Equivalence Testing the Easy Way

Vesna Lužar-Stiffler[1] and Charles Stiffler[2]

[1]University Computing Centre and CAIR Research Center, Zagreb, Croatia
[2]CAIR Research Center, Zagreb, Croatia

The purpose of the article is to demonstrate an equivalence testing software application written under SAS[1] Institute software designed for use by pharmaceutical and other medical research professionals. Besides making the entire equivalence testing procedure easier and more efficient, the application "EquivEasy" offers three main advantages over similar software: a) testing for 3×3 in addition to 2×2 crossover designs, b) familiar SAS user environment, and c) export flexibility (MS Word, PDF, HTML). Two case studies are presented with report results provided in tabular and graphical form.

Keywords: equivalence testing, SAS application, crossover design, TOST, nonparametric tests, clinical trials.

1. Introduction

The purpose of this paper is to describe a new SAS software-based statistical application "EquivEasy", that has been designed and developed with the aim of making the actual execution of equivalence testing procedures easier for a typical final user (e.g., a researcher in a pharmaceutical company, etc.), while at the same time ensuring both a high level of statistical competency and access to other powerful features (e.g., import formats, other statistical procedures, etc.) of the SAS software.

Equivalence testing now represents one of the most frequently used routine applications in clinical pharmacokinetic studies.

Although any clinical statistician can easily perform an equivalence testing procedure using almost any reliable statistical software, less experienced researchers can benefit from having a quick, easy, programming-free, "shrink wrapped" application for performing both routine and some of the more demanding equivalence tests.

In some simple situations one can use SAS Analyst Application's "equivalence testing" option, but for more complex situations (e.g., 3×3 crossover design, nonparametric tests), one has to resort to programming.

Among other software solutions for equivalence testing the best known is the EquivTest from Statistical Solutions Ltd., Ireland [10].

The application described in this paper offers three main advantages as compared to existing solutions for equivalence testing. They are as follows:

1) the possibility to perform equivalence testing for 3×3 (3 treatments, 3 periods) crossover designs (in addition to 2×2 crossover and parallel designs),
2) the added flexibility and power of the SAS environment (the "de facto" standard for statistical analysis in pharmaceutical industry), and
3) exporting flexibility (results can be created in simple listing and various graphics formats, MS word, PDF, and HTML formats).

2. Equivalence Testing

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".

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

Therefore, in equivalence studies the null hypothesis tested is "Treatment 1 is NOT equivalent to treatment 2" versus the alternative hypothesis "Treatment 1 is equivalent to treatment 2". (Note: Treatment 2 (T_2) is usually denoted as "R", for "reference" treatment.)

	Test for Difference	**Test for Equivalence**
Null Hypothesis	T_1 equal to T_2	T_1 not equivalent to T_2
Alternative Hypothesis	T_1 different from T_2	T_1 equivalent to T_2

Table 1. Equivalence vs. superiority trials.

There are, generally speaking, two types of equivalence studies: clinical equivalence and bioequivalence studies. Each type has its own purpose:

In Clinical Equivalence (CE) studies treatments are proclaimed "similar" with respect to clinical outcome (e.g., response rates, BP, survival). CE can take into account different therapeutic measures with different mechanisms of action.

In Bioequivalence (BE) studies, drugs are "similar" with respect to pharmacokinetic characteristics (e.g., AUC, CMAX, TMAX). BE is limited to drugs with the same mechanism of action.

Although there are several different approaches to testing bioequivalence, the most common one (also recommended by FDA[2] [11]) is based on equivalence region, and ultimately on confidence sets. The equivalence region E can be defined as follows:

Let δ denote the difference between treatments T_1 and T_2.

Let E be a set of "small" differences.

If δ lies in E (i.e., if δ is "small"), then we say that T_1 and T_2 are equivalent.

If δ does not lie in E (i.e., if δ is "large"), then we say that T_1 and T_2 differ in a clinically relevant way. (That is, it cannot be claimed that they are "substitutes".)

[2] Food and Drug Administration, USA

δ	**Equivalence region (2 sided)**	**Equivalence region (1 sided)**
Mean ratio	$\lvert\mu_1/\mu_2-1\rvert<\varepsilon$	$\mu_1/\mu_2<1+\varepsilon$ or $\mu_1/\mu_2>1-\varepsilon$
Difference in means	$\lvert\mu_1-\mu_2\rvert<\varepsilon$	$\mu_1-\mu_2<\varepsilon$ or $\mu_1-\mu_2>\varepsilon$

Table 2. Typical equivalence regions in bioequivalence studies.

δ	**Equivalence region (2 sided)**	**Equivalence region (1 sided)**
Hazard ratio	$\lvert\lambda_1/\lambda_2-1\rvert<\varepsilon$	$\lambda_1/\lambda_2<\varepsilon$ or $\lambda_1/\lambda_2>-\varepsilon$
Difference in means	$\lvert\pi_1-\pi_2\rvert<\varepsilon$	$\pi_1-\pi_2<\varepsilon$ or $\pi_1-\pi_2>-\varepsilon$

Table 3. Typical equivalence regions in clinical equivalence studies.

Examples of δ and corresponding equivalence regions in **bioequivalence** and **clinical equivalence** studies are given in Table 2 and 3, respectively.

Confidence interval $\mathrm{CI}_\alpha(\delta)$ provides a way to test if "δ lies in E". A $100(1-\alpha)\%$ confidence interval $\mathrm{CI}_\alpha(\delta)$ for a parameter δ is defined by

$$\mathrm{Prob}\{\mathrm{CI}_\alpha(\delta) \text{ contains } \delta\} >= 1-\alpha.$$

The width of the confidence $\mathrm{CI}_\alpha(\delta)$ (for a given δ) depends on the estimate of the standard error of δ estimate (i.e., standard error of difference in means or of mean ratio) and the type of hypothesis being tested (1 or 2 sided). The estimate of the standard error, on the other hand, depends on the experimental design used, statistical model applied, estimation method, software, etc.

There is a certain amount of controversy and misunderstanding around the issue of using $100(1-2\alpha)\%$ vs. $100(1-\alpha)\%$ confidence interval for testing 2-sided equivalence hypothesis at the level α (see e.g., [1]). These and other issues most relevant to this research topic will be covered in the following paragraphs.

In this paper we will be focused on the issues related to bioequivalence studies only (although most of it also applies to clinical equivalence studies).

3. Bioequivalence Testing

Hypotheses that specify only that the population means should be "close" are called average bioequivalence hypotheses. Hypotheses that state that the whole distribution of bioavailabilities is the same for the test and reference populations are called population bioequivalence hypotheses. Sometimes bioequivalence is defined in terms of parameters that more directly measure equivalence of response within an individual. This is called individual bioequivalence. Since the FDA currently requires that only the results of the average bioequivalence tests be submitted, in this paper we will not consider either individual or population bioequivalence tests.

As mentioned, methods for average bioequivalence are based on either the raw data model (untransformed) or the log-transformed model, and are derived under the assumptions of normality or lognormality for between subject (intersubject) and within subject (intrasubject) variabilities.

To claim bioequivalence in average bioavailability it is commonly required that the ratio of the two true formulation averages μ_T/μ_R be within (80%, 120%) limits (or the difference $\mu_T - \mu_R$ be within $\pm 20\%$ of μ_R, where T = test formulation, R = reference formulation). However, for the logarithmic transformation of pharmacokinetic responses, the FDA guidance requests that to claim average bioequivalence, the ratio of the two formulation averages on the original scale be within (80%, 125%) limit (which corresponds to a symmetric interval around 0 in the log scale, i.e., $\log(.80) = -\log(1.25)$). The FDA requires that the bioequivalence be concluded with 90% assurance. In pursuit of this goal, several methods have been proposed in the past two decades. These methods include:

1) the confidence interval approach,

2) the method of interval hypothesis testing,

3) the bayesian approach, and

4) nonparametric methods.

In this paper, we will discuss the 3 most common (currently recommended by FDA, and implemented in EquivEasy) methods: 1) the "classical" (shortest) confidence interval for μ_T/μ_R, 2) Schuirmann's two one sided tests (TOST, [7], [8]), and 3) a nonparametric method.

It can be shown that the 90% confidence interval method is equivalent to carrying out two one sided tests at 5% significance level. Both of the methods will be discussed and implemented for the raw data model (additive) and then for the log-transformed data model (multiplicative). However, some caution is advised with regard to selection of the appropriate test when the analysis is performed on raw data (see below).

In the next two paragraphs typical methods for the analysis of data coming from a crossover design will be briefly introduced. The analysis for data from a parallel design is easily found in standard statistical texts, and will not be dealt with here.

4. Raw Data Analysis

4.1. Confidence Interval Approach

If the analysis is performed in the original scale, then the "classic" confidence interval for the ratio μ_T/μ_R is obtained from the difference condition:

$$\delta_L < \mu_T - \mu_R < \delta_U \qquad (1)$$

using the standard t statistic for $\mu_T - \mu_R$ and converting it to the $(1 - 2\alpha) \times 100\%$ confidence interval for the ratio μ_T/μ_R by dividing the limits for the difference by the least squares estimate of the reference mean (assuming that the estimate is the true reference mean μ_R). Although intuitively appealing and regularly used, the "classical" procedure may not have the desired level of assurance required by the FDA in cases where there is a large coefficient of variation (CV) or a large intrasubject variability (e.g., CV greater than 20%). In other words, the probability of correctly concluding bioequivalence may not be of the desired level. In this case (i.e., when CV is greater than 20%) it is suggested that a simulation study (e.g., parametric bootstrap, see [4]) be conducted to evaluate the finite sample performance of the confidence limits before a decision on average bioequivalence is made. Yet another alternative in the case of high variability is to use some other confidence interval method, such as the one based on Fieller's theorem ([5]).

4.2. TOST

Schuirmann's ([7],[8]) proposed two one-sided (TOST) procedure suggests taking the conclusion of average bioequivalence at the α level of significance if, and only if, both H_{01} and H_{02} are rejected at a predetermined α level of significance:

$$H_{01}: \mu_T - \mu_R \leq \delta_L \qquad H_{02}: \mu_T - \mu_R \geq \delta_U$$
$$H_{a1}: \mu_T - \mu_R > \delta_L \qquad H_{a2}: \mu_T - \mu_R < \delta_U$$

The tests are performed using two one-sided t tests as follows:

$$T_L = (D - \delta_L)/SE(D) > t_{\alpha,r} \quad (2)$$
$$T_U = (D - \delta_U)/SE(D) < -t_{\alpha,r}, \quad (3)$$

where $t_{\alpha,r}$ is the upper 100α percentile of t-distribution with r degrees of freedom, D is the estimated difference between test and reference means, and SE(D) is the standard error of the difference.

5. Log-transformed Data Analysis

There are several reasons or rationales for applying log transformation to AUC and CMAX data. They are labeled "Clinical", "Pharmacokinetic", and "Statistical". The statistical rationale is that much of pharmacokinetic data is skewed in original scale and that it appears more lognormal than normal. Besides, by log transformation, the ratio condition in original scale changes into a difference condition in the log scale:

$$\delta_L < \mu_T/\mu_R < \delta_U \implies$$
$$\log(\delta_L) < \eta_T - \eta_R < \log(\delta_U), \quad (4)$$

where $\mu_i = \exp(\eta_i + \sigma^2/2)$, i=T,R and η_i, σ are parameters of the lognormal distribution. Here we assume that the log transformed data are distributed according to the normal distribution with means μ_T, μ_R and a common variance σ^2 (i.e., that the original data are distributed according to lognormal distribution with parameters η_i and δ). As mentioned earlier, recommended limits for logged data are $\delta_L = .80$ and $\delta_U = 1.25$, which yield symmetric limits $\log(.80) = -.223$ and $\log(1.25) = .223$, respectfully, in log scale.

5.1. Confidence Interval Approach

The confidence interval approach in the case of logged data is straightforward:

$$CI = [D - t_{\alpha,r}SE(D), D + t_{\alpha,r}SE(D)], \quad (6)$$

where $t_{\alpha,r}$ is the upper 100α percentile of t-distribution with r degrees of freedom, D is the estimated difference between test and reference means, and SE(D) is the standard error of the difference. The interval for the ratio of means in the original scale is obtained by exponentiating the interval limits (6).

5.2. TOST

In the case of log transformed data, TOST tests the following two hypotheses:

$$H_{01}: \eta_T - \eta_R \leq \log(\delta_L)$$
$$H_{02}: \eta_T - \eta_R \geq \log(\delta_U)$$
$$H_{a1}: \eta_T - \eta_R > \log(\delta_L)$$
$$H_{a2}: \eta_T - \eta_R < \log(\delta_U).$$

The tests are performed by comparing T_L and T_U to the percentiles of the t-distribution as follows:

$$T_L = (D - \log(\delta_L))/SE(D) > t_{\alpha,r} \quad (7)$$
$$T_U = (D - \log(\delta_U))/SE(D) < -t_{\alpha,r}. \quad (8)$$

It can be shown that the $100(1-2\alpha)$ (and not the $100(1-\alpha)$ confidence interval approach is operationally identical to TOST performed at α level.

The width of the confidence interval (and the decision from TOST) depend on (given fixed δ_L, δ_U and the selected bioequivalence method) the estimate of the difference between two drug means and its standard error. This estimate, on the other hand, depends on a number of other things such as:

- experimental design (e.g., crossover vs. parallel)
- treatment of effects (fixed vs. random)
- effects included in model (e.g., with or without carryover effects)
- unbalanced data (missing data and/or incomplete designs)

- modeling covariance structure (e.g., different variances)
- estimation method (software, procedure, etc.)

5.3. Nonparametric Test

Statistical methods for assessment of bioequivalence are developed under the following assumptions:

S_{ik} (the effect of the ith subject in the kth sequence) are iid normal with mean 0 and variance σ_s^2 ("intersubject" variability),

e_{ijk} (the (within subject) random error) are iid normal with mean 0 and variances σ_e^2 ("intrasubject" variability), and

S_{ik} and e_{ijk} are mutually independent.

It is important to check these assumptions, and several approaches based on examining inter- and intrasubject residuals have been suggested in the literature. However, they are not rigorous statistical tests for normality and should be used with caution due to the small number of subjects usually used in bioequivalence studies (see [9]).

If normality (or lognormality) is seriously violated, TOST (even for log transformed data) is no longer justified. Application EquivEasy uses a distribution-free approach proposed by Hauschke D., Steinijans VW, and Diletti E. [6]. The advantage of this approach is its applicability even in the case of unequal period effects.

The procedure proposed in [6] uses Mann- Whitney-Wilcoxon tests and the corresponding distribution-free $100(1-2\alpha)\%$ confidence interval, and the Hodges-Lehman estimator (as a point estimator for the logged ratio of means). The approach is specially appealing because it yields results in the form of $100(1-2\alpha)\%$ confidence intervals (as with parametric approach), which are then easily compared to the pre-specified (bio)equivalence range (see Table 5 in paragraph 6 for an example of the results generated by the nonparametric option in EquivEasy).

6. EquivEasy Application

Although SAS GLM and MIXED procedures can be used for "standard" bioequivalence testing (e.g., 90% confidence intervals and TOST), their use is not straightforward because the tests results have to be calculated (using appropriate formulas) from SAS procedures outputs. Also, it is usually necessary to first examine the results from the model with carryover effects (in case of crossover design), and then (in the absence of significant carryover effects) from the model without carryover effects. Furthermore, reporting on bioequivalence studies require that some standard tables, figures and listings (TFL) (e.g., means, CV, ratios, estimates of inter- and intrasubject variability, etc.) be supplied in addition to the bioequivalence test(s) results. These usually require that appropriate manipulations and transformations be applied to the data before TFLs are made.

The purpose of EquivEasy application is:

- to raise the likelihood of proper reporting on bioequivalence studies (for data from 2 treatments, 2 periods crossover design and 3x3 crossover (Williams) design):
- to minimize the errors in report preparation (increased quality),
- to minimize the maximum time required for studies (increased efficiency),
- to reduce the need for in-house SAS expertise (i.e., so as to simplify use, and to reduce training costs),
- to maximize the uniformity of reporting (standardization), and
- to minimize additional validation costs by using pre-validated SAS Institute software procedures wherever possible.

Using the application is straightforward: appropriate selections (such as location of the data, response variable name, level names, log/original scale, GLM/ MIXED /nonparametric procedure, with/without carryover effects, limits, output format: MS Word/ HTML/ PDF, etc.) have to be supplied (as shown in Figure 1) prior to pushing the "OK" button, which then creates typical output (tables and graphs) in the selected format.

Data for the first example (3×3 crossover "Williams" design) is taken from [3] (Table 10.3.13). The results are generated by selecting GLM (parametric) procedure, no carryover

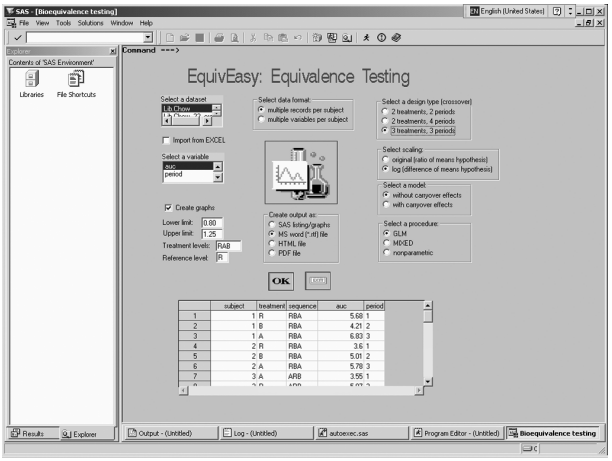

Fig. 1. EquivEasy Application frame.

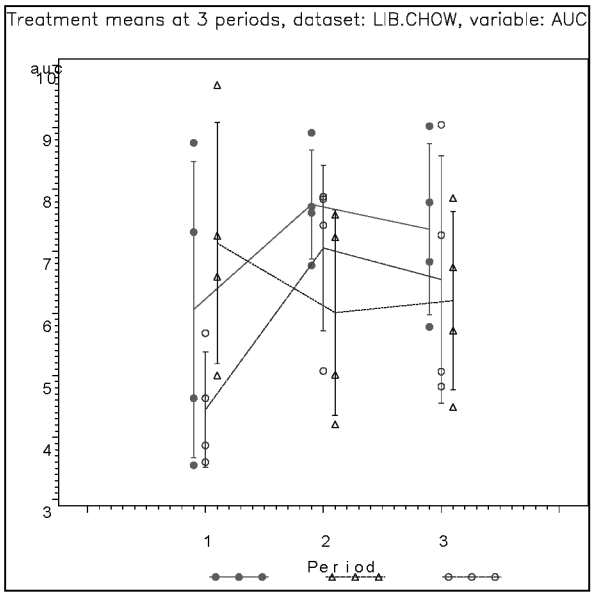

Fig. 2. "Treatment means by period" plot.

effects, original scale, and (0.80,1.20) bioequivalence limits in the EquivEasy Application frame. The output was requested in .rtf (MS Word) format. The key results, a graph of treatment means by periods, and a table with 90% confidence intervals and TOST are shown in Figure 2 and Table 4, respectively. It can be shown that the application yields results identical to the results given in the original Table 10.3.16. in [3] (n.b., original table contains two clerical errors). Incidentally, the results (both the 90% confidence interval approach and TOST) show that equivalence with the reference treatment (R) can be concluded only for treatment B.

The second example (2×2 crossover design) is also based on data from [3] (Table 3.6.1.). The nonparametric procedure, log scale, and the usual (.80, 1.25) bioequivalence range for the logged data were selected to analyze the second data set. The results for the 90% (HSD) confidence interval based on Mann-Whitney-Wilcoxon test, a Hodges-Lehman (HSD) point estimate, and an exact confidence coefficient are given in Table 5. The stated bioequivalence limits (0.90, 1.25) encompass the 90% HSD confidence interval (0.94, 1.18). Hence, it is concluded that the two treatments are equivalent.

Lower 90% HSD Confidence Limit	Upper 90% HSD Confidence Limit	HSD Point Estimate	Exact Confidence Coefficient
0.9430	1.1789	1.0523	0.91127

Table 5. Nonparametric test results.

7. Summary

Equivalence testing need not consume excessive researcher time and company resources. Although not simple, the testing procedure can be made much easier with the aid of "expert

Treatments	Estimate	Standard Error	Estimate in orig.scale	Lower 90% Confidence Limit	Upper 90% Confidence Limit	t for lower TOST	P>t for lower TOST (0.8)	t for upper TOST	P<-t for upper TOST (1.25)
A - R	1.04250	0.43896	1.04250	1.04746	1.29922	5.11511	.00002	-1.05030	0.15305
B - R	0.43333	0.43896	0.43333	0.94617	1.19794	3.72736	.00066	-2.43805	0.01211
A - B	0.60916	0.43896	0.60917	0.97541	1.22717	4.12793	.00026	-2.03748	0.02753

Table 4. 90% parametric confidence intervals and TOST.

system" software application assistance. The EquivEasy application (created using SAS Institute software) makes equivalence testing both faster and easier for the pharmaceutical research professional.

References

[1] BERGER RL, HSU JC. Bioequivalence trials, Intersection-Union Tests and Equivalence Confidence Sets (with discussion). *Statistical Science* 1996; 11(4), 283–319.

[2] CHEN, ML. Individual bioequivalence – a regulatory update. *J. Biopharm. Stat.* 1997; 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] EFRON, B. *The Jackknife, Bootstrap and Other Resampling Plans.* SIAM, Philadelphia; 1982.

[5] FIELLER, E. Some problems in interval estimation. *J. R. Stat. Soc.* B 1954; 16: 175–185.

[6] HAUSCHKE D., STEINIJANS VW., DILETTI E. A distribution-free procedure for the statistical analysis of bioequivalence studies. *Int. J. Of Clin. Pharm., Therapy and Toxic.*, 1990; 28(2): 72–78.

[7] SCHUIRMANN, DJ. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics* 1981; 37: 617.

[8] SCHUIRMANN, DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokin. Biopharm.* 1987; 15: 657–680.

[9] SHAPIRO, SS., WILK MB. An analysis of variance test for normality (complete samples). *Biometrika* 1965; 52: 591–611.

[10] Statistical Solutions Ltd., Ireland. EquivTest. `http://www.statsol.ie/equivtest/equivtest.html` [3/30/2002].

[11] 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.

Received: June, 2002
Accepted: September, 2002

Contact address:
Vesna Lužar-Stiffler
University Computing Centre
and CAIR Research Center
Zagreb, Croatia
e-mail: `vluzar@srce.hr`

Charles Stiffler
CAIR Research Center
Zagreb, Croatia
e-mail: `c|harles.stiffler@cair-center.hr`

VESNA LUŽAR-STIFFLER is a senior researcher at the University Computing Centre, University of Zagreb and the Director of Statistical Methods at CAIR Research Center in Zagreb, Croatia. She obtained her B.Sc. in mathematics, her Ph.D. in computer science/computational statistics at the University of Zagreb, and was awarded a Fulbright Postdoctoral Grant for research at the Department of Statistics, Stanford University, specializing in computational statistics and multivariate analysis. Her teaching experience includes statistical courses at the University of Zagreb, Stanford University, University of Maryland, University of Naples, SAS Institute and various companies in Central and Eastern Europe. She has consulted in the area of statistical/graphical applications, data mining, marketing research, and SAS business intelligence software support with various companies (including pharmaceutical, aeronautic, automotive, semiconductor manufacturing technology, insurance, FMCG retail, beverage, telecommunications) and is currently involved with a variety of companies and government organizations in the US, Italy, Croatia, Slovenia, Macedonia, Romania, etc.

CHARLES STIFFLER obtained his MBA in 1980 at San Diego State University, specializing in consumer psychology and market research, and his doctorate (Ph.D.) in business and administration from the University of Colorado, Boulder in 1985, specializing in market research and business strategy. He taught at the University of Colorado, Boulder until 1987, where over a 7 year period he completed 120+ market research projects. He has consulted in industry, taught in the area of entrepreneurship and was owner/director of a research firm (Environmental Marketing Inc.) during the late 80's. From 1990 until 1995 he taught with the University of Maryland at various locations in Europe. For the past several years he has worked as a private consultant in Europe and the US, and has periodically taught at the University of Zagreb, SAS Institute, University Computing Centre, Zagreb and various institutions in Europe. Recent areas of research activity include beverage industry, food retailing, insurance marketing, automotive sales, telecommunications, customer relationship applications, business analytical solutions, consumer behavior, TQM and excellence in business intelligence systems.