_i = N(k-1)T_{1N} + NT_{2N}, \qquad (7)$$

where

$$T_{1N} = \frac{(e_i - \bar{e}_i 1)'(e_i - \bar{e}_i 1)}{\Sigma_s (e_s - \bar{e}_s 1)'(e_s - \bar{e}_s 1)}$$
$$T_{2N} = \frac{\bar{e}_i^2}{\Sigma_s \bar{e}_s^2} \qquad (8)$$

N is the number of subjects, k is the number of treatments, and $e_i = (e_{i1}, e_{i2}, \ldots e_{ik})$ is the residual vector for the ith subject in a standard (nonreplicated) crossover design.

A table of 90%, 95%, and 99% critical values of $\max_{1\le i\le N} D_i$ is given in [15] for various N (up to 50).

5. Liu and Weng Procedure Adjusted to Replicated 2x2 Crossover Design

To adjust the above procedure to the fully replicated 2x2 crossover design, we denote by $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})$ a 4x1 vector of the responses to 2 formulations repeatedly observed on subject i (i.e., Y_{i1}, Y_{i2} corresponds to the first replication (periods 1,2), while Y_{i3}, Y_{i4} corresponds to the second replication (periods 3,4)). Thus, Y_i are 4-dimensional multivariate normal (MVN) random vectors with mean vector $\boldsymbol{\alpha}$ and covariance matrix Λ, where

$\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\alpha_3,\alpha_4)$, and Λ has the structure assumed for the covariance matrix of Y, as defined in the statistical model Section above.

The procedure explained above is then applied to the case of replicated design in the following way:

Let $(W_1^2,\ldots W_N^2)$ be a vector of N Hotelling T^2 statistics computed from a sample of size N from a 4-dimensional multivariate normal distribution with mean 0 and covariance matrix I_4. Hypothesis $H_{0(i)}$ is rejected if

$$P\left\{\max_{1\le j\le i} W_j^2 > T_{(i)}^2\right\} < \alpha, \qquad (9)$$

provided that $H_{0(N)},\ldots H_{0(i+1)}$ are rejected at the α level of significance, where the sampling distribution of $T_{(i)}^2$ under $H_{0(i)}$ is empirically evaluated by Monte Carlo simulation using standard four-dimensional multivariate normal vectors.

Note: Application of the procedures described in Sections 4 and 5 requires that either special tables of quantiles or a program/ application for developing the appropriate Monte Carlo sampling distribution be made available to the researcher.

6. Empirical Power Study

In this section we compare the power of the new proposed (adjusted Liu and Weng) procedure for testing subject outliers in a replicated 2x2 crossover design against that of the original Liu and Weng procedure for a standard 2x2 crossover design.

The simulation research experiment consisted of generating random samples from both a standard and a replicated 2x2 crossover model based on a procedure used in Wang and Chow [15]. Random samples y_{ijr} are generated based on the following expression:

$$y_{ijr} = \gamma\,(z_{i0} + z_{ijr}) + \mu_j, \qquad (10)$$

where z_{i0} and z_{ijr} are i.i.d. standard normal, accounting for inter- and intra-subject variability, respectively (i=1,n; j=T,R; r (replication) =1 or r=1,2). Without loss of generality, the reference mean μ_R is set to 100. The test formulation is then set to the following levels: μ_T = 80 ,90, 100, 110, 125. The values for γ are selected in such a way that the coefficients of intra-subject variation for the reference formulation are 10%, 20%, and 30%, respectively.

Table 2. Percentage of correctly identifying the outlying subject

		Reference formulation intra-subject variation					
		10%		20%		30%	
N	p (%)	replicated	standard	replicated	standard	replicated	standard
14	10	92.70	98.60	9.60	15.90	1.10	0.90
	30	62.90	76.80	2.50	3.80	0.60	0.00
	50	19.00	26.40	0.50	0.80	0.20	0.10
	100	0.10	0.20	0.40	0.10	0.10	0.10
	130	10.40	9.90	2.50	3.00	1.20	0.80
	150	37.40	39.80	8.80	9.00	6.20	5.20
	200	93.10	96.20	43.90	47.30	30.00	24.40
20	10	99.10	100.00	14.60	16.50	0.50	0.50
	30	83.70	91.80	3.00	2.40	0.00	0.10
	50	30.80	36.40	0.30	0.40	0.00	0.00
	100	0.00	0.00	0.30	0.00	0.40	0.00
	130	12.80	11.30	3.10	0.20	1.70	0.16
	150	49.30	48.10	14.20	11.30	7.50	0.50
	200	97.60	96.70	58.80	53.30	39.50	29.60

For simplicity, in all simulated samples the first subject is made an outlying subject by multiplying the responses y_{1Tl} and y_{1Rl} by a constant p. Constant p is permitted to vary from 10% to 200%, so as to consider a full range of various outlier intensities/scenarios.

Table 2 shows the percentage of times (out of 200 simulated samples) that the outlying subject was identified correctly. The results show that the proposed procedure for a replicated crossover design is, on average, as powerful as the original method for the standard crossover design.

All the tests, and the simulation study were developed in SAS[1] Version 8.2 using intensive macro programming.

7. Conclusion and Further Research

All the tests considered in this paper are sequential, in the sense that the procedures are first applied to all subjects and then the subjects are ordered according to some statistic that measures how "unusual" ("outlying") each individual subject is under the assumed model. Tests are applied sequentially, starting from the most "unusual" (largest order statistic) subject until no additional subject can be identified as outlying. Procedures usually stop after one or two steps (i.e., identifying none or one subject as "outlying"). In other words, the tests considered here do not test whether any one specific subject is "outlying", but rather, they identify which (if any) of the subjects may be identified as "outlying" under the assumed model.

The proposed procedure for detecting outlying subjects using a fully repeated crossover design has been found to be as powerful as the equivalent procedure for the standard design. Hence, further research will be directed toward development of tables (and the associated software application/applet) of ordered T^2 statistics upper quantiles (for varying sample sizes, number of outlying subjects, and formulations) to be used specifically in the situation of tests for outlying subjects in higher-order crossover designs.

8. References

[1] Berger R.L., Hsu J.C. Bioequivalence trials, Intersection-Union Tests and Equivalence Confidence Sets (with discussion). Statistical Science 1996; Vol. 11(4), 283-319.

[2] Chen, M.L. Individual bioequivalence – a regulatory update. J. Biopharm. Stat. 1997; Vol.7: 5-11.

[3] Chow, S.-C., Liu J.-P. Design and Analysis of Bioavailability and Bioequivalence Studies. New York: Marcel Dekker, Inc.; Second Edition; 2000.

[4] Chow, S.-C., Tse S.-K. Outlier detection in bioavailability/bioequivalence studies. Statistics in Medicine, 1990; Vol 9: 549-558.
[5] Efron, B. The Jackknife, Bootstrap and Other Resampling Plans. SIAM, Philadelphia; 1982.
[6] Fieller, E. Some problems in interval estimation. J. R. Stat. Soc. B 1954; 16: 175-185.
[7] Hauschke D., Steinijans V.W., Diletti E. A distribution-free procedure for the statistical analysis of bioequivalence studies. Int. J. of Clin. Pharm., Therapy and Toxic., 1990; Vol. 28(2): 72-78.
[8] Liu, J.-P., Weng C.-S. Detection of outlying data in bioavailability/bioequivalence studies. Statistic in Medicine, 1991; Vol. 10: 1375-1389.
[9] Lund, R.E. Tables for an approximate test for outliers in linear models. Technometrics, 17, 473-476.
[10] Lužar-Stiffler V., Stiffler C. Equivalence Testing the Easy Way. CIT, 2002. Vol 10(3); 233-239.
[11] Schuirmann, D.J. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. Biometrics 1981; Vol. 37: 617.
[12] Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J. Pharmacokin. Biopharm. 1987; Vol. 15: 657-680.
[13] US Food and Drug Administration (FDA). Guidance on Statistical Procedures for Bioequivalence Using a Standard Two-treatment Crossover Design. Division of Bioequivalence, Office of Generic Drugs, Center for Drug Evaluation and Research, US Food and Drug Administration, Rockville, MD.; 1992.
[14] US Food and Drug Administration (FDA). Guidance for Industry. Statistical Approaches to Establishing Bioequivalence, Center for Drug Evaluation and Research, US Food and Drug Administration, Rockville, MD.; 2000.
[15] Wang W., Chow S.-S. Examining outlying subjects and outlying records in bioequivalence trials. J. Biopharm. Stat. 2003; Vol.13(1): 43-56.

[1] SAS is a registered trademark of SAS Institute Inc. in the USA and other countries