ISSN: 1696-2281
SORT 32 (2) July-December (2008)

SORT

Statistics and Operations Research Transactions

Sponsoring institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Supporting institution

Spanish Region of the International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Published on paper
bearing the certificate
of sustainable forest
management

SORT

Volume 32 (2), July-December 2008

Formerly Qüestiió

SORT 32 (2) July-December 2008, 115-140

Assessing influence in survival data with a cure fraction and covariates

Edwin M. M. Ortega[1], Vicente G. Cancho[2] and Victor Hugo Lachos[3]

[1]*ESALQ-USP,* [2]*ICMC-USP and* [3]*IMECC-UNICAMP*

Abstract

Diagnostic methods have been an important tool in regression analysis to detect anomalies, such as departures from error assumptions and the presence of outliers and influential observations with the fitted models. Assuming censored data, we considered a classical analysis and Bayesian analysis assuming no informative priors for the parameters of the model with a cure fraction. A Bayesian approach was considered by using Markov Chain Monte Carlo Methods with Metropolis-Hasting algorithms steps to obtain the posterior summaries of interest. Some influence methods, such as the local influence, total local influence of an individual, local influence on predictions and generalized leverage were derived, analyzed and discussed in survival data with a cure fraction and covariates. The relevance of the approach was illustrated with a real data set, where it is shown that, by removing the most influential observations, the decision about which model best fits the data is changed.

MSC: 62F15, 62H10, 62J20, 62N01, 62N02

Keywords: Cure fraction, Bayesian inference, local influence, generalized leverage, survival data.

1 Introduction

Models for survival analysis typically assume that every subject in a population is susceptible to the event under study and will eventually experience it if follow-up is sufficiently long. However, there are situations where a fraction of individuals are

[1] ESALQ, University of São Paulo. Departamento de Ciências Exatas, USP Av. Pádua Dias 11 - Caixa Postal 9, 13418-900 Piracicaba - São Paulo - Brazil. e-mail: edwin@esalq.usp.br

[2] ICMC, University of São Paulo, São Carlos, São Paulo, Brazil. E-mail: garibay@icmc.usp.br

[3] IMECC, University of Campinas, Campinas, Brazil. E-mail: vlachos@ime.unicamp.br

Received: March 2007
Accepted: January 2008

not expected to experience the event of interest, that is, those individuals are cured or insusceptible. For example, researchers may be interested in analyzing the recurrence of a disease. Many individuals may never experience a recurrence; therefore, a cured fraction of the population exists. Cure rate models have been utilized to estimate the cured fraction.

Cure rate models are survival models which allow for a fraction of cured individuals. These models extend the understanding of time-to-event data by allowing for the formulation of more accurate and informative conclusions. These conclusions are otherwise unobtainable from an analysis which fails to account for a cured or insusceptible fraction of the population. If a cured component is not present, the analysis reduces to standard approaches of survival analysis.

Cure rate models have been used for modeling time-to-event data for various types of cancers, including breast cancer, non-Hodgkins lymphoma, leukemia, prostate cancer and melanoma. Perhaps the most popular type of cure rate models is the mixture model introduced by Berkson and Gage (1958). In this model, the population is divided into two subpopulations so that an individual either is cured with probability p or has a proper survival function $S(t)$, with probability $1-p$. This gives an improper population survivor function $G(t)$ in the form of mixture, that is,

$$G(t) = p + (1-p)S(t), \quad S(\infty) = 0, \quad G(\infty) = p, \tag{1}$$

A common choice of the $S(t)$ in (1) is exponential and the Weibull distribution. With those choices, we have respectively an exponential model with a cured fraction and a Weibull model with a cured fraction. This mixture model has been studied by several authors, including Farrell (1982), Goldman (1984), Greenhouse (1998) and Sy and Taylor (2000). The book by Maller and Zhou (1996) provides a wide range of applications of the long-term survivor mixture model. We considered a classical analysis for model Weibull with a cured fraction and covariates. The inferential part was carried out using the asymptotic distribution of the maximum likelihood estimators, which in situations when the sample is small, may present difficult results to be justified. As an alternative for classical analysis, we explored the use of techniques of the Markov Chain Monte Carlo (MCMC) method to develop a Bayesian inference for the Weibull model with a cure fraction.

The development of influence diagnostics is an important step in the analysis of a data set as it provides us with an indication of bad model fitting or of influential observations. However, there are no applications of influence diagnostics to survival data with a cured fraction and covariates. Cook (1986) proposed a diagnostic approach named local influence to assess the effect of small perturbations in the model and/or data on the parameter estimates. Several authors have applied the local influence methodology in more general regression models than the normal regression model (see, for example, Paula 1993, Galea et al., 2000 and Dias, et al., 2003). Also, some authors have investigated the assessment of local influence in survival analysis models:

for instance, Pettit and Bin Daud (1989) investigate local influence in proportional hazard regression models; Escobar and Meeker (1992) adapt local influence methods to regression analysis with censoring, Ortega et al. (2003) considered the problem of assessing local influence in generalized log-gamma regression models with censored observations, Ortega et al. (2006) derived curvature calculations under various perturbation schemes in exponentiated-Weibull regression models with censored data and Fachini et al. (2007) adapt local influence methods to polyhazard models under the presence of covariates.

In this article, we present diagnostic methods based on local influence and residual analysis in survival data with a cure fraction and covariates, where the covariates are modeled through p via a binomial regression model. In section 2, we present the Weibull model with a cured fraction and covariates and discuss the process of estimation for the parameters in the model. Section 3 deals with a Bayesian analysis using MCMC methodology under informative priors. In Section 4, 5 and 6, we discuss the local influence method, local influence on predictions and generalized leverage. Likelihood displacement is used to evaluate the influence of observations on the maximum likelihood estimators. Section 7 presents the results of an analysis with a real data set and residual analysis.

2 The Weibull model with a cure fraction and covariates

Let a binary random variable Y_i, $i = 1,\ldots,n$ indicate that the ith individual in a population is at risk or not with respect to a certain type of failure, that is, $Y_i = 1$ indicates that the ith individual will eventually experience a failure event (uncured) and $Y_i = 0$ indicates that the individual will never experience such event (cured). For an individual with covariate vector $\mathbf{x}_i$, the proportion of uncured p can be specified to be a logistic link of $\mathbf{x}$ such that the conditional distribution of Y is given by

$$Pr\Big(Y_i = 1 \,|\, \mathbf{x}_i\Big) = \frac{1}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})} = 1 - p_i$$

where $\boldsymbol{\beta} = (\beta_1,\beta_2,\ldots,\beta_p)^T$ is a vector p-dimensional parameter. Note that the cure probability varies from individual to individual so that the probability that individual i is cured is modeled by $p_i = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}$. The logistic link keeps each p_i strictly between 0 and 1.

Letting T_i be the ith time of occurrence of the failure event and considering that T_i´s are independent and identically distributed with the Weibull distribution, the density function is given by

$$f(t;\alpha,\lambda \,|\, Y_i = 1) = \alpha t^{\alpha-1}\exp\{\lambda - t^{\alpha}e^{\lambda}\}, \tag{2}$$

where $\alpha > 0$ is a shape parameter and $\lambda \in R$ is a scale parameter. Thus, the contribution of an individual that failed at t_i to the likelihood function is given by $(1-p_i)\alpha t_i^{\alpha-1}\exp\{\lambda - t_i^{\alpha}e^{\lambda}\}$, and the contribution of an individual that is at risk at time t_i is $p_i + (1-p_i)\exp\{-t_i^{\alpha}e^{\lambda}\}$.

Given a sample $t_1, \ldots, t_n$, where we observed $t_i = \min(T_i, C_i)$ where T_i is the lifetime for the ith individual and C_i is the censoring time for the ith individual. In this case the log-likelihood function corresponding to the parameter vector $\boldsymbol{\theta} = (\alpha, \lambda, \boldsymbol{\beta}^T)^T$ is given by

$$\begin{aligned} l(\boldsymbol{\theta}) \quad \propto \quad & \mathrm{r}\log(\alpha) + \mathrm{r}\lambda + \sum_{i\,\epsilon\,F}\log(1-p_i) + (\alpha-1)\sum_{i\,\epsilon\,F}\log(t_i) - \\ & \exp\{\lambda\}\sum_{i\,\epsilon\,F} t_i^{\alpha} + \sum_{i\,\epsilon\,C}\log\Big[p_i + (1-p_i)\exp\{-t_i^{\alpha}\mathrm{e}^{\lambda}\}\Big], \end{aligned} \tag{3}$$

where r is the number of uncensored observations (failures), F denotes the set of uncensored observations, C denotes the set of censored observations. Maximum likelihood estimates for parameter vector $\boldsymbol{\theta}$ can be obtained by maximizing the likelihood function, while Bayesian estimation is discussed. In this paper, software Ox (MAXBFGS subroutine) (see Doornik, 1996) was used to compute maximum likelihood estimates (MLE). Covariance estimates for maximum likelihood estimators $\widehat{\boldsymbol{\theta}}$ can also be obtained by using the Hessian matrix. Confidence intervals and hypothesis testing can be conducted by using the large sample distribution of MLE, which is a normal distribution with the covariance matrix as the inverse of Fisher information as long as regularity conditions are satisfied. More specifically, the asymptotic covariance matrix is given by $\mathbf{I}^{-1}(\boldsymbol{\theta})$ with $\mathbf{I}(\boldsymbol{\theta}) = -E[\ddot{\mathbf{L}}(\boldsymbol{\theta})]$ such that $\ddot{\mathbf{L}}(\boldsymbol{\theta}) = \left\{\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right\}$.

Since it is not possible to compute the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ due to the censored observations (censoring is random and noninformative), the matrix of second derivatives of the log likelihood, $-\ddot{\mathbf{L}}(\boldsymbol{\theta})$, evaluated at MLE $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, which is consistent, can be used. Then

$$\ddot{\mathbf{L}}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{L}_{\alpha\alpha} & \mathbf{L}_{\alpha\lambda} & \mathbf{L}_{\alpha\beta} \\ . & \mathbf{L}_{\lambda\lambda} & \mathbf{L}_{\lambda\beta} \\ . & . & \mathbf{L}_{\beta\beta} \end{pmatrix}$$

with the submatrices in appendix A.

3 A Bayesian analysis using MCMC

In this section, we consider a Bayesian approach to the MCMC methodology for approximating the posterior distribution for quantities of interest in survival data with a cure fraction and covariates. As seen in the previous section, likelihood based inference in small samples can be somewhat misleading. Thus, Bayesian inference may play an

important role in such cases. Since the derivation of exact posterior densities is not feasible for the Weibull model with a cure fraction and covariates, we make use of the MCMC methodology to obtain approximation for such densities. We consider the joint prior density for $\boldsymbol{\theta} = (\alpha, \lambda, \boldsymbol{\beta}^T)^T$ of the form

$$\pi(\boldsymbol{\theta}) = \prod_{i=1}^{p} \left(\phi(\beta_i \,|\, \mu_{\beta_i}, \sigma^2_{\beta_i})\right) \phi(\lambda \,|\, \mu_\lambda, \sigma_\lambda) \Gamma(\alpha \,|\, a, b), \tag{4}$$

where $\phi(\cdot \,|\, \mu, \sigma^2)$ denotes the probability density function of the Normal distribution with mean μ and variance σ^2 and $\Gamma(\cdot \,|\, a, b)$ denoting the Gamma distribution with shape parameter $a > 0$ and scale $b > 0$. Here all the hyperparameters are specified.

Combining likelihood function $L(\boldsymbol{\theta}) \propto \exp\{l(\boldsymbol{\theta})\}$ and prior to specification (4), the joint posterior distribution for $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta} \,|\, D) \propto T^{\alpha-1}\alpha^{r+a-1} \exp\left\{-b\alpha - \frac{\lambda^2}{2\sigma^2_\lambda} + r\lambda - \frac{1}{2}\sum_{j=1}^{p}\frac{\beta_j^2}{\sigma^2_{\beta_j}} - \right. \tag{5}$$

$$\left. e^\lambda \sum_{i\in F} t_i^\alpha + \sum_{i\in F}\log(1-p_i) + \sum_{i\,\epsilon\, C}\log\left[p_i + (1-p_i)\exp\{-t_i^\alpha e^\lambda\}\right]\right\},$$

where r is the number of uncensored observations, $T = \prod_{i\in F} t_i$, $i = 1, 2, \ldots, n$ and D denotes the observed data.

To implement the MCMC methodology, we consider Gibbs within the Metropolis-Hasting sampler, which requires the derivation of the complete set of conditional posterior distributions. After some algebraic manipulations, it follows that the conditional posterior densities are given by

$$\pi(\alpha \,|\, \boldsymbol{\beta}, \lambda, D) \propto T^{\alpha-1}\alpha^{r+a-1} \exp\left\{-b\alpha - e^\lambda \sum_{i\in F} t_i^\alpha + \sum_{i\,\epsilon\, C}\log\left[p_i + (1-p_i)\exp\{-t_i^\alpha e^\lambda\}\right]\right\}$$

$$\pi(\lambda \,|\, \alpha, \boldsymbol{\beta}, D) \propto \exp\left\{-\frac{(\lambda-\mu_\lambda)^2}{2\sigma^2_\lambda} + r\lambda - e^\lambda \sum_{i\in F} t_i^\alpha + \sum_{i\,\epsilon\, C}\log\left[p_i + (1-p_i)\exp\{-t_i^\alpha e^\lambda\}\right]\right\}$$

$$\pi(\boldsymbol{\beta} \,|\, \alpha, \lambda, D) \propto \left\{-\frac{1}{2}\sum_{j=1}^{p}\frac{(\beta_j-\mu_{\beta_j})^2}{\sigma^2_{\beta_j}} + \sum_{i\in F}\log(1-p_i) + \right.$$

$$\left. \sum_{i\,\epsilon\, C}\log\left[p_i + (1-p_i)\exp\{-t_i^\alpha e^\lambda\}\right]\right\}$$

Since the conditional posteriors do not present standard forms, the use of the Metropolis-Hasting sampler is required.

4 Influence diagnostics

4.1 Local influence

Let $l(\boldsymbol{\theta})$ denote the log-likelihood function from the postulated model, where $\boldsymbol{\theta} = (\alpha, \lambda, \boldsymbol{\beta}^T)^T$, and let $\boldsymbol{\omega}$ be an $n \times 1$ vector of perturbations restricted to some open subset $\boldsymbol{\Omega} \subset \mathbb{R}^n$. The perturbations are made in the log-likelihood function. We will assume, in particular, the case-weights perturbation scheme such that the log-likelihood function takes the form

$$l(\boldsymbol{\theta} \,|\, \boldsymbol{\omega}) = \sum_{i \in F} \omega_i \log\left[(1 - p_i)\alpha t_i^{\alpha-1} \exp\{\lambda - t_i^{\alpha} e^{\lambda}\}\right] + \sum_{i \in C} \omega_i \log\left[p_i + (1 - p_i) \exp\{-t_i^{\alpha} e^{\lambda}\}\right],$$

where $0 \leq \omega_i \leq 1$ and $\boldsymbol{\omega}_0 = (1, 1, \ldots, 1)^T$ is the vector of no perturbation. Note that $l(\boldsymbol{\theta} \,|\, \boldsymbol{\omega}_0) = l(\boldsymbol{\theta})$. To assess the influence of the perturbations in the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, we consider the likelihood displacement

$$LD(\boldsymbol{\omega}) = 2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{\omega})\},$$

where $\hat{\boldsymbol{\theta}}_{\omega}$ denotes the maximum likelihood estimate under the model $l(\boldsymbol{\theta} \,|\, \boldsymbol{\omega})$.

The idea of local influence (Cook, 1986) is concerned with characterizing the behavior of $LD(\boldsymbol{\omega})$ around $\boldsymbol{\omega}_0$. The procedure consists in selecting a unit direction $\mathbf{d}$, $\| \mathbf{d} \| = 1$, and then considering the plot of $LD(\boldsymbol{\omega}_0 + a\mathbf{d})$ against a, where $a \in R$. This plot is called lifted line. Note that, since $LD(\boldsymbol{\omega}_0) = 0$, $LD(\boldsymbol{\omega}_0 + a\mathbf{d})$ has a local minimum at $a = 0$. Each lifted line can be characterized by considering the normal curvature $C_{\mathbf{d}}(\boldsymbol{\theta})$ around $a = 0$. This curvature is interpreted as the inverse radius of the best fitting circle at $a = 0$. The suggestion is to consider direction $\mathbf{d}_{max}$ corresponding to the largest curvature $C_{\mathbf{d}_{max}}(\boldsymbol{\theta})$. The index plot of $\mathbf{d}_{max}$ may reveal those observations which, under small perturbations, exercise notable influence on $LD(\boldsymbol{\omega})$. Cook(1986) showed that the normal curvature at direction $\mathbf{d}$ takes the form $C_{\mathbf{d}}(\boldsymbol{\theta}) = 2\,|\,\mathbf{d}^T \boldsymbol{\Delta}^T (\ddot{\mathbf{L}})^{-1} \boldsymbol{\Delta} \mathbf{d}\,|$ where $-\ddot{\mathbf{L}}$ is the observed Fisher information matrix for the postulated model ($\boldsymbol{\omega} = \boldsymbol{\omega}_0$) and $\boldsymbol{\Delta}$ is the $(p+1) \times n$ matrix with elements $\Delta_{ji} = \partial^2 L(\boldsymbol{\theta} \,|\, \boldsymbol{\omega}) / \partial\theta_j \, \partial\omega_i$, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, $j = 1, \ldots, p + 2$ and $i = 1, \ldots, n$. Then, $C_{\mathbf{d}_{max}}$ is the largest eigenvalue of the matrix $\mathbf{B} = \boldsymbol{\Delta}^T (\ddot{\mathbf{L}})^{-1} \boldsymbol{\Delta}$, and $\mathbf{d}_{max}$ is the corresponding eigenvector. The index plot of $\mathbf{d}_{max}$ for the matrix $\boldsymbol{\Delta}^T (\ddot{\mathbf{L}})^{-1} \boldsymbol{\Delta}$ can show how to perturb the log-likelihood function to obtain larger changes in the estimate of $\boldsymbol{\theta}$.

Another procedure is the total local curvature corresponding to the *ith* element, which follows by taking $\mathbf{d}_i$ or an $n \times 1$ vector of zeros with one at the *ith* position. Thus, the curvature at the direction $\mathbf{d}_i$ assumes the form

$$C_i = 2\,|\,\boldsymbol{\Delta}_i^T \ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1} \boldsymbol{\Delta}_i\,| \tag{6}$$

where $\mathbf{\Delta}_i^T$ denotes the *ith* row of $\mathbf{\Delta}$. This is named total local influence (see, for example, Lesaffre and Verbeke, 1998). It is suggested looking at the index plot of C_i.

We find, after some algebraic manipulation, the following expressions for the weighted log-likelihood function and for the elements of the matrix $\mathbf{\Delta}$:

In this case the perturbed log-likelihood function takes the form

$$\begin{aligned} l(\boldsymbol{\theta}\,|\,\boldsymbol{\omega}) &= \Big[\log(\alpha)+\lambda\Big]\sum_{i\,\epsilon\,F}\omega_i + \sum_{i\,\epsilon\,F}\omega_i\log(1-p_i) + (\alpha-1)\sum_{i\,\epsilon\,F}\omega_i\log(t_i) \\ &\quad - \exp\{\lambda\}\sum_{i\,\epsilon\,F}\omega_i t_i^{\alpha} + \sum_{i\,\epsilon\,C}\omega_i\log\Big[p_i+(1-p_i)\exp\{-t_i^{\alpha}e^{\lambda}\}\Big] \end{aligned} \tag{7}$$

Let us denote $\mathbf{\Delta} = (\mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_{p+2})^T$.

Then the elements of vector $\mathbf{\Delta}$ is given in appendix B.

However, if the interest is only in vector $\boldsymbol{\beta}$, the normal curvature in direction $\mathbf{d}$ is given by $C_{\mathbf{d}}(\boldsymbol{\beta}) = 2\,|\,\mathbf{d}^T\mathbf{\Delta}^T(\ddot{\mathbf{L}}^{-1} - \mathbf{B}_{22})\mathbf{\Delta}\mathbf{d}\,|$ (see Cook, 1986), where

$$\mathbf{B}_{22} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddot{\mathbf{L}}_{22}^{-1} \end{pmatrix}$$

with $\ddot{\mathbf{L}}_{22}$ denoting the submatrix of $\ddot{\mathbf{L}}$ obtained according to partition

$$\ddot{\mathbf{L}}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix}$$

The index plot of the largest eigenvector of $\mathbf{\Delta}^T(\ddot{\mathbf{L}}^{-1} - \mathbf{B}_{22})\mathbf{\Delta}$ can reveal those observations to be most influential on $\hat{\boldsymbol{\beta}}$.

4.2 Local influence on predictions

Let $\mathbf{z}$ be a $p \times 1$ vector of values of the explanatory variables, for which we do not necessarily have an observed response. Then, the prediction at $\mathbf{z}$ is $\hat{\mu}(\mathbf{z}) = \sum_{j=1}^{p} z_j\hat{\beta}_j$. Analogously, the point prediction at $\mathbf{z}$ based on the perturbed model becomes $\hat{\mu}(\mathbf{z},\boldsymbol{\omega}) = \sum_{j=1}^{p} z_j\hat{\beta}_{j\omega}$, where $\hat{\boldsymbol{\beta}}_{\omega} = (\hat{\beta}_{1\omega},\ldots,\hat{\beta}_{p\omega})^T$ denotes the maximum likelihood estimate from the perturbed model. Thomas and Cook (1990) have investigated the effect of small perturbations in predictions at some particular point $\mathbf{z}$ in continuous generalized linear models and by assuming ϕ known or estimated separately from $\hat{\beta}$. ϕ^{-1} is defined as a dispersion parameter. For more details, see McCullagh and Nelder (1989). They defined three objective functions based on different residuals. Because the diagnostic calculations were identical for the proposed functions, they concentrated the application of the methodology on the objective function $f(\mathbf{z},\boldsymbol{\omega}) = \{\hat{\mu}(\mathbf{z}) - \hat{\mu}(\mathbf{z},\boldsymbol{\omega})\}^2$.

Similarly, we will concentrate our study on investigating the normal curvature of the surface formed by vector $\boldsymbol{\omega}$ and function $f(\mathbf{z}, \boldsymbol{\omega})$, around $\boldsymbol{\omega}_0$. The normal curvature at unit direction $\mathbf{d}$ takes, in this case, the form $C_{\mathbf{d}}(\mathbf{z}) = 2 \mid \mathbf{d}^T \ddot{\mathbf{f}} \mathbf{d} \mid$, where $\ddot{\mathbf{f}} = \partial^2 f / \partial \boldsymbol{\omega} \, \partial \boldsymbol{\omega}^T$ is evaluated at $\boldsymbol{\omega}_0$ and $\hat{\boldsymbol{\beta}}$. From Thomas and Cook (1990) one has that

$$\ddot{\mathbf{f}} = \boldsymbol{\Delta}^T (\ddot{\mathbf{L}}_{\beta\beta}^{-1} \mathbf{z}\mathbf{z}^T \ddot{\mathbf{L}}_{\beta\beta}^{-1}) \boldsymbol{\Delta},$$

where $\boldsymbol{\Delta} = \partial^2 l(\boldsymbol{\theta} \mid \boldsymbol{\omega}) / \partial \boldsymbol{\beta} \, \partial \boldsymbol{\omega}^T$. Consequently

$$\mathbf{d}_{max}(\mathbf{z}) \propto -\boldsymbol{\Delta}^T \ddot{\mathbf{L}}_{\beta\beta}^{-1} \mathbf{z}.$$

In the sequence we discuss the calculation of $\mathbf{d}_{max}(\mathbf{z})$ under additive perturbations for the response and for each continuous explanatory variable.

4.2.1 Response perturbation

Consider regression model (3) by assuming now that each t_i is perturbed as $t_i \rightarrow t_i + (S_t)\omega_i = t_i^*$, $i = 1, \ldots, n$, where (S_t) is a scale factor that can be the estimated standard deviation of T and $w_i \epsilon \mathbb{R}$. Below we give the expressions for the log-likelihood function

Here the perturbed log-likelihood function is expressed as

$$\begin{aligned} l(\boldsymbol{\theta} \mid \boldsymbol{\omega}) \quad = \quad & r \log(\alpha) + r\lambda + \sum_{i \, \epsilon \, F} \log(1 - p_i) + (\alpha - 1) \sum_{i \, \epsilon \, F} \log(t_i^*) - \\ & \exp\{\lambda\} \sum_{i \, \epsilon \, F} t_i^{*\alpha} + \sum_{i \, \epsilon \, C} \log \left[p_i + (1 - p_i) \exp\{-t_i^{*\alpha} e^{\lambda}\} \right] \end{aligned} \tag{8}$$

where $t_i^* = t_i + (S_t)\omega_i$.

Matrix $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \ldots, \Delta_{p+2})^T$ is given in appendix C.

Vector $\mathbf{d}_{max}(\mathbf{z})$ is constructed by taking $\mathbf{z} = \mathbf{x}_i$, which corresponds to the $n \times 1$ vector

$$\mathbf{d}_{max}(\mathbf{x}_i) \propto -\boldsymbol{\Delta}^T \ddot{\mathbf{L}}_{\beta\beta}^{-1} \mathbf{x}_i. \tag{9}$$

A large value for the ith component of (15), $\mathbf{d}_{max_i}(\mathbf{x}_i)$, indicates that the ith observation should have substantial local influence on $\hat{y}_i$. Then, the suggestion is to take the index plot of the $n \times 1$ vector $(\mathbf{d}_{max_1}(\mathbf{x}_1), \ldots, \mathbf{d}_{max_n}(\mathbf{x}_n))^T$ in order to identify those observations with high influence on its own fitted value.

4.2.2 Explanatory variable perturbation

Consider now an additive perturbation on a particular continuous explanatory variable, namely X_t, by making $x_{it\omega} = x_{it} + \omega_i S_x$, where S_x is a scaled factor that can be the estimated standard deviation of X_t. This perturbation scheme leads to the following expressions for the log-likelihood function and for the elements of matrix $\boldsymbol{\Delta}$:

The perturbed log-likelihood function is, in this case, expressed as

$$\begin{aligned} l(\boldsymbol{\theta}\,|\,\boldsymbol{\omega}) &= r\log(\alpha) + r\lambda + \sum_{i\,\epsilon\,F}\log(1-p_i^*) + (\alpha-1)\sum_{i\,\epsilon\,F}\log(t_i) - \\ &\quad \exp\{\lambda\}\sum_{i\,\epsilon\,F} t_i^{\alpha} + \sum_{i\,\epsilon\,C}\log\left[p_i^* + (1-p_i^*)\exp\{-t_i^{\alpha}e^{\lambda}\}\right] \end{aligned} \tag{10}$$

where $p_i^* = \frac{\exp\{\mathbf{x}_i^{*T}\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^{*T}\boldsymbol{\beta}\}}$ and $\mathbf{x}_i^{*T}\boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_t(x_{it} + \omega_i S_x) + \cdots + \beta_p x_{ip}$.

Matrix $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \ldots, \Delta_{p+2})^T$ is given in appendix D.

Similarly to the response perturbation case, the suggestion here is to evaluate the largest curvature at $\mathbf{z} = \mathbf{x}_i$, which leads to

$$C_{max}(\mathbf{x}_i) = 2\,|\,\mathbf{d}_{max}^T\,\ddot{\mathbf{f}}\mathbf{d}_{max}\,|\,,$$

and consequently

$$\mathbf{d}_{max}(\mathbf{x}_i) \propto -\boldsymbol{\Delta}^T\ddot{\mathbf{L}}_{\beta\beta}^{-1}\mathbf{x}_i.$$

To see for which observed values of X_t the prediction is most sensitive under small changes in X_t, we can perform the plot of $C_{max}(\mathbf{x}_i)$ against x_{it}. The index plot of the $n \times 1$ vector $(\ell_{max_1}(\mathbf{x}_1), \ldots, \ell_{max_n}(\mathbf{x}_n))^T$ can indicate those observations for which a small perturbation in the value of X_t leads to a substantial change in the prediction.

4.3 Generalized leverage

Let $l(\boldsymbol{\theta})$ denote the log-likelihood function from the postulated model in equation (10), $\widehat{\boldsymbol{\theta}}$ the MLE of $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ the expectation of $\mathbf{T}$, then, $\widehat{\mathbf{t}} = \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}})$ will be the predicted response vector.

The main idea behind the concept of leverage (see, for instance, Cook and Weisberg, 1982; Wei et al., 1998) is that of evaluating the influence of t_i on its own predicted value. This influence may well be represented by derivative $\frac{\partial\widehat{t_i}}{\partial ti}$ that equals h_{ii} is the i-th principal diagonal element of the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{X}$ is the model matrix. Extensions to more general regression models have been given, for instance, by St. Laurent and Cook (1992), and Wei, et al. (1998) and Paula (1999), when $\boldsymbol{\theta}$ is restricted with inequalities. Hence, it follows from Wei et al.(1998) that the nxn matrix $(\frac{\partial\widehat{\mathbf{t}}}{\partial\mathbf{t}})$ of generalized leverage can be expressed as:

$$\mathbf{GL}(\widehat{\boldsymbol{\theta}}) = \left\{\mathbf{D}_{\boldsymbol{\theta}}\left[-\ddot{\mathbf{L}}(\boldsymbol{\theta})\right]^{-1}\ddot{\mathbf{L}}_{\boldsymbol{\theta}\mathbf{t}}\right\} \tag{11}$$

evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$.

Matrix $\mathbf{D}_{\boldsymbol{\theta}} = \left(\mathbf{D}_{\alpha}, \mathbf{D}_{\lambda}, \mathbf{D}_{\boldsymbol{\beta}}\right)$ is given in appendix E.

5 Residual analysis

In order to study departures from the error assumption as the well as the presence of outliers, we will first consider the martingale residual proposed by Barlow and Prentice (1988) (see also Therneau et al., 1990). This residual was introduced in counting processes and can be written for the Weibull model with a cure fraction and covariates as

$$r_{M_i} = \begin{cases} 1 + \log\left[\hat{p}_i + (1-\hat{p}_i)\exp\{-t_i^{\hat{\alpha}}e^{\hat{\lambda}}\}\right], & \text{if} \quad i \in F; \\ \log\left[\hat{p}_i + (1-\hat{p}_i)\exp\{-t_i^{\hat{\alpha}}e^{\hat{\lambda}}\}\right], & \text{if} \quad i \in C. \end{cases} \tag{12}$$

Due to the skewness distributional form of r_{M_i}, it has maximum value +1 and minimum value $-\infty$, and transformations to achieve a more normal shaped form would be more appropriate for residual analysis. Another possibility is to use the deviance residual (see, for instance, McCullagh and Nelder, 1989, Section 2.4), which has been largely applied in generalized linear models (GLMs). Various authors have investigated the use of deviance residuals in GLMs (see, for instance, Williams, 1987; Hinkley et al., 1991; Paula 1995; Ortega et al., 2007) as well as in other regression models (see, for example, Farhrmeir and Tutz, 1994). In the Weibull model with a cure fraction and covariates, the modified residual deviance is expressed here as

$$r_{D_i} = \begin{cases} \operatorname{sgn}(r_{M_i})\Bigg[-2-2\log\Big\{\Big[\hat{p}_i + (1-\hat{p}_i)\exp\{-t_i^{\hat{\alpha}}e^{\hat{\lambda}}\}\Big]\times \\ \log\left[\hat{p}_i + (1-\hat{p}_i)\exp\{-t_i^{\hat{\alpha}}e^{\hat{\lambda}}\}\right]^{-1}\Big\}\Bigg]^{1/2}, & \text{if} \quad i \in F; \\ \operatorname{sgn}(r_{M_i})\Big\{-2\log\left[\hat{p}_i + (1-\hat{p}_i)\exp\{-t_i^{\hat{\alpha}}e^{\hat{\lambda}}\}\right]\Big\}^{1/2}, & \text{if} \quad i \in C, \end{cases} \tag{13}$$

where r_{M_i} is the residual martingale corresponding to the Weibull model with a cure fraction and covariates.

6 Application

In this section, the application of the local influence theory to a set of real data on cancer recurrence is discussed. The data are part of an assay on cutaneous melanoma (a type of malignant cancer) for the evaluation of postoperative treatment performance with a high dose of a certain drug (interferon alfa-2b) in order to prevent recurrence. Patients were included in the study from 1991 to 1995, and follow-up was conducted until 1998. The data were collected by Ibrahim et al. (2001); variable T represented the time until the patient's death. The original size of the sample was n = 427 patients , 10 of whom did not present a value for covariable tumor thickness, herein denominated as Breslow. When such cases were removed, a sample of size n = 417 patients was considered. The

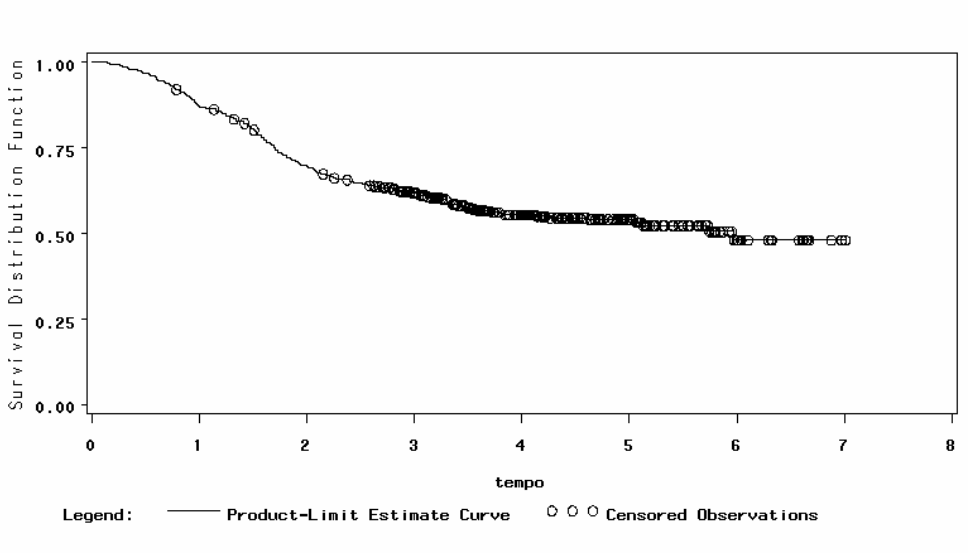


Figure 1: *Plot of the Survivor Function.*

percentage of censored observations was 56%. The following data were associated with each participant, $i = 1, 2, \ldots, n$.

- t_i: observed time (in years);
- δ_i: censoring indicator (0=censoring, 1=lifetime observed);
- x_{i1}: treatment (0=observation, 1=interferon);
- x_{i2}: age (in years);
- x_{i3}: nodule (nodule category: to 4);
- x_{i4}: sex (0=male, 1=female);
- x_{i5}: p.s. (performance status-patient's functional capacity scale as regards his daily activities: 0=fully active, 1=other);
- x_{i6}: Breslow (tumor thickness in mm).

The survival function graph, Kaplan-Meier estimate, is presented in Figure 1, from where a fraction of survivors can be observed.

6.1 Maximum likelihood results

To obtain the maximum likelihood estimates for the parameters in the Weibull model we use the subroutine MAXBFGS in Ox, whose results are given in the Table following.

The mean cure fraction estimated was $\hat{p} = 0.5162$.

In Table 1, it is estimated that the only significant variable is x3(nodule). Also, the information criteria based on the decision theory which penalize models with a large

Table 1: *Maximum likelihood estimates for the complete data set of the Weibull model with a cure fraction and covariates*

Parameter	Estimate	SE	p-value
α	1.6104	0.1066	—
λ	−1.2877	0.1217	—
β_0	2.2656	0.5811	< 0.0001
β_1	−0.1603	0.2247	0.4756
β_2	−0.0142	0.0086	0.0977
β_3	−0.5392	0.1142	< 0.0001
β_4	0.2019	0.2315	0.3832
β_5	−0.1509	0.3352	0.6527
β_6	−0.0599	0.0391	0.1253
Statistics	Value	Statistics	Value
AIC	1045.578	BIC	1081.876

number of parameters were used. The used criteria are based on the AIC statistics (Akaike Information Criterion) and BIC (Bayesian Information Criterion) (see Table 1).

6.2 Bayesian analysis

We consider now a Bayesian analysis for the data considering the following independent prior (4) with values of the hyperparameters given for $a = b = 0,1$, $\mu_\lambda = \mu_{\beta_j} = 0$ and $\sigma^2_\lambda = \sigma^2_{\beta_j} = 100$, $j = 0,1,\ldots,6$. Considering those prior densities we generated two parallel independent runs of the Gibbs sampler chain with size 40,000 for each parameter, discarding the first 5,000 iterations. To eliminate the effect of the initial values and to avoid correlation problems, we considered a spacing of size 10, obtaining a sample of size 3,500 from each chain. To monitor the convergence of the Gibbs

Table 2: *Bayesian estimates. Posterior summary results of fitting the Weibull model with a cure fraction and covariates to the data set.*

Parameters	Mean	SD	95% credible interval	$\hat{R}$
α	1.5760	0.1123	(1.353 ; 1.793)	1.017
λ	−1.3020	0.1227	(−1.544 ; −1.071)	1.000
β_0	2.2870	0.5962	(1.164 ; 3.508)	1.002
β_1	−0.1506	0.2325	(−0.603 ; 0.299)	1.001
β_2	−0.0136	0.0086	(−0.031 ; 0.002)	1.001
β_3	−0.5700	0.1268	(−0.826 ; −0.339)	1.005
β_4	0.2095	0.2377	(−0.259 ; 0.674)	1.072
β_5	−0.1508	0.3446	(−0.839 ; 0.509)	1.001
β_6	−0.0681	0.0432	(−0.159 ; 0.009)	1.011

samples, we used the between and within sequence information, following the approach developed in Gelman and Rubin (1992) to obtain the potential scale reduction, $\hat{R}$. In all cases, these values were close to one, indicating the convergence of the chain. In Table 2 we report posterior summaries for the parameters of the Weibull, mixture model and, in Figure 2, we have the approximate marginal posterior densities considering 7,000 Gibbs samples.

In Table 2, we observe that only the covariate nodule (x_3) presents significant effect on lifetime. It is interesting to note that the Bayesian analysis is very similar to the classical analysis. The computational implementation of the algorithm was developed in the software package R jointly with package R2Winbug (see Gelman, 2004), and the programs can be requested from the authors.

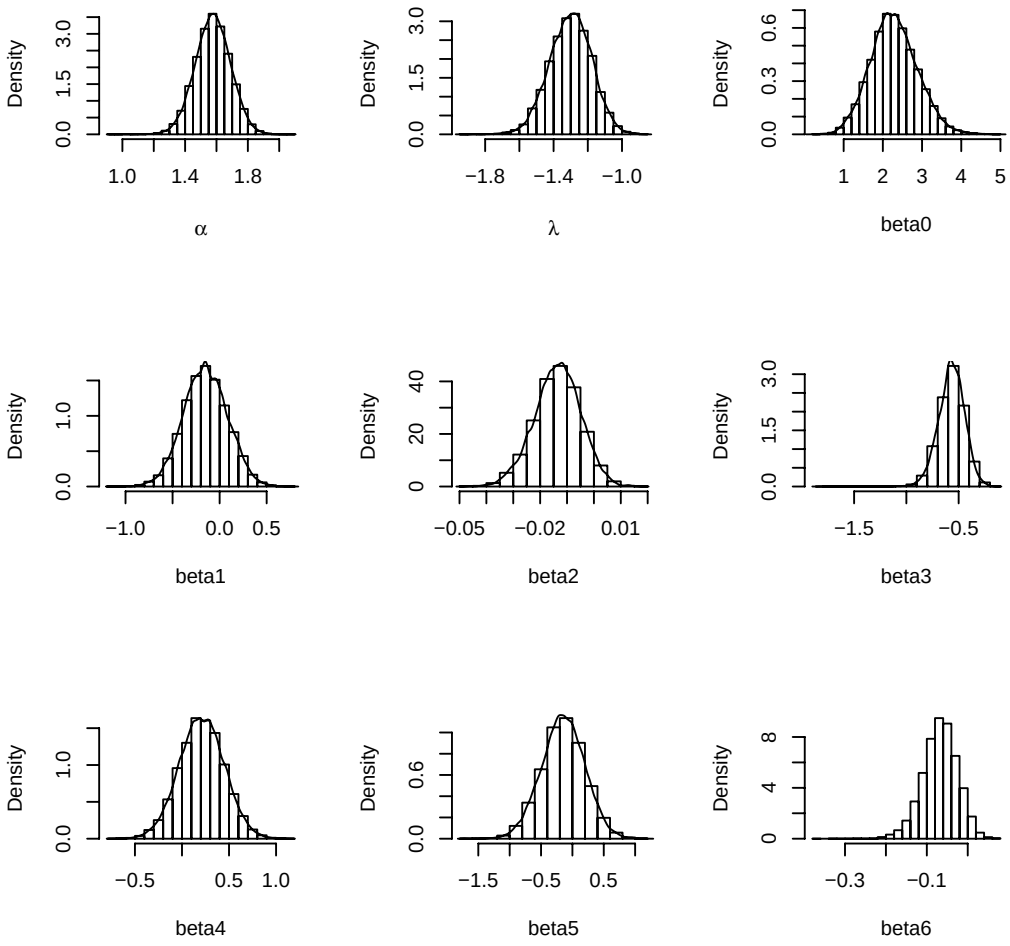


Figure 2: *Approximate marginal posterior densities for parameters of the Weibull model with a cure fraction and covariates.*

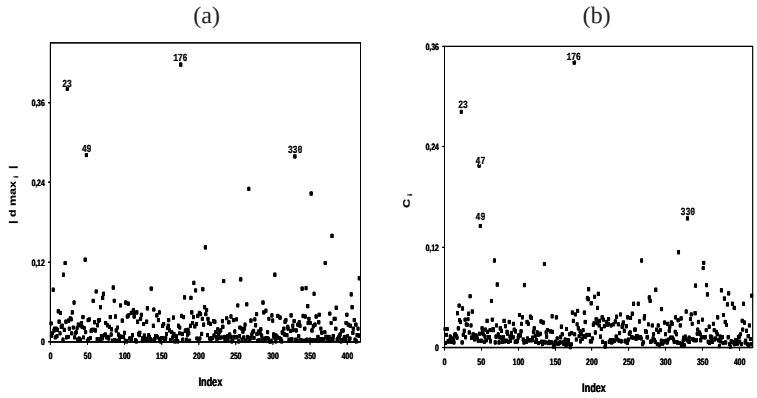


Figure 3: *(a) Index plot of* $\mathbf{d}_{max}$ *for* $\boldsymbol{\theta}$ *(case-weights perturbation). (b) Total local influence on the estimates* $\boldsymbol{\theta}$ *(case-weights perturbation)*

6.3 Local influence analysis

In this section, we will make an analysis of local influence for the data set given in Ibrahim et. al. (2001), using a cure fraction in the Weibull model.

6.3.1 Case-weights perturbation

By applying the local influence theory developed in Section 3, where case-weight perturbation is used, value $C_{\mathbf{d}_{max}} = 1.5820$ was obtained as maximum curvature. In Figure 3(a), the graph of the eigenvector corresponding to $C_{\mathbf{d}_{max}}$ is presented, and total influence C_i is shown in Figure 3(b). Observations ♯23 and ♯176 are the most distinguished in relation to the others.

6.3.2 Prediction of influence using the response variable perturbation

Next, the influence of perturbations on the observed survival times will be analyzed. The value for the maximum curvature calculated was $C_{\mathbf{d}_{max}} = 11.21$. Figure 4 (a), containing the graph for $|\mathbf{d}_{max}|$ versus the observation index, shows that some points were distinguished from the others, among which are points ♯279 and ♯341. The same applies to Figure 4(b), which corresponds to total local influence (C_i). By analyzing the data associated with these two observations, it is noted that the highlighted observations refer to patients with shorter non-censored survival times.

6.3.3 Prediction of influence using the explanatory variable perturbation

The perturbation of vectors for covariates age (x_2) and Breslow (x_6) are investigated here. For perturbation of covariable age, value $C_{\mathbf{d}_{max}} = 1.0374$ was obtained as maximum curvature, and for the perturbation of covariable Breslow, value $C_{\mathbf{d}_{max}} = 1.2864$ was achieved. The respective graphs of $|\mathbf{d}_{max}|$ as well as total local influence C_i against the

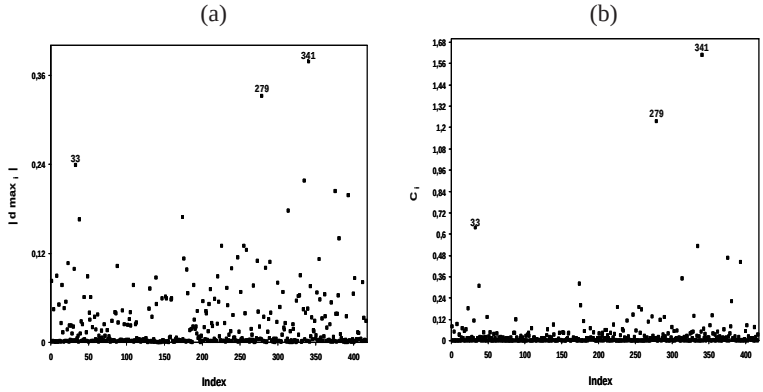


Figure 4: *(a) Index plot of* $\mathbf{d}_{max}$ *for* $\boldsymbol{\theta}$ *(response perturbation). (b) Total local influence on the estimates* $\boldsymbol{\theta}$ *(response perturbation)*

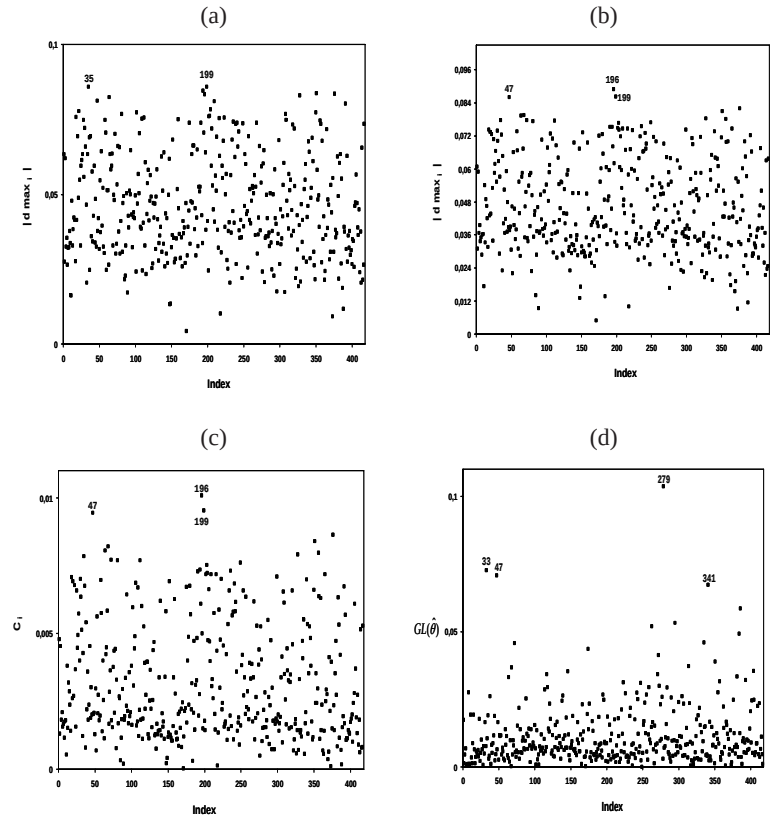


Figure 5: *(a)Index plot of* $\mathbf{d}_{max}$ *for* $\boldsymbol{\theta}$ *(age explanatory variable perturbation). (b)Index plot of* $\mathbf{d}_{max}$ *for* $\boldsymbol{\theta}$ *(Breslow explanatory variable perturbation). (c)Total local influence on the estimates* $\boldsymbol{\theta}$ *(Breslow explanatory variable perturbation). (d) Generalized leverage for* $\boldsymbol{\theta}$

observation index are shown in Figures 5(a), 5(b) and 5(c). In these three graphs, we can see no influential observation.

6.3.4 Generalized leverage analysis

Figure 5(d) exhibits the index plot of $GL(\boldsymbol{\theta})$, using the model given in equation (12). The generalized leverage graph presented in Figure 5(d) confirms the tendencies observed under local and total influence methods. Observations with large and small values for **t** tend to have a high influence on these own-fitted values. We note outstanding influence observations ♯33, ♯279 and ♯341. The graph for $GL(\boldsymbol{\theta})$ is very similar to the one given in Figure 4(a).

6.4 Residual analysis

In order to detect possible outlying observations as well as departures from the assumptions generalized log-gamma regression models with a cure fraction, we present, in Figure 6(a) and 6(b), the graphs of r_{M_i} and r_{D_i} against the order observations.

By analyzing the martingale residual and modified deviance residual graph, a random behavior is observed for the data. A tendency to form two groups is also noted; however, this results from considering the logistic function to introduce covariables. Such problems are also observed in the logistic regression. For further details, refer to Hosmer et al. (2003), Mccullagh et al. (1989), among others.

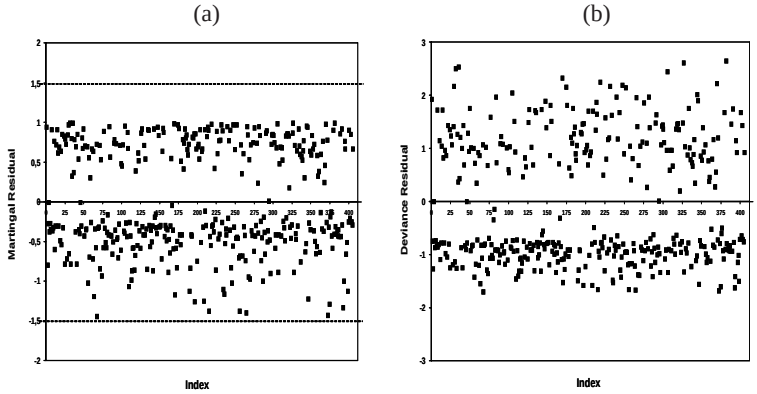


Figure 6: *(a) Index plot of the martingal residual r_{M_i}. (b) Index plot of the modified deviance residual r_{D_i}.*

6.5 Impact of the detected influential observations

Therefore, diagnostic analysis (local influence, local influence on predictions, generalized leverage and residual analysis) detected the following four cases ♯23, ♯176, ♯279 and ♯341 as potentially influential. In order to reveal the impact of these three observations on the parameter estimates, we refitted the model under some situations. First, we individually eliminated each one of these three cases. In Table 4, we have

Table 3: *Relative changes [-**RC**- in %], parameter estimates and their p-values in parentheses for the indicated set.*

Propping	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
	–	–	–	–	–	–	–
all observations	2.27	−0.16	−0.01	−0.54	0.20	−0.15	−0.06
	(< 0.01)	(0.48)	(0.10)	(< 0.01)	(0.38)	(0.65)	(0.13)
	[1]	[−5]	[−3]	[−1]	[3]	[−3]	[−1]
♯23	2.23	−0.15	−0.01	−0.53	0.20	−0.15	−0.06
	(< 0.01)	(0.50)	(0.11)	(< 0.01)	(0.40)	(0.64)	(0.12)
	[1]	[−4]	[−2]	[−2]	[4]	[−7]	[−4]
♯176	2.25	−0.17	−0.01	−0.53	0.19	−0.16	−0.06
	(< 0.01)	(0.45)	(0.10)	(< 0.01)	(0.40)	(0.63)	(0.13)
	[1]	[−2]	[−2]	[−1]	[2]	[−14]	[0]
♯279	2.24	−0.16	−0.01	−0.53	0.20	−0.13	−0.06
	(< 0.01)	(0.47)	(0.11)	(< 0.01)	(0.39)	(0.70)	(0.13)
	[1]	[−11]	[−2]	[−1]	[9]	[−8]	[−2]
♯341	2.29	−0.18	−0.01	−0.54	0.22	−0.16	−0.06
	(< 0.01)	(0.43)	(0.09)	(< 0.01)	(0.34)	(0.63)	(0.12)
	[0]	[−3]	[−1]	[−4]	[10]	[−11]	[−5]
♯23/♯176	2.26	−0.16	−0.01	−0.52	0.18	−0.17	−0.06
	(< 0.01)	(0.48)	(0.09)	(< 0.01)	(0.42)	(0.61)	(0.13)
	[1]	[−5]	[−1]	[−3]	[9]	[−9]	[−2]
♯23/♯279	2.25	−0.15	−0.01	−0.52	0.18	−0.14	−0.06
	(< 0.01)	(0.49)	(0.10)	(< 0.01)	(0.42)	(0.68)	(0.12)
	[1]	[−3]	[−3]	[−2]	[3]	[−12]	[0]
♯23/♯341	2.30	−0.17	−0.01	−0.53	0.21	−0.17	−0.06
	(< 0.01)	(0.46)	(0.08)	(< 0.01)	(0.37)	(0.61)	(0.12)
	[2]	[−6]	[−5]	[−3]	[6]	[−7]	[−4]
♯176/♯279	2.22	−0.17	−0.01	−0.52	0.19	−0.14	−0.06
	(< 0.01)	(0.44)	(0.11)	(< 0.01)	(0.41)	(0.67)	(0.13)
	[0]	[−15]	[0]	[−3]	[5]	[−14]	[−2]
♯176/♯341	2.27	−0.18	−0.01	−0.52	0.21	−0.17	−0.06
	(< 0.01)	(0.41)	(0.09)	(< 0.01)	(0.35)	(0.60)	(0.12)
	[0]	[−12]	[−1]	[−2]	[7]	[−6]	[−1]
♯279/♯341	2.26	−0.18	−0.01	−0.53	0.22	−0.14	−0.06
	(< 0.01)	(0.42)	(0.10)	(< 0.01)	(0.35)	(0.67)	(0.12)

the relative changes (in percentage) of each parameter estimate, defined by: $RC_{\boldsymbol{\theta}_j} = [(\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}_j(I))/\hat{\boldsymbol{\theta}}_j] \times 100$, and the corresponding p-values, where $\hat{\boldsymbol{\theta}}_j(I)$ denotes the MLE of $\boldsymbol{\theta}_j$ after that "set I" of observations has been removed.

Table 4: Continuation Relative changes [-RC- in %], parameter estimates and their p-values in parentheses for the indicated set.

Propping	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
	[1]	[−1]	[−3]	[−5]	[13]	[−3]	[−6]
♯23/♯176/♯279	2.24	−0.16	−0.01	−0.51	0.18	−0.15	−0.06
	(< 0.01)	(0.47)	(0.10)	(< 0.01)	(0.44)	(0.65)	(0.13)
	[0]	[−5]	[0]	[−3]	[0]	[−2]	[−1]
♯23/♯279/♯341	2.27	−0.17	−0.01	−0.52	0.20	−0.15	−0.06
	(< 0.01)	(0.45)	(0.09)	(< 0.01)	(0.38)	(0.66)	(0.12)
	[1]	[−17]	[−3]	[−4]	[3]	[−1]	[−2]
♯176/♯279/♯341	2.24	−0.19	−0.01	−0.52	0.21	−0.15	−0.06
	(< 0.01)	(0.40)	(0.10)	(< 0.01)	(0.36)	(0.65)	(0.12)
	[0]	[−10]	[−2]	[−5]	[4]	[−5]	[−4]
♯23/♯176/♯279/♯341	2.26	−0.18	−0.01	−0.51	0.19	−0.16	−0.06
	(< 0.01)	(0.42)	(0.10)	(< 0.01)	(0.39)	(0.63)	(0.12)

From Tables 3 and 4 we can notice some robust aspects of the maximum likelihood estimates from the Weibull model with a cure fraction and covariates. In general, the significance of the parameter estimates does not change after removing set I at the level of 5 %. A significant change was found when observations 23 and 341 were removed, from which it was noted that covariate age was significant if an 8% level were taken into account. Therefore, we did not encounter inferential change after removing the observations given in the diagnostic graphs.

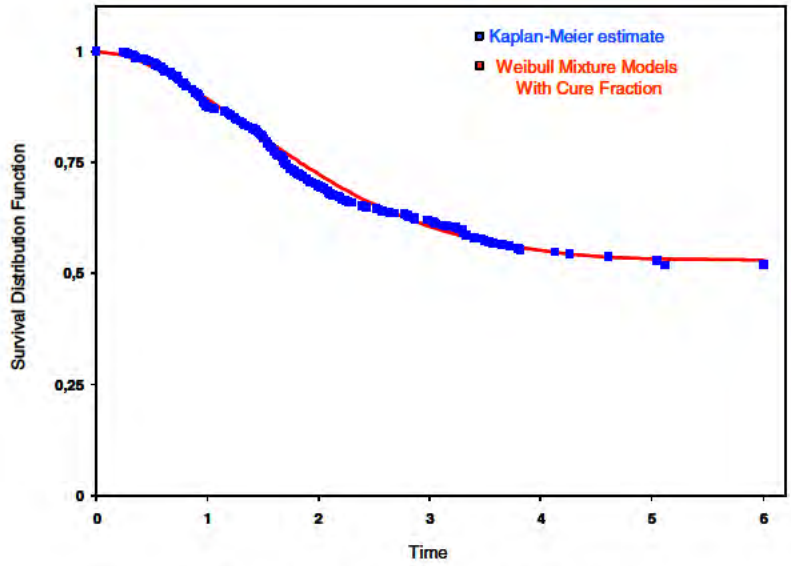


Figure 7: Theoretical survival curve and Kaplan-Meier curve

6.6 Quality of fitting

In order to measure quality of fitting, a Kaplan-Meier survival graph and a survival graph estimated by the Weibull model with a cure fraction were plotted. Good model fitting was observed.

7 Concluding Remarks

The local influence theory (Cook (1986) and Thomas and Cook (1990)), that of generalized leverage proposed by Wei et al. (1998) and a study based on martingale and deviance residual in a survival model with a cure fraction were discussed in this study by using two estimation approaches: the maximum likelihood and the Bayesian approaches. The matrices necessary for application of the technique were obtained by taking into account various types of perturbations to the data elements and to the models. By applying such results to a data set, indication was found of which observations or set of observations would sensitively influence the analysis results. This fact is illustrated in Application (Section 7). By means of a real data set, it was observed that, for some perturbation schemes, the presence of certain observations could considerably change the levels of significance of certain variables. The results of the applications indicate that the local influence technique as well as that of generalized leverage in models with a cure fraction can be rather useful in the detection of possibly influential points by admitting two types of estimation methods: maximum likelihood and Bayesian. In order to measure quality of fitting, martingale and deviance residuals were used, which showed that the model fitting was correct. The Kaplan-Meier survival function was also plotted with the survival function for the proposed model, indicating good model fitting.

Appendix A: Matrix of second derivatives $\ddot{\mathbf{L}}(\boldsymbol{\gamma})$

Here we derive the necessary formulas to obtain the second order partial derivatives of the log-likelihood function. After some algebraic manipulations, we obtain

$$\begin{aligned}
\mathbf{L}_{\alpha\alpha} &= -\frac{\mathrm{r}}{\alpha^2} - \exp\{-\lambda\}\sum_{i\,\epsilon\,F} t_i^{\alpha}\Big[\log(t_i)\Big]^2 \\
&\quad -\sum_{i\,\epsilon\,C}\frac{(1-p_i)\Big[\log(t_i)\Big]^2\Big[-\log(\mathrm{h}_i)\Big]\mathrm{h}_i\Big[p_i\{1+\log(h_i)\}+(1-p_i)h_i\Big]}{\Big[p_i+(1-p_i)h_i\Big]^2}. \\
\mathbf{L}_{\alpha\lambda} &= -\exp\{\lambda\}\sum_{i\,\epsilon\,F} t_i^{\alpha}\log(t_i) \\
&\quad +\sum_{i\,\epsilon\,C}\frac{(1-p_i)\Big[\log(t_i)\Big]\Big[\log(h_i)\Big]h_i\Big[p_i\{1+\log(h_i)\}+(1-p_i)h_i\Big]}{\Big[p_i+(1-p_i)h_i\Big]^2}.
\end{aligned}$$

$$\mathbf{L}_{\alpha\beta} = -\sum_{i\epsilon C}\frac{(x_{ij})p_i\Big[\log(t_i)\Big]\Big[\log(h_i)\Big]h_i}{\Big[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\Big]\Big[p_i+(1-pi)h_i\Big]^2}.$$

$$\mathbf{L}_{\lambda\lambda} = -\exp\{\lambda\}\sum_{i\epsilon F}t_i^{\alpha} + \sum_{i\epsilon C}\frac{(1-p_i)h_i\log(h_i)\Big[1+\log(h_i)+(1-p_i)h_i\{-\log(h_i)\}\Big]}{\Big[p_i+(1-pi)h_i\Big]^2}.$$

$$\mathbf{L}_{\lambda\beta} = -\sum_{i\epsilon C}\frac{(x_{ij})p_i\Big[-\log(h_i)\Big]h_i}{\Big[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\Big]\Big[p_i+(1-pi)h_i\Big]^2}.$$

$$\mathbf{L}_{\beta\beta} = \sum_{i\epsilon F}\frac{-(x_{ij}^2)p_i\Big[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}(p_i-1)\Big]}{(1-pi)^2\Big[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\Big]^2} + \sum_{i\epsilon C}\frac{(x_{ij}^2)p_i\Big[1-h_i\Big]\Big\{\Big[1-\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\Big]\Big[p_i+(1-pi)h_i\Big]-p_i\Big[1-h_i\Big]\Big\}}{\Big[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\Big]^2\Big[p_i+(1-p_i)h_i\Big]^2}.$$

where $h_i=\exp\{-t_i^{\alpha}e^{\lambda}\}$, $p_i=\dfrac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}$, $i=1,2,\ldots,n$ and $j=1,2,\ldots,p$.

Appendix B: Local influence: Case-weight perturbation $\ddot{\mathbf{L}}(\boldsymbol{\gamma})$

Here, we provide the derivatives of the elements considering the case-weight perturbation scheme. Then the elements of vector $\boldsymbol{\Delta}_1$ take the form

$$\Delta_{1i}=\begin{cases}\dfrac{1}{\hat{\alpha}}+\log(t_i)[1+\log(\hat{h}_i)], & \text{if} \quad i\epsilon F;\\ \dfrac{(1-\hat{p}_i)[\log(\hat{h}_i)][\log(t_i)]\hat{h}_i}{[\hat{p}i+(1-\hat{p}_i)\hat{h}_i]}, & \text{if} \quad i\epsilon C.\end{cases}$$

The elements of vector $\boldsymbol{\Delta}_2$ take the form

$$\Delta_{2i}=\begin{cases}1+\log(\hat{h}_i), & \text{if} \quad i\epsilon F;\\ \dfrac{(1-\hat{p}_i)[\log(\hat{h}_i)]\hat{h}_i}{[\hat{p}i+(1-\hat{p}_i)\hat{h}_i]}, & \text{if} \quad i\epsilon C.\end{cases}$$

The elements of vector $\mathbf{\Delta}_j$, for $j = 3, \ldots, p+2$, can be expressed as

$$\Delta_{ji} = \begin{cases} -\dfrac{(x_{ij})\hat{p}_i}{(1-\hat{p}_i)[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}]}, & \text{if} \quad i \epsilon F; \\ \dfrac{(x_{ij})\hat{p}_i[1-\hat{h}_i]}{[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}][\hat{p}_i+(1-\hat{p}_i)\hat{h}_i]}, & \text{if} \quad i \epsilon C. \end{cases}$$

where

$$\hat{h}_i = \exp\{-t_i^{\hat{\alpha}} e^{\hat{\lambda}}\} \quad \hat{p}_i = \frac{\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}.$$

Appendix C: Local influence on predictions: Response perturbation

Here we provide the derivatives of elements $\mathbf{\Delta}_{ij}$ of matrix $\mathbf{\Delta}$ considering the response variables perturbation scheme. The elements of vector $\mathbf{\Delta}_1$ take the form

$$\Delta_{1i} = \begin{cases} \dfrac{(S_t)}{t_i} + (S_t)\log(\hat{h}_i)(t_i)^{-1}[(\hat{\alpha})\log(t_i)+1], & \text{if} \quad i \epsilon F; \\ (1-\hat{p}_i)(S_t)\log(\hat{h}_i)(t_i)^{-1}\hat{h}_i\Big\{\dfrac{(\hat{\alpha})\log(t_i)[1+\log(\hat{h}_i)]+1}{\hat{p}_i+(1-\hat{p}_i)\hat{h}_i} - \\ \dfrac{(\hat{\alpha})\log(\hat{h}_i)(1-\hat{p}_i)\log(t_i)\hat{h}_i}{[\hat{p}_i+(1-\hat{p}_i)\hat{h}_i]^2}\Big\}, & \text{if} \quad i \epsilon C. \end{cases}$$

the elements of vector $\mathbf{\Delta}_2$ are expressed as

$$\Delta_{2i} = \begin{cases} (S_t)(\hat{\alpha})\log(\hat{h}_i)(t_i)^{-1}, & \text{if} \quad i \epsilon F; \\ (S_t)(1-\hat{p}_i)(\hat{\alpha})\log(\hat{h}_i)(t_i)^{-1}\Big\{\dfrac{1+\log(\hat{h}_i)}{\hat{p}_i[\hat{h}_i-1]+1} - \dfrac{(1-\hat{\hat{p}}_i)\log(\hat{h}_i)}{[\hat{p}_i(\hat{h}_i-1)+1]^2}\Big\}, & \text{if} \quad i \epsilon C. \end{cases}$$

and the elements of the vector $\mathbf{\Delta}_j$, $j = 3, \ldots, p+2$ are expressed as

$$\Delta_{ji} = \begin{cases} 0, & \text{if} \quad i \epsilon F; \\ -(x_{ij})(S_t)(\hat{p}_i)(\hat{\alpha})\log(\hat{h}_i)(t_i)^{-1}\hat{h}_i \\ \Big\{\dfrac{1}{[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}][\hat{p}i+(1-\hat{p}i)\hat{h}_i]} + \dfrac{(1-\hat{p}_i)(1-\hat{h}_i)}{[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}][\hat{p}_i+(1-\hat{p}_i)\hat{h}_i]^2}\Big\}, & \text{if} \quad i \epsilon C. \end{cases}$$

where

$$\hat{h}_i = \exp\{-t_i^{\hat{\alpha}} e^{\hat{\lambda}}\} \quad \hat{p}_i = \frac{\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}.$$

Appendix D: Local influence on predictions: Explanatory variable perturbation

In this appendix we provide the derivatives of elements $\mathbf{\Delta}_{ij}$ of matrix $\mathbf{\Delta}$ considering the explanatory variables perturbation scheme. The elements of vector $\mathbf{\Delta}_1$ are expressed as

$$\Delta_{1i} = \begin{cases} 0, & \text{if} \quad i \epsilon F; \\ -\dfrac{\hat{\beta}_t(S_x)(\hat{p}_i)\log(\hat{h}_i)\log(t_i)}{[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}]}\left\{\dfrac{\hat{h}_i}{\hat{p}_i+(1-\hat{p}_i)\hat{h}_i}+\dfrac{(1-\hat{p}i)\hat{h}_i(1-\hat{h}_i)}{[\hat{p}_i+(1-\hat{p}_i)\hat{h}_i]^2}\right\}, & \text{if} \quad i \epsilon C. \end{cases}$$

the elements of vector $\mathbf{\Delta}_2$ are expressed as

$$\Delta_{2i} = \begin{cases} 0, & \text{if} \quad i \epsilon F; \\ -\dfrac{(\beta_t)(S_x)\log(\hat{h}_i)(\hat{h}_i)(\hat{p}_i)}{[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}][\hat{p}_i+(1-\hat{p}_i)\hat{h}_i]^2}, & \text{if} \quad i \epsilon C. \end{cases}$$

the elements of vector $\mathbf{\Delta}_j$, for $j = 3, \ldots, p+2$ and $j \neq t$, take the forms

$$\Delta_{ji} = \begin{cases} -\dfrac{x_{ij}(\hat{\beta}_t)(S_x)(\hat{p}_i)}{[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}]}, & \text{if} \quad i \epsilon F; \\ -x_{ij}(\hat{p}_i)(S_x)(\hat{\beta}_t)(1-\hat{h}_i)\left\{\dfrac{\hat{p}_i(1-\hat{h}_i)}{[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}]^2[\hat{p}_i+(1-\hat{p}i)\hat{h}_i]^2}\right. & \\ \left.-\dfrac{[1-\exp\{x_i^T\hat{\boldsymbol{\beta}}\}]}{[\hat{p}_i+(1-\hat{p}_i)\hat{h}_i][1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}]^2}\right\}, & \text{if} \quad i \epsilon C. \end{cases}$$

the elements of vector $\mathbf{\Delta}_t$ are given by

$$\Delta_{ti} = \begin{cases} -(S_x)(\hat{p}_i)\left[1+\dfrac{x_{it}\hat{\beta}_t}{[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}]}\right], & \text{se} \quad i \epsilon F; \\ \dfrac{(S_x)(\hat{p}_i)(1-\hat{p}_i)^2(1-\hat{h}_i)}{[\hat{p}_i+(1-\hat{p}_i)\hat{h}_i]}\left\{1+\hat{\beta}_t x_{it}[1-\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}]\right\} & \\ -\dfrac{(x_{it})(S_x)\hat{\beta}_t(\hat{p}_i)(1-\hat{p}_i)^2[1-\hat{h}_i]^2}{[\hat{p}_i+(1-\hat{p}i)\hat{h}_i]^2}, & \text{se} \quad i \epsilon C. \end{cases}$$

where

$$\hat{h}_i = \exp\{-t_i^{\hat{\alpha}}e^{\hat{\lambda}}\} \quad \hat{p}_i = \frac{\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}.$$

Appendix E: Generalized leverage

In this appendix we provide the derivatives of elements $\mathbf{D}_\alpha, \mathbf{D}_\lambda, \mathbf{D}_{\boldsymbol{\beta}}$, of matrix $\mathbf{D}_{\boldsymbol{\theta}}$ considering generalized leverage.

The elements of vector $\mathbf{D}_{\boldsymbol{\theta}}$ are expressed as

$$\mathbf{D}_\alpha = (1-\hat{p}_i)(\hat{\alpha})^{-2}\exp\Big\{-\frac{\hat{\lambda}}{\hat{\alpha}}\Big\}\Big\{\log\Big(\frac{\hat{\alpha}+1}{\hat{\alpha}}\Big)+\Big(\frac{\hat{\alpha}+1}{\hat{\alpha}}\Big)\Big\}.$$

$$\mathbf{D}_\lambda = (1-\hat{p}_i)(\hat{\alpha}^{-1})\Big(-\exp\Big\{-\frac{\hat{\lambda}}{\hat{\alpha}}\Big\}\Big)\Big\{\log\Big(\frac{\hat{\alpha}+1}{\hat{\alpha}}\Big)\Big\}.$$

$$\mathbf{D}_{\beta_j} = (x_{ij})(\hat{p}_i)\Big[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}\Big]^{-1}\exp\Big\{-\frac{\hat{\lambda}}{\hat{\alpha}}\Big\}\log\Big(\frac{\hat{\alpha}+1}{\hat{\alpha}}\Big).$$

where

$$\ddot{\mathbf{L}}_{\boldsymbol{\theta}\mathbf{t}} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\mathbf{t}^T} = \begin{pmatrix} \ddot{\mathbf{L}}_{\alpha t_i} \\ \ddot{\mathbf{L}}_{\lambda t_i} \\ \ddot{\mathbf{L}}_{\beta_j t_i} \end{pmatrix}$$

with

$$\ddot{\mathbf{L}}_{\alpha t_i} = \begin{cases} t_i^{-1} - \exp\{\hat{\lambda}\}t_i^{\hat{\alpha}-1}[\hat{\alpha}\log(t_i)+1], & \forall i: i\,\epsilon\, F; \\ -\hat{g}_i^{-2}(1-\hat{p}_i)\exp\{\hat{\lambda}\}\hat{h}_i t_i^{\hat{\alpha}-1}\log(t_i) & \\ \Big\{\hat{g}_i(-\exp\{\hat{\lambda}\}\hat{\alpha}t_i^{\hat{\alpha}}+\hat{\alpha}+[\log(t_i)]^{-1})- & \\ (1-\hat{p}_i)\hat{h}_i\exp\{\hat{\lambda}\}\hat{\alpha}t_i^{\hat{\alpha}}\Big\}, & \forall i: i\,\epsilon\, C. \end{cases}$$

$$\ddot{\mathbf{L}}_{\lambda t_i} = \begin{cases} -\hat{\alpha}t_i^{\hat{\alpha}-1}\exp\{\hat{\lambda}\}, & \forall i: i\,\epsilon\, F; \\ \hat{g}_i^{-2}(1-\hat{p}_i)\exp\{\hat{\lambda}\}\hat{h}_i\hat{\alpha}t_i^{\hat{\alpha}-1}\Big[\hat{g}_i\hat{\alpha}t_i^{\hat{\alpha}}+(1-\hat{p}_i)\exp\{\hat{\lambda}\}t_i^{\hat{\alpha}}\hat{h}_i\Big], & \forall i: i\,\epsilon\, C. \end{cases}$$

$$\ddot{\mathbf{L}}_{\beta_j t_i} = \begin{cases} 0, & \forall i: i\epsilon F; \\ \hat{g}_i^{-2}\hat{p}_i x_{ij}[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}]^{-1}\hat{h}_i\exp\{\hat{\lambda}\}\hat{\alpha}t_i^{\hat{\alpha}-1}\times\Big\{\hat{g}_i-(1-\hat{p}_i)[1-\hat{h}_i]\Big\}, & \forall i: i\,\epsilon\, C. \end{cases}$$

where

$$\hat{h}_i = \exp\Big\{-t_i^{\hat{\alpha}}\exp\{\hat{\lambda}\}\Big\},\quad \hat{g}_i = \hat{p}_i+(1-\hat{p}_i)\hat{h}_i \quad\text{and}\quad \hat{p}_i = \frac{\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}}{1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}}.$$

References

Barlow, W. E., and Prentice, R. L.(1988). Residual for relative risk regression. *Biometrika*, 75, 65-74.

Beckman, R. J., Nachtsheim, C. J. and Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, 29, 413-426.

Bolfarine, H. and Cancho, V. (2001). Modelling the presence of immunes by using the exponentiated-Weibull model. *Journal of Applied Statistics*, 28, 659-671.

Cancho, V., Bolfarine, H. and Achcar, J. A. (1999). A Bayesian analysis for the Exponentiated-Weibull distribution. *Journal Applied Statistical Science*, 8, 227-242.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression.* New York: John Wiley.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society B*, 30, 248-275.

Collet, D. (1994). *Modelling Survival Data in Medical Research.* Chapman and Hall: London.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* , 19 15-18.

Cook, R. D. (1986). Assesment of local influence (with discussion). *Journal of the Royal Statistical Society*, 48 2, 133-169.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hill.

Davison, A. C. and Gigli, A. (1989). Deviance residuals and normal scores plots. *Biometrika*, 76, 211-221.

Doornik, J. (1996). Ox: An Object-Oriented Matrix Programming Language. International Thomson Business Press.

Escobar, L. A. and Meeker, W. Q. (1992). Assessing influence in regression analysis with censored data. *Biometrics*, 48, 507-528.

Fachini, J. B., Ortega, E. M. M. and Louzada-Neto, F. (2007). Influence diagnostics for polyhazard models in the presence of covariates. Statistical Methods and Applications. In Press.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer-Verlag: New York.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Process and Survival Analysis.* Wiley: New York.

Fung, W. K. and Kwan, C. W. (1997). A note on local influence based on normal curvature. *Journal of the Royal Statistical Society B*, 59, 839-843.

Galea, M., Paula, G. A. and Bolfarine, H. (1997). Local influence in elliptical linear regression models. *The Statistician*, 46, 71-79.

Gelman, A. (2004). Running WingBUGS from R, Available for download at (http:carn.us.r-project. org/src/contrib/Descriptions/R2WinBUGS.html).

Gu, H. and Fung, W. K. (1998). Assessing local influence in canonical analysis. *Annals of the Institute of Statistical Mathematics*, 50, 755-772.

Hinkley, D. V., Reid, N. and Snell, E. J.(1991). *Statistical Theory and Modelling-In honour of Sir David Cox*. Chapman & Hall, London.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag: New York.

Kim, M. G. (1995). Local influence in multivariate regression. *Communications in Statistics: Theory and Methods*, 20, 1271-1278.

Kwan, C. W. and Fung, W. K. (1998). Assesing local influence for specific restricted likelihood: Applications to factor analysis. *Psychometrika*, 63, 35-46.

Lawrence, A. J. (1988). Regression transformation diagnostics using local influence. *Journal of the American Statistical Association*, 83, 1067-1072.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley: New York.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, 54, 570-582.

Liu, S. Z. (2000). On local influence for elliptical linear models. *Statistical Papers*, 41, 211-224.

Maller, R. and Zhou, X. (1996). *Survival Analysis with Long-term Survivors*. New York: Wiley.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Edition. Chapman and Hall: London.

Mudholkar, G. S., Srivastava, D. K. and Friemer, M. (1995). The exponentiated Weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, 37, 436-445.

Nelson, W. B. (1990). *Accelerated Testing; Statistical Models, Test Plans and Data Analysis*. New York: John Wiley.

O'Hara, R. J., Lawless, J. F. and Carter, E. M. (1992). Diagnostics for a cumulative multinomial generalized linear model with application to grouped toxicological mortality data. *Journal of the American Statistical Association*, 87, 1059-1069.

Ortega, E. M. M., Bolfarine, H. and Paula G. A. (2003). Influence diagnostics in generalized log-gamma regression models. *Computational Statistics and Data Analysis*, 42, 165-186.

Ortega, E. M. M., Cancho, V. G., Bolfarine, H. (2006). Influence diagnostics in exponentiated -Weibull regression models with censored data. *Statistics and Operation Reserch Transactions*, 30, 171-192.

Ortega, E. M. M., Paula, G. A. and Bolfarine, H. (2007). Deviance Residuals in Generalized Log-Gamma Regression Models with Censored Observations. *Journal of Statistical Computation and Simulation*, 77. In press.

Paula, G. A. (1993). Assessing local influence in restricted regressions models. *Computational Statistics and Data Analysis*, 16, 63-79.

Paula, G. A. (1995). Influence residuals in restricted generalized linear models. *Journal of Statistical Computation and Simulation*, 51, 63-79.

Paula, G. A. (1999). Leverage in inequality constrained regression models. *The Statistician*, 48, 529-538.

Pettitt, A. N. and Bin Daud, I. (1989). Case-weight measures of influence for proportional hazards regression. *Applied Statistics*, 38, 51-67.

Prentice, R. L. (1974). A log-gamma model and its maximum likelihood estimation. *Biometrica*, 61, 539-544.

Stacy, E. W. (1962). A generalization of the gamma distribution. *Ann. Math. Stat.*, 33, 1187-1192.

St. Laurent, R. T. and Cook, R. D. (1992). Leverage and superleverage in nonlinear regression. *Journal of the American Statistical Association*, 87, 985-990.

Thomas, W. and Cook, R. D. (1990). Assessing influence on predictions from generalized linear models. *Technometrics*, 32, 59-65.

Tsai, C. and Wu, X. (1992). *Transformation-model diagnostics*. *Technometrics*, 34, 197-202.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77, 147-60.

Williams, D. A. (1987). Generalized linear model diagnostic using the deviance and single case deletion. *Applied Statistics*, 36, 181-191.

Wei, B., Hu, Y. Q. and Fung, W. K. (1998). Generalized leverage and its applications. *Scandinavian Journal of statistics*, 25, 25-37.

SORT 32 (2) July-December 2008, 141-150

A microbiology application of the skew-Laplace distribution

Olga Julià[1] and Josep Vives-Rego[2]

Universitat de Barcelona

Abstract

Flow cytometry scatter are ofen used in microbiology, and their measures are related to bacteria size and granularity. We present an application of the skew-Laplace distribution to flow cytometry data. The goodness of fit is evaluated both graphically and numerically. We also study skewness and kurtosis values to assess usefulness of the skew-Laplace distribution.

MSC: 62F10, 62GP10

Keywords: skew-Laplace distribution, goodness of fit, bacteria size.

1 Introduction

The counting, sizing and distribution analysis of particles is a task performed in such diverse fields as archaeology, medicine, geology, biology and technology. The distributions most commonly proposed for describing particle size or its logarithm are the normal, the hyperbolic and the skew-Laplace distributions. The normal distribution, widely used in other fields, is unsuitable for the distribution of bacterial size distribution in axenic cultures (Koch *et al.*, 1987; Vives-Rego *et al.*, 1994). Barndorff-Nielsen (1977) and Bagnold (1980) proposed log-hyperbolic as a suitable model for particle size distribution. However this four-parameter model presents some computational difficulties and shows nearly identical hyperbolic distributions for different parameter value combinations (Fieller (1992)). In 1992, Fieller *et al.* presented the skew-Laplace

[1] Departament de Probabilitat, Lògica i Estadística. Universitat de Barcelona. Spain.

[2] Departament de Microbiologia. Universitat de Barcelona. Spain.

Received: November 2007
Accepted: April 2008

distribution as a simple but effective model for particle sizes. It is easily computed and flexible enough, with the flexibility to handle complex data sets. Kotz *et al.,* (1998) have reported several properties, generalizations and applications of the skew-Laplace distribution. More recently, Puig and Stephens (2007) presented two useful goodness-of-fit tests for this distribution.

We applied the skew Laplace to microbiological data in Julià and Vives-Rego (2005), where we reported the suitable skew-Laplace model for the flow cytometry measures (specifically for the side light scatter) of different microorganisms. In the present paper this study is expanded through the introduction of skewness and kurtosis as goodness-of-fit indicators, following the ideas of Puig and Stephens (2007). The flow cytometry data are introduced in Section 2. In Section 3 the skew-Laplace distribution is described along with some properties and maximum likelihood estimators of its parameters. The results of the skew-Laplace and log-skew-Laplace fitting our data are shown in Section 4 together with different ways of assessing the goodness of fit. The biological relevance and potential applications of the fit of cellular parameter to the skew-Laplace distribution is analyzed and discussed in a forthcoming paper (Vives-Rego *et al.* 2008).

2 The forward and scatter flow cytometry data

The data come from flow cytometry which generates two kinds of measure: the forward (FS) and the side scatter (SS). The forward scatter (FS) sensor is a photodiode that collects the laser light scattered at narrow angles (typically 2-11°) from the axis of the laser beam. When light reaches the FS sensor, the sensor generates voltage pulse signals that are proportional to the amount of light that the sensor receives. Sensitivity is enough to detect 0.5 μm particles. The side scatter (SS) is a photodiode sensor that collects the amount of laser light scattered at an angle of about 90° from the axis of the laser beam. The amount of SS is proportional to the granularity of the cell that scatters the laser light. Forward scatter is preferred to side scatter because it shows high signal intensity and is insensitive to sub-cellular structure. Forward scatter is normally assumed to be proportional to bacterial size.

Three microorganisms have been analyzed: strain 31 and strain 41 from the intestinal faeces of laboratory mice Mus musculus, and *Escherichia coli* strain 536. All strains were analyzed after 24 hours of incubation and no treatment was applied. The flow cytometer distributed the forward (or scatter) measures in 1024 channels, giving a number between 1 and 1024 for each cell. Our data are organized in frequency tables. The sample sizes range between 10,000, for *E. coli* and 120,000 for the other two bacteria. For more microbiological details see Julià and Vives-Rego (2005).

3 The skew-Laplace distribution and maximum likelihood estimation

In this Section we introduce the definition of the skew-Laplace distribution and some properties, useful to fit this distribution. An extensive study of this distribution and its properties can be found in Kotz *et. al* (2001).

The skew-Laplace distribution has the following density:

$$f(x;\alpha,\beta,\mu) = \begin{cases} \exp\left(\dfrac{x-\mu}{\alpha}\right)\Big/(\alpha+\beta) & x \le \mu \\ \exp\left(\dfrac{\mu-x}{\beta}\right)\Big/(\alpha+\beta) & x > \mu \end{cases} \tag{1}$$

where α, $\beta > 0$ and $\mu \in \mathbb{R}$. When the logarithm is applied we obtain two straight lines with $1/\alpha$ and $-1/\beta$ slopes that intersect at μ. This fact can be used to check approximately the goodness of fit. The parameter μ is the mode, and in the symmetric case, when $\alpha = \beta$, is also the mean. If X is a random variable skew-Laplace distributed, the mean and variance are:

$$\mathbf{E}[X] = \mu + \beta - \alpha$$

$$\sigma^2 = \alpha^2 + \beta^2.$$

The coefficients of skewness and kurtosis are the following:

$$\gamma_1 = \frac{\mathbf{E}[(X-\mathbf{E}(X))^3]}{\sigma^2} = 2\frac{\beta^3-\alpha^3}{(\alpha^2+\beta^2)^{\frac{3}{2}}}$$

$$\gamma_2 = \frac{\mathbf{E}[(X-\mathbf{E}(X))^4]}{\sigma^4} = 3 + 6\frac{\alpha^4+\beta^4}{(\alpha^2+\beta^2)^2}$$

As it is reported in Puig and Stephens (2007) the skewness value determines the kurtosis value, but not viceversa because the same kurtosis corresponds to γ_1 and $-\gamma_1$. This relationship can be used to assess whether the skew-Laplace is appropriate. The possible values of skewness and kurtosis are $\gamma_1 \in (-2, 2)$ and $\gamma_2 \in [6, 9)$.

The maximum likelihood estimators

The maximum likelihood method is used to estimate the skew-Laplace parameters of (1). The mathematical derivation of those estimators can be found for example in Kotz *et al.* (2001) or Puig and Stephens (2007). The maximum likelihood estimator of μ,

denoted by $\hat{\mu}$, can be obtained by a simple algorithm since $\hat{\alpha}$ and $\hat{\beta}$, maximum likelihood estimators of α and β respectively, have an explicit expression depending on $\hat{\mu}$. Indeed, let $x_1, \ldots, x_n$ be a sample coming from a skew-Laplace distribution: we then consider the following functions:

$$\Delta(\mu) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \mu|$$

$$\psi(\mu) = \Delta(\mu) + \sqrt{\Delta^2(\hat{\mu}) - (\bar{x} - \hat{\mu})^2}.$$

Then, the maximum likelihood estimators are given by:

$$\hat{\mu} = x_j \tag{2}$$

$$\hat{\alpha} = \frac{1}{2}\left(\Delta(\hat{\mu}) - \bar{x} + \hat{\mu} + \sqrt{\Delta^2(\hat{\mu}) - (\bar{x} - \hat{\mu})^2}\right) \tag{3}$$

$$\hat{\beta} = \frac{1}{2}\left(\Delta(\hat{\mu}) + \bar{x} - \hat{\mu} + \sqrt{\Delta^2(\hat{\mu}) - (\bar{x} - \hat{\mu})^2}\right) \tag{4}$$

where x_j is any sample value where the function $\psi(x_j)$ attains its single minimum. Note that ψ could attain its single minimum for two or more x_j sample values.

A simple proof of the derivation of maximum likelihood estimators for the skew-Laplace distribution can be found in Puig and Stephens (2007).

4 Results

In order to see if the flow cytometric scatter data fit the skew-Laplace distribution, we first plot frequencies logarithms versus size values in Figure 1. As we noted in Section 3, two straight lines will appear when the Laplace distribution is appropriate. Even though the parameters of skew-Laplace can be estimated by fitting two straight lines in plots of Figure 1, the maximum likelihood method is preferable. The maximum likelihood estimators are calculated following the steps described in Section 3. In all cases the minimum of function Ψ is reached at only one sample value. Histograms with their estimated skew-Laplace density can be found in Figure 2. Samples with good fits and samples with not such good fits can clearly be seen. Our explanation to the fact that some data sets are not well fitted by the skew-Laplace distribution is that some unknown biological factors are modifying the standard biomass distribution in the culture.

The sample sizes of our data range between 10,000 and 120,000, therefore the p-values of any test of goodness of fit are too small to be useful for comparing goodness of fit. In order to assess the usefulness of the skew-Laplace distribution in Julià and Vives-Rego (2005) we calculated the critical size, N_{crit}. This statistic, proposed by Fieller *et al.*

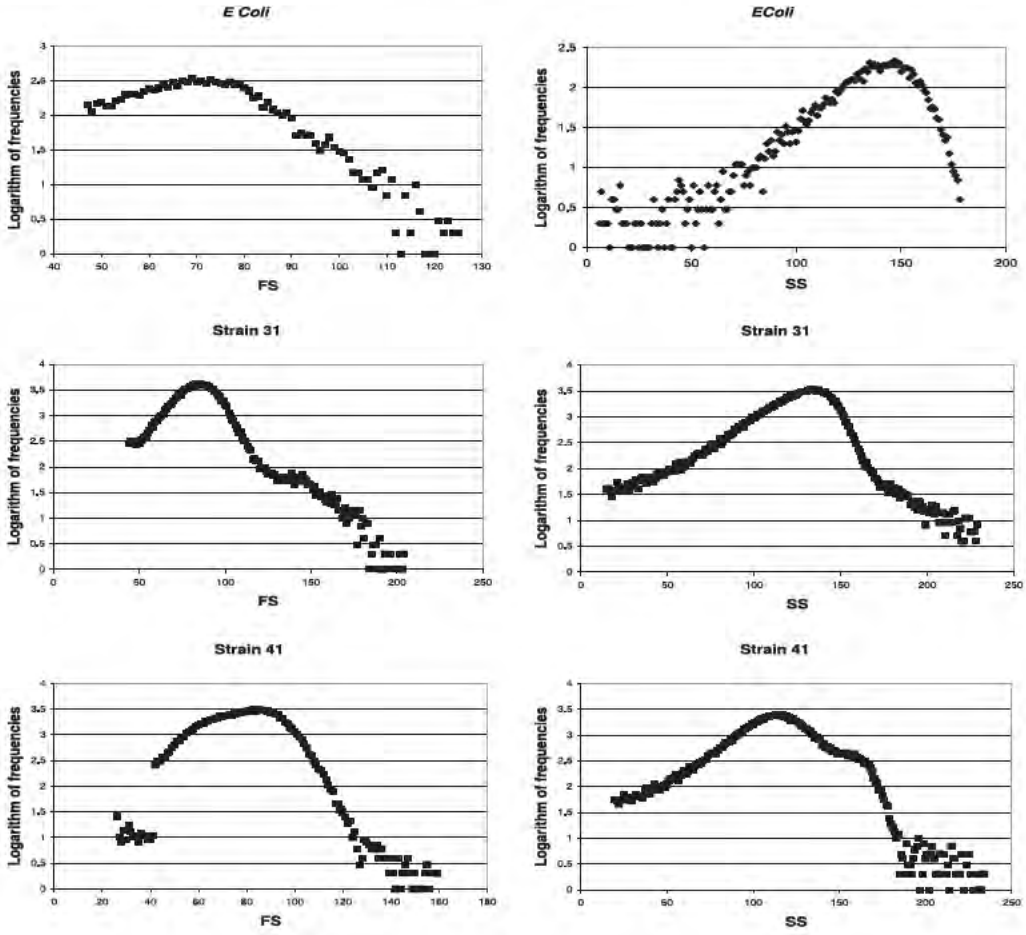


Figure 1: Logarithm of frequencies versus cytometry FS a SS values.

Table 1: *Empirical and skew-Laplace theoretical values of skewness and kurtosis.*

Table 1	empirical		estimated skew-Laplace	
	skewness	kurtosis	skewness	kurtosis
E. Coli FS	0.7570	4.9947	0.5236	6.1838
E. Coli SS	−1.5885	7.4991	−1.5655	7.7273
Strain 31 FS	0.9114	7.4101	0.0167	6.0002
Strain 31 SS	−0.8608	3.9256	−1.7929	8.3221
Strain 41 FS	−0.0981	2.8955	−0.9082	6.5588
Strain 41 SS	−0.3166	5.2102	−0.3040	6.0617

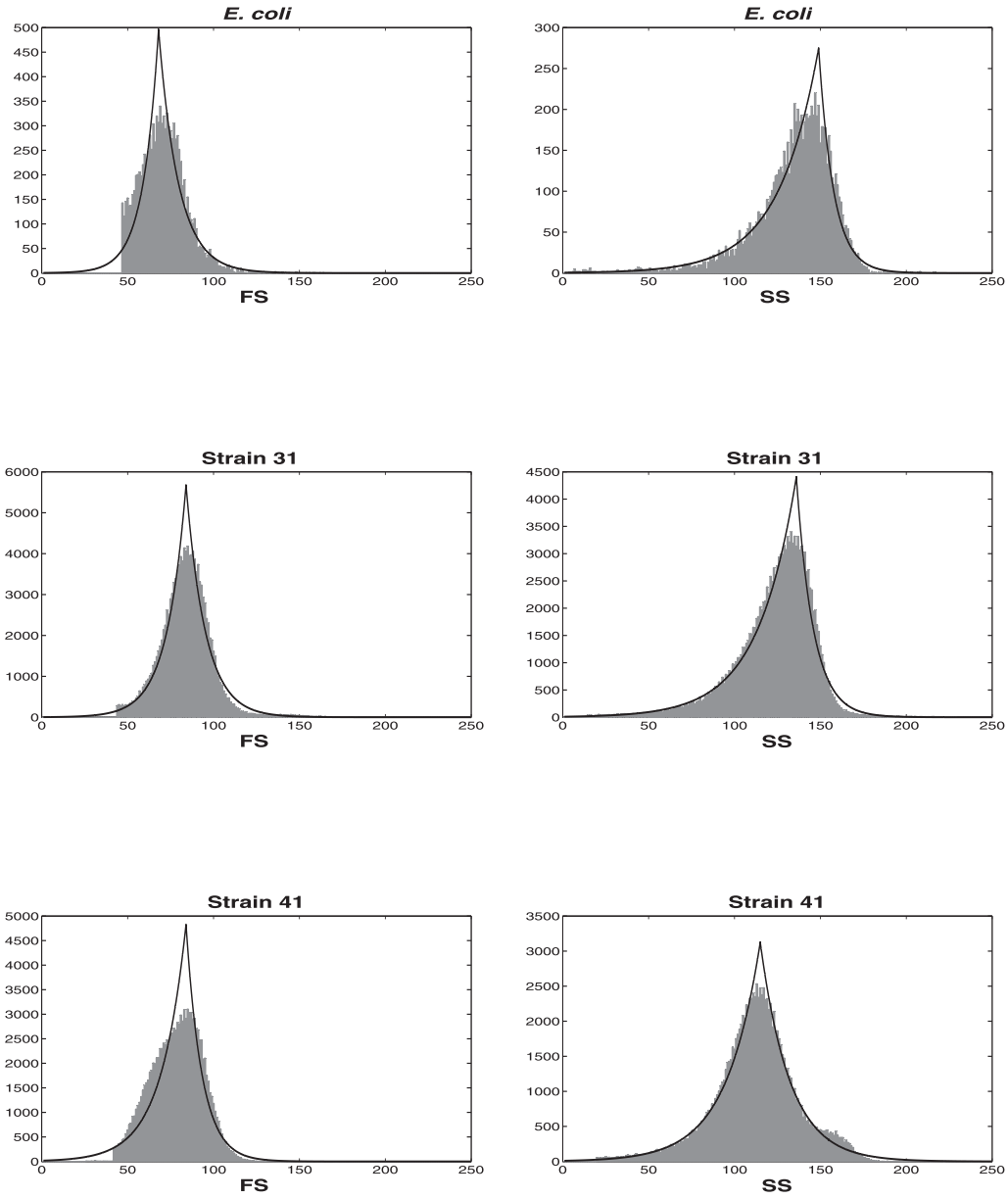


Figure 2: *Skew-Laplace fitting of cytometry FS a SS data. The histogram is shaded in grey and the continuous profile is the estimated skew-Laplace.*

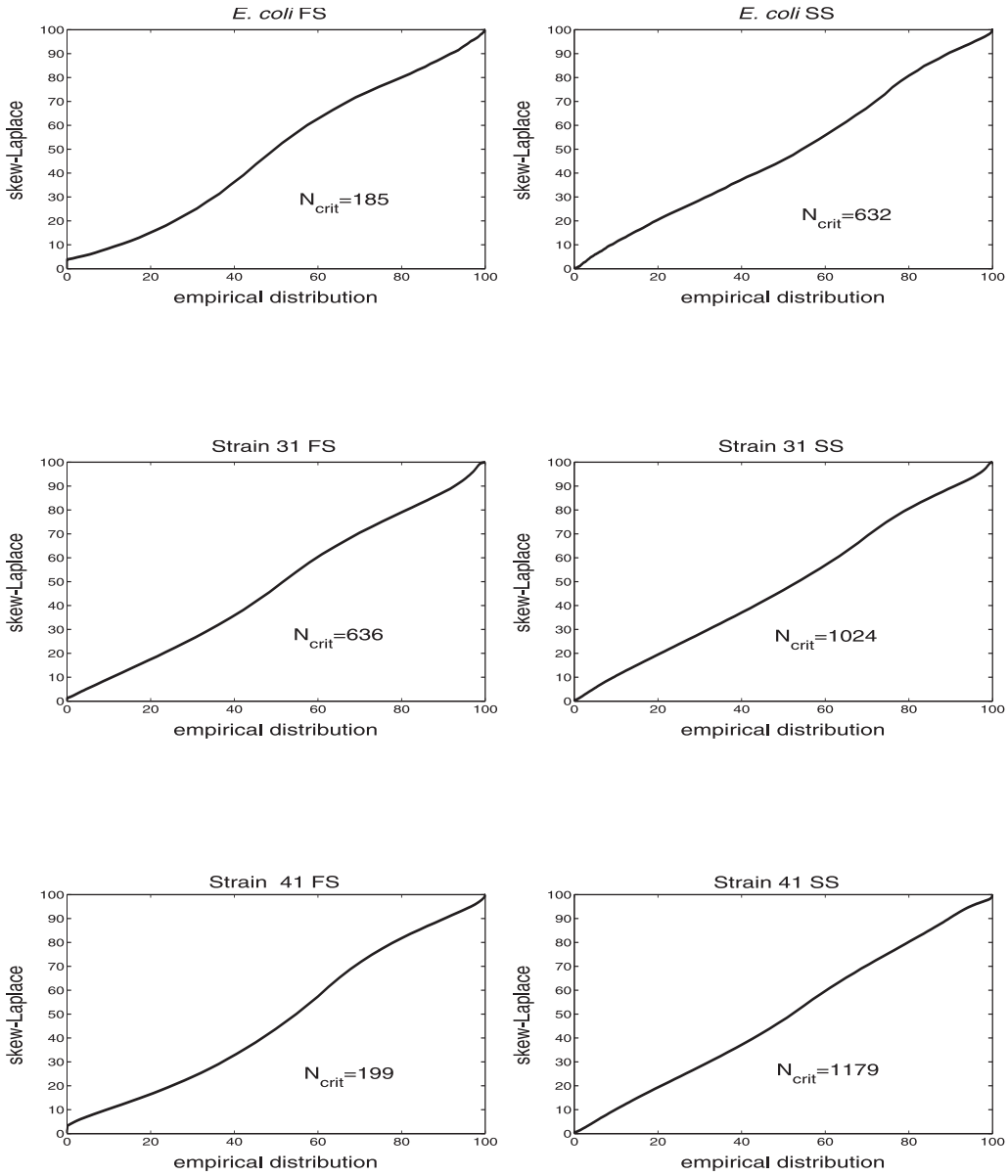


Figure 3: *The q-q plot together N_{crit} values.*

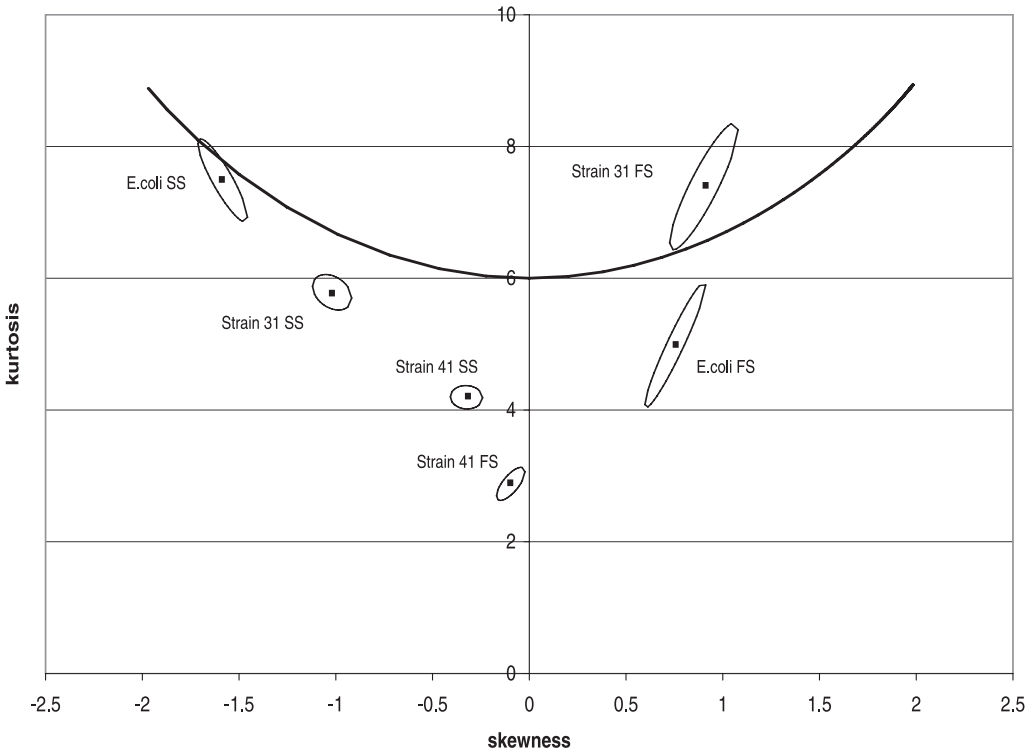


Figure 4: *Skewness and kurtosis of skew-Laplace are represented in solid line, and empirical skewness and kurtosis of cytometry FS and SS values plotted in dots. For each point the 95% confidence region is showed.*

(1992), is based on the chi-square goodness-of-fit test. The N_{crit} statistic is defined as:

$$N_{\text{crit}} = \frac{\chi^2_{k-m-1,\ 0.95}}{\sum_1^k \left(r_i - p_i(\hat{\theta})\right)^2 / p_i(\hat{\theta})}$$

where k represents the number of intervals, m the number of estimated parameters, r_i and $p_i(\hat{\theta})$ are the sample proportion and the estimated skew-Laplace probability of the respective interval. In order to standardize the procedure, we took 40 identical probability intervals for each sample. This statistic can be interpreted as the critical sample size, required just to detect a lack of fit at the 5% level, disregarding the fact that maximum likelihood estimations could be calculated from the grouping data. In Figure 3 the N_{crit} values are shown on each q-q plot. As we can see, greater values of N_{crit} correspond to straighter lines in the q-q plot. In Puig and Stephens (2007) the skewness and kurtosis values are used to build a goodness-of-fit test. Although this test is not appropriate in our case because we have grouped data, we can use this idea to connect the proximity of theoretical skewness and kurtosis values to the empirical values, with

the goodness of fit. The theoretical curve of skew-Laplace skewness and kurtosis and the empirical values are shown in Figure 4. We have also added for each empirical point the 95 % confidence region obtained using the bootstrap method. It can be seen that values near the curve belong to samples with good fit, but it is difficult to assess the goodness of fit according only to their proximity. Table 1 shows the empirical skewness and kurtosis values together with the skew-Laplace values using he estimated parameter.

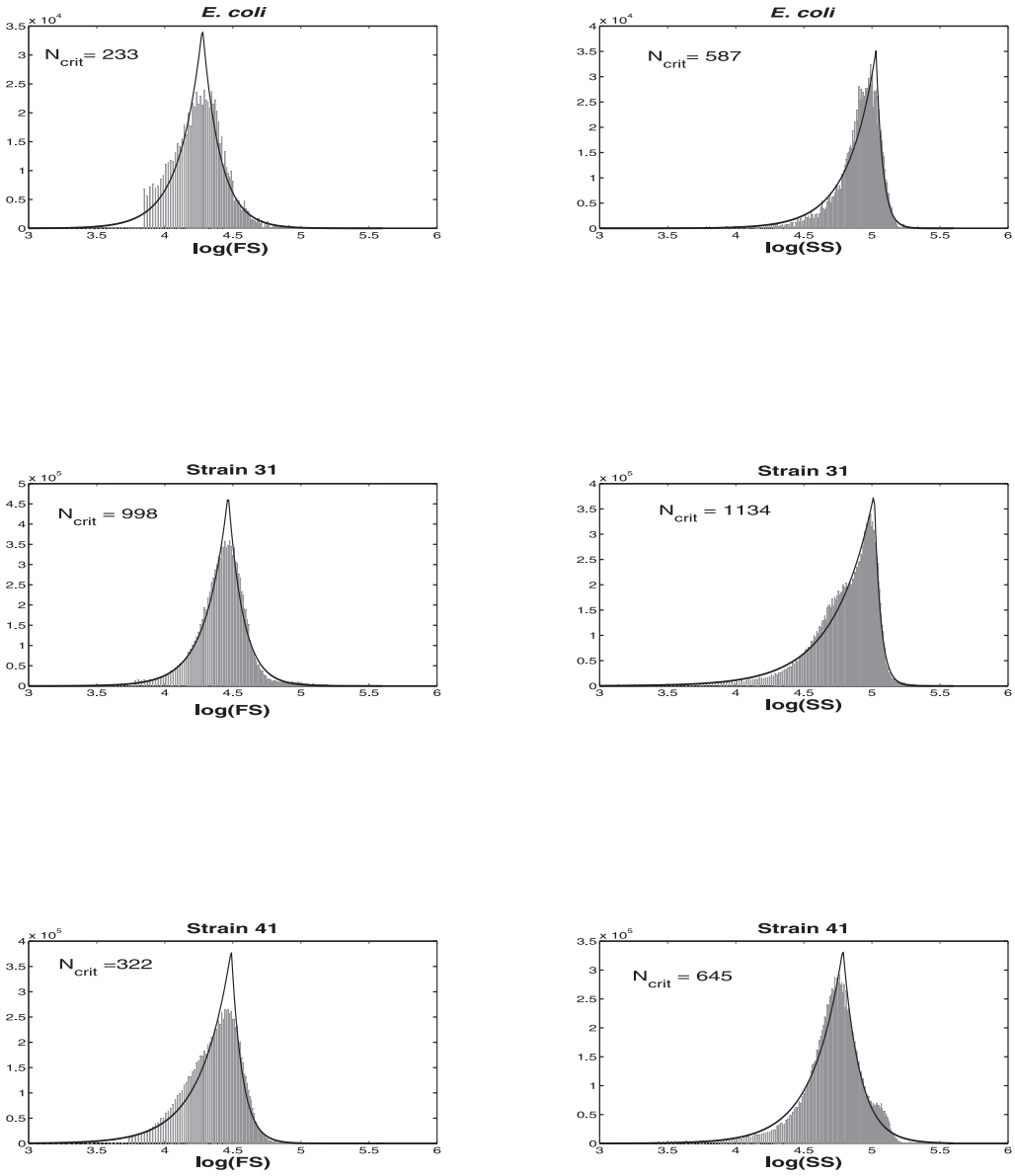


Figure 5: *Log-skew-Laplace fitting of cytometry FS a SS data with N_{crit} values. The histogram is shaded in grey and the continuous profile is the estimated log-skew-Laplace.*

Skew-Laplace versus log-skew-Laplace

According to Fieller (1992) the models based on log-size are more useful not only due to their wide range of particle size, but also because of the multiplicative process of breakage underlying particle production. Even though this multiplicative effect is not clearly applicable to our bacteria size data, we found that in some cases the goodness of fit improves if logarithms are taken. In Figure 5 we can see the histogram and the estimated log-skew-Laplace density. We have also added the respective N_{crit}.

As we can see only in the case of *E. coli* SS and Strain 31 SS no improvement is observable.

Acknowledgments

We thank the referee for their comments and suggestions. This work has been partially supported by grant MTM2005-08886 from Ministerio de Ciencia y Tecnología.

References

Bagnold R. A. and Barndorff-Nielsen, O. (1980). The pattern of natural size distributions. *Sedimentology*, 27, 199-207.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London* A, 353, 401-419.

Fieller, N. R. J., E. C. Flenley and Olbricht, W. (1992). Statistics of particle size data. *Applied Statistics*, 41, 127-146.

Julià, O. and Vives-Rego, J. (2005). Skew-Laplace distribution in Gram-negative bacterial axenic cultures: new insights into intrinsic cellular heterogeneity. *Microbiology*, 151, 749-755.

Koch, A. L. (1987). The variability and individuality of the bacteria. In: Neidhart, C. (Ed.), *Escherichia coli and Salmonella typhymurium Cellular and Mollecular Biology*. American Society for Microbiology, Washington, DC, 1606-1614.

Kotz, S., T. J. Kozubowski and Podgorski, K. (2001). *The Laplace distribution and generalizations*. Birkhäuser, Berlin.

Puig, P. and Stephens, M. A. (2007). Goodness of fit tests for the skew-Laplace distribution. *Statistics and Operations Research Transactions*, 31, 45-54.

Vives-Rego J, López-Amorós, R. and Comas, J. (1994). Flow cytometric narrow-angle light scatter and cell size during starvation of E.coli in artificial sea water. *Letters in Applied Microbiology*, 19, 374-376.

Vives-Rego, J., Julià, O., Vidal-Mas, J. and Panikov, N. S. (2008). *The flow cytometric scatters of bacterial axenic cultures fit the skew-Laplace distribution pattern: biological consequences*. Submitted, preprint No.400 IMUB (http://www.imub.ub.es/publications/preprints/pdf/PreprintN400.pdf)

SORT 32 (2) July-December 2008, 151-176

On equivalence and bioequivalence testing

Jordi Ocaña[1], M. Pilar Sánchez O.[2], Alex Sánchez[1] and Josep Lluís Carrasco[1]

[1]*Universitat de Barcelona and* [2]*Universidad de Chile*

Abstract

Equivalence testing is the natural approach to many statistical problems. First, its main application, bioequivalence testing, is reviewed. The basic concepts of bioequivalence testing (2×2 crossover designs, TOST, interval inclusion principle, etc.) and its problems (TOST biased character, the carry-over problem, etc.) are considered. Next, equivalence testing is discussed more generally. Some applications and methods are reviewed and the relation of equivalence testing and distance-based inference is highlighted. A new distance-based method to determine whether two gene lists are equivalent in terms of their annotations in the Gene Ontology illustrates these ideas. We end with a general discussion and some suggestions for future research.

MSC: 62F03, 62P10, 62F12, 62F15.

Keywords: crossover designs, TOST, intersection-union, distance-based inference, validation of simulation models, Gene Ontology.

1 Introduction and motivation

Consider the following situation: an experimenter has obtained some data under each of two distinct experimental conditions (e.g. control and treated patients), with the objective of demonstrating the *non-existence* (or more properly, the *near non-existence*) of differences between the two experimental conditions –e.g. the absence of an adverse drug reaction. In the subsequent statistical analysis, a known (but still frequent in practice) error is to test a null hypothesis stating equality (e.g. the corresponding means are equal) vs. an alternative hypothesis stating the existence of differences.

[1] Universitat de Barcelona.
[2] Universidad de Chile.

Received: February 2008
Accepted: May 2008

Suppose that the above test has been properly applied. If the computed *p*-value is greater than the previously stated significance level (e.g. a *p*-value of 0.12 when the significance level was 0.05), the null hypothesis will not be rejected. As is well known (but not always taken into account in practice), non-rejection of the null hypothesis is not "proof" of its validity. In such a situation, the *p*-value may be accompanied by some post-experiment or post-hoc power calculations, in order to give more "credibility" to the null hypothesis. Such observed power computations consist of calculating the probability of rejecting the null hypothesis under a distributional setting compatible with the test assumptions, and under parameter values defined by appropriate summary statistics obtained from the data. For example, under the typical Student's *t*-test for comparing the means of two separate groups, the post-experiment power calculations will assume normality and a nature state compatible with the alternative hypothesis, characterised by the estimated difference of means and the value of the pooled variance estimate. The assumed rationale of these power calculations is that a high observed power (e.g. greater than 0.9 or 0.95) "reinforces the credibility" of the null hypothesis, which could not be rejected "even" under such high power or low type II error probability. In fact, observed power calculations do not have any evidential value: see, for example, Hoenig and Heisey (2001).

It might also be possible to find a situation where the *p*-value leads to significant results with a very high power, possibly due to a very large sample size. The null hypothesis of effect *non-existence* would, therefore, be rejected, even if such an effect was negligible, i.e. the effect is statistically but not practically (e.g. clinically) significant.

More tenable approaches are based on improving the evidential value of *p*-values, for example using some sort of *p*-value calibration (preferably under a Bayesian point of view), as in Sellke *et al.* (2001) or in Girón *et al.* (2006), or, in a fully Bayesian setting, using Bayes factors and posterior probabilities, as in Moreno and Girón (2006). Under a frequentist approach, the best policy would be to recognise the inherent asymmetry of the risks associated with both hypotheses, and to invert their roles. If the end goal is to "demonstrate" the non-existence of effect (or more generally, of differences), a more dependable approach would be to establish an alternative hypothesis of "near equality" (not of strict equality) *versus* a complementary null hypothesis of "sufficiently large difference". This approach, where the alternative hypothesis defines the effect *non-existence* as equivalence of parameters rather than strict equality, is taken in equivalence testing. It is also compatible with a Bayesian point of view.

The rest of the paper is organised as follows. The second section is devoted to bioequivalence (BE) testing, by far the most common equivalence situation. This will provide a reference case sufficient to establish the main ideas and problems. The third section is devoted to evaluating the potentialities of this approach in more general terms, to establish their relations with a distance-based approach to statistics and to illustrate these ideas with a new distance-based equivalence test in bioinformatics. The last section brings together these ideas in a final discussion.

2 Bioequivalence testing

2.1 Statement of the bioequivalence problem

When the patent period of a drug is going to expire, the company that developed the brand-name "innovator" product based on this drug may try to develop a new formulation or dosage form with the same active ingredient, in order to extend its market exclusivity. Concurrently, other companies may try to develop generic forms based on the same active principle as the innovator product. To obtain approval of these alternatives, most regulatory agencies, including the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMEA), require proof of "equivalence of average bioavailabilities" or average bioequivalence (ABE), between the brand (innovator) product, commonly referred to as the "reference" (R) product, and the new dosage form or generic copy, commonly referred to as the "test" (T) product. An equivalence trial is much less expensive and easier to perform than a clinical trial for the development of a brand new drug. The former is based on sample sizes usually of 24 to 36 healthy volunteers as experimental subjects, in comparison with the sample sizes of thousands of patients commonly required in the latter.

The concept of bioavailability refers to the rate and extent by which the drug is available at its site of action. This is a complex and multidimensional concept. Quantitatively, it is expressed by several measures obtained from the curve of the concentration of drug in blood or plasma *versus* time, observed in each subject after a single-dose administration. The main bioavailability measures are t_{max}, the time until the maximum concentration is reached, C_{max}, the maximum concentration, AUC_{0-t}, the area under the curve from the dose administration to the last observation time, and $AUC_{0-\infty}$, the area under the curve until infinity. The underlying assumption in the requirement of equivalence of average bioavailability is that, as the drug or active principle is the same in all formulations under comparison, its therapeutic effect depends mainly on its concentration at the site of action, which should be similar for all products.

As summarised in Chow and Liu (2000), several criteria of ABE have been used since the 1970s, according to several regulatory recommendations. These include what are known as the 80/20 rule, the 75/75 rule, the ±20% rule and the (currently most widely used) 80%/125% rule.

The 80/20 rule states that, to declare bioequivalence, these two conditions must be fulfilled:

1. The test and reference means should not be statistically significantly different (commonly at a 5% level).
2. There should be at least an 80% power of detecting differences if the true difference were at least as large as 20% of the *observed* reference average.

Note that condition (ii) is an example of "observed power" computation.

The 75/75 states that at least 75% of the subjects must show a bioavailability value for the new formulation that is at least 75% of the corresponding bioavailability measurement for the reference formulation.

The ±20% rule concludes bioequivalence if the mean bioavailability of the test formulation, μ_T, is within ±20% of the mean of the reference formulation, μ_R, i.e., in terms of a ratio of means, if $0.8 < \mu_T / \mu_R < 1.20$.

Most regulatory agencies (e.g. CDER, 2001) recommend making all analyses for log-transformed data. The 80%/125% rule adapts the preceding criterion to analyses made at a logarithmic scale and, at the same time, enables bioequivalence to be stated in terms of a difference rather than a ratio. If, at the original bioavailability scale, a geometric means ratio between 0.8 and 1.25 is admissible, i.e. $0.8 < m_T / m_R < 1.25 = 1 / 0.8$, assuming that the means of the log-transformed variables correspond to the log-transformed geometric means, this inequality becomes $-0.22314 = \log(0.8) < \mu_T - \mu_R < \log(1.25) = +0.22314$. This is the basis of the ±0.223 rule on the logarithmic scale, equivalent to the 80%/125% rule on the original scale.

The ±20% and the 80%/125% criteria are used in conjunction with inferential procedures to ensure type I and type II error control. The first one requires inferential methods on the ratio of means; and the second one, on the difference.

Metzler (1974) was possibly the first author to recognise the inadequacy of the classical testing approach in bioavailability studies and the need for an equivalence approach, though the need for such an approach in a more general context can be traced back to Lehmann (1959). The reviews by Senn (2001) and Zapater and Horga (1999) are to some extent complementary to the present paper. In the next subsections we review bioequivalence in its most common setting: under a fixed sample size crossover design and for normal log-transformed data.

2.2 Average Bioequivalence. Design and basic statistical analysis

The commonest experimental design in bioequivalence studies is a 2×2 crossover design. In it, each experimental subject receives a single dose of both formulations, *R* and *T*, in one and only one of two possible orders or treatment sequences, *RT* or *TR*. There is always a "washout period" between dose administrations, in order to avoid "carry-over" effects, a possible influence or interaction of the first dose on the second. A sample of $N = n_1 + n_2$ subjects are randomly allocated, n_1 to sequence *RT* and n_2 to sequence *TR*. For a given variable *Y on the logarithmic scale*, say $Y = \log C_{max}$ or $Y = \log AUC_{0-t}$, Y_{ijk} will designate an observation made on the *i*-th individual, in the *j*-th period and the *k*-th sequence, $i = 1, \ldots, n_k$, $j = 1,2$ and $k = 1,2$.

With slight variants, all authors follow the linear model and basic analysis for 2×2 crossover trials proposed by Grizzle (1965). We consider the following underlying linear

model:

$$Y_{ijk} = \mu + P_j + F_{(j,k)} + C_{(j-1,k)} + S_{i(k)} + e_{ijk} \tag{1}$$

where μ is an overall mean, P_j is the fixed effect of the administration period j, $F_{(j,k)}$ is the fixed effect of the formulation administered on the k-th sequence and j-th period, and $C_{(j-1,k)}$ corresponds to the fixed effect of carry-over. It can only occur during the second period.

The possible carry-over effect of the reference formulation from the first period to the second period in sequence 1 is denoted by C_R, while the equivalent effect of the test formulation in sequence 2 is denoted by C_T. Therefore:

$$C_{(j-1,k)} = \begin{cases} C_R \text{ if } j = 2 \text{ and } k = 1 \\ C_T \text{ if } j = 2 \text{ and } k = 2 \\ 0 \quad \text{otherwise} \end{cases}$$

with $C_R = -C_T = C$. Similarly,

$$F_{(j,k)} = \begin{cases} F_R \text{ if } j = k \\ F_T \text{ if } j \neq k \end{cases}$$

with $F_R = F_T = F$, and $P_1 = P_2 = P$ as we consider $\sum_{j=1}^{2} P_j = 0$.

We will designate the formulation effect as $\phi = F_T - F_R = -2F$, the period effect as $\pi = P_2 - P_1 = -2P$ and the carry-over effect as $\kappa = C_T - C_R = -2C$.

$S_{i(k)} \sim N(0, \sigma_S^2)$ represents the random effect of the i-th subject nested in the k-th sequence and $e_{ijk} \sim N(0, \sigma^2_{\tau(j,k)})$ is the random error or residual, or disturbance term. Additionally, we assume independence between all $S_{i(k)}$ and all e_{ijk}, and mutual independence between $\{S_{i(k)}\}$ and $\{e_{ijk}\}$.

Subindex $\tau(j,k)$ in the residual variance indicates a possible dependence on the experimental conditions. Obviously one possibility is constant variance, $\sigma^2 = \sigma^2_{\tau(j,k)}$. We will assume a slightly more general model, with possible dependence on the administered formulation:

$$\sigma^2_{\tau(j,k)} = \begin{cases} \sigma_R^2 \text{ if } j = k \\ \sigma_T^2 \text{ if } j \neq k. \end{cases} \tag{2}$$

The inference on the formulation effect is based on the period difference contrasts for each subject i within each sequence k, $d_{ik} = 0.5\,(Y_{i2k} - Y_{i1k})$. Its expectation and variance are:

$$\begin{aligned} E(d_{ik}) &= \begin{cases} \frac{1}{2}(\pi + \phi + C_R) \text{ if } k = 1 \\ \frac{1}{2}(\pi - \phi + C_T) \text{ if } k = 2 \end{cases} \\ var(d_{ik}) &= \tfrac{1}{4}(\sigma_R^2 + \sigma_T^2). \end{aligned} \tag{3}$$

If $\bar{d}_k = n_k^{-1} \sum_{i=1}^{n_k} d_{ik}$ are the sample means of the period differences, its difference:

$$\bar{D} = \bar{d}_1 - \bar{d}_2 \tag{4}$$

is an unbiased estimate of the formulation effect ϕ, provided that no carry-over is present, *i.e.* if $\kappa = 0$.

The standard error of $\bar{D}$ can be independently estimated by

$$\widehat{se}_{\bar{D}} = \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \hat{\sigma}_d \sqrt{\frac{N}{n_1 n_2}} \tag{5}$$

where

$$\hat{\sigma}_d^2 = \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left(d_{ik} - \bar{d}_k \right)^2 \tag{6}$$

estimates the variance in (3). Alternatively, one may consider that $\hat{\sigma}_d^2$ corresponds to half the ANOVA estimate of the disturbance terms variance:

$$\sigma^2 = \frac{1}{2} \left(\sigma_R^2 + \sigma_T^2 \right), \tag{7}$$

sometimes denoted $\hat{\sigma}_{Res}^2$(for "residual") or $\hat{\sigma}_W^2$ (for "within" subjects).

An alternative way of defining $\bar{D}$ is based on the "least squares" means of the test and reference formulation:

$$\bar{Y}_R = \frac{1}{2} \left(\bar{Y}_{\cdot 11} + \bar{Y}_{\cdot 22} \right) \quad \text{and} \quad \bar{Y}_T = \frac{1}{2} \left(\bar{Y}_{\cdot 21} + \bar{Y}_{\cdot 12} \right),$$

where $\bar{Y}_{\cdot jk} = (1/n_k) \sum_{i=1}^{n_k} Y_{ijk}$ is the average of all observations in the j-th period and k-th sequence. Its difference coincides with (4):

$$\bar{D} = \bar{Y}_T - \bar{Y}_R. \tag{8}$$

The establishment of ABE is stated in terms of an equivalence test for the formulation effect ϕ:

$$\begin{aligned} H_0 &: \phi \leq \theta_1 \quad \text{and} \quad \phi \geq \theta_2 \\ H_1 &: \theta_1 < \phi < \theta_2 \end{aligned} \tag{9}$$

where normally $-\theta_1 = \theta_2 = \theta = 0.223$. In the following, if nothing more is specified, we will assume symmetrical equivalence limits, $\pm\theta = \pm 0.223$ for data on the logarithmic scale.

Schuirmann (1987) suggested decomposing the above hypothesis testing problem in two one-tail hypothesis testing problems:

$$\begin{array}{ll} H_{01} : \phi \leq \theta_1 & \\ H_{11} : \phi > \theta_1 & \end{array} \quad \text{and} \quad \begin{array}{ll} H_{02} : \phi \geq \theta_2 & \\ H_{12} : \phi < \theta_2 & \end{array} \tag{10}$$

and to conclude ABE, if and only if both H_{01} and H_{02} were rejected at a chosen α nominal level of significance (e.g. 0.05). The one-sided tests are easily implemented, since the statistic

$$T = \frac{\bar{D} - \phi}{\widehat{se}_{\bar{D}}} \tag{11}$$

follows a Student's central distribution, with $N - 2$ degrees of freedom. This provides an α level test, as a direct consequence of the intersection-union principle: see Berger (1982) and Berger and Hsu (1996).

The above procedure, known as the Two One-Sided Test (TOST) procedure, is operationally equivalent to the "confidence interval inclusion principle", say, to declare ABE if the usual $1 - 2\alpha$ shortest confidence interval:

$$\bar{D} \pm t_{(\alpha,N-2)}\ \widehat{se}_{\bar{D}} \tag{12}$$

where $t_{(\alpha,N-2)}$ is the $1 - \alpha$ quantile of a Student's t distribution with $N - 2$ degrees of freedom, is fully included in the bioequivalence limits, $[\theta_1, \theta_2]$. This principle was first pointed out by Westlake (1972). See Wellek (2003) for a discussion in more general terms. Declaring ABE if the 90 % interval (12) for log-transformed data is included between the limits ±0.233 is the current methodological mainstream in bioequivalence studies.

The use of a $1 - 2\alpha$ interval for a test of size α may seem counter-intuitive. As is shown in Munk and Pflüger (1999) in more general terms, this relation between confidence and test size requires the fulfilment of two conditions: convexity of the parametric region associated with the alternative hypothesis and an equivariance property of the confidence interval. If $I_{1-2\alpha}$ stands for a $1–2\alpha$ confidence interval, particularising the equivariance condition to the inference problem considered here, it may be stated as $I_{1-2\alpha}\left(d_\phi\left(\bar{D}\right), \widehat{se}_{\bar{D}}\right) = 2\phi - I_{1-2\alpha}\left(\bar{D}, \widehat{se}_{\bar{D}}\right)$ with respect to the transformation $d_\phi(x) = 2\phi - x$. This equivariance condition is fulfilled by (12) but relaxing this requirement in other confidence intervals leads to $1 - \alpha$ confidence intervals associated to α size tests. This is the case for the three confidence intervals described below.

Westlake (1976), from controversial considerations on the need of symmetry for any bioequivalence decision rule, introduced the following confidence interval:

$$\left[\bar{D} - t_2\ \widehat{se}_{\bar{D}}, \quad \bar{D} - t_1\ \widehat{se}_{\bar{D}}\right] \tag{13}$$

where t_1 and t_2 must satisfy the equations

$$\begin{aligned} &\Pr\{t_1 < T < t_2\} = 1 - \alpha \\ &(t_1 + t_2)\,\widehat{se}_{\bar{D}} = 2\bar{D} \end{aligned} \tag{14}$$

in order to define a symmetric around-zero interval, with 100 % coverage if the true formulation effect (in the logarithmic scale) is null, $\phi = 0$, and coverage tends to $1 - \alpha$ as ϕ tends to infinity. It ensures a bioequivalence test of size α when the interval inclusion rule is applied. The computation of (13) needs a trial-and-error iteration.

Hsu *et al.* (1994) introduced the following intervals not requiring any trial and error step, also with confidence level $1 - \alpha$ and associated with bioequivalence tests of size α:

$$\pm\left(\left|\bar{D}\right| + t_{(\alpha, n_1+n_2-2)}\widehat{se}_{\bar{D}}\right) \tag{15}$$

and

$$\left[\min\left(0,\ \bar{D} - t_{(\alpha, n_1+n_2-2)}\widehat{se}_{\bar{D}}\right),\quad \max\left(0,\ \bar{D} + t_{(\alpha, n_1+n_2-2)}\widehat{se}_{\bar{D}}\right)\right]. \tag{16}$$

Interval (15) is symmetrical and both have asymptotic confidence level $1 - \alpha$, and 100 % coverage if $\phi = 0$. By construction, there is an inclusion relation in the sense of (16) $\subset$ (15) $\subset$ (13). Thus, from (13) to (16), the power of the corresponding α level tests is not decreasing, and possibly increases.

The properties of the above intervals, and their relation to α level tests, are summarized in Chow and Shao (2002).

Rodda and Davis (1980) interpreted the confidence interval inclusion principle from a Bayes point of view. Under model (1) and no carry-over effect, the statistics $\bar{d}_1$, $\bar{d}_2$ and $(N-2)\,\hat{\sigma}_d^2$ are independently distributed, $\bar{d}_k \sim N\left(\xi_k, \frac{\sigma^2}{2}\right)$ with $\xi_1 = 0.5\,(\pi + \phi)$, $\xi_2 = 0.5\,(\pi - \phi)$ and $(N-2)\,\hat{\sigma}_d^2 \sim \frac{\sigma^2}{2}\chi^2_{N-2}$. Assuming independent and locally uniform non-informative priors for ξ_1, ξ_2 and σ^2, it is finally found that the posterior distribution of

$$\frac{\phi - \bar{D}}{\widehat{se}_{\bar{D}}} \tag{17}$$

is a central Student's t with $N - 2$ degrees of freedom. This allows a probabilistic interpretation in terms of credible intervals. For example, the $1 - 2\alpha$ highest density interval is computationally identical to the shortest confidence interval (12). But now, declaring bioequivalence when it is included within the bioequivalence limits $\pm\theta$ may be interpreted as imposing the condition that the posterior probability of $-\theta < \phi < +\theta$ must be no less than $1 - 2\alpha$.

2.3 An ABE study example

We illustrate the preceding basic bioequivalence analyses with the results of Al Mohizea *et al.* (2007), a bioequivalence study on two forms (the new form gemifloxacin 320 mg/tablet vs. the reference form factive 320 mg/tablet) of the antibiotic Gemifloxacin. The study was performed in 24 healthy volunteers under a 2×2 crossover design.

		AUC_{0-t}		$AUC_{0-\infty}$		C_{max}	
		Interval limits					
		Lower	Upper	Lower	Upper	Lower	Upper
Interval type	"Shortest" (12)	87.48	107.83	88.72	108.19	92.08	113.47
	Westlake (13)	87.12	114.79	88.15	113.44	87.55	114.22
	Symmetric (15)	87.48	114.32	88.71	112.72	88.13	113.47
	"Optimal" (16)	87.48	107.83	88.71	108.19	92.08	113.47

The $\bar{D}$ values for log AUC_{0-t}, log $AUC_{0-\infty}$ and log C_{max} were –0.0292, –0.0205 and 0.0220, respectively. The standard errors $\widehat{se}_{\bar{D}}$ for the same pharmacokinetic measures were 0.0609, 0.0578 and 0.0608. These results lead to the following standard "shortest" 90 % confidence interval for log AUC_{0-t}:

$$\bar{D} \pm t_{(\alpha,N-2)}\ \widehat{se}_{\bar{D}} = -0.0292\ \pm\ 1.7171 \times 0.0609 = [-0.1338,\ 0.0754]$$

which leads to the following confidence interval for the ratio in the original scale: $[\exp(-0.1338), \exp(0.0754)] = [0.8748, 1.0783]$ or [87.48, 107.83] in percentage terms.

As this interval is included in the bioequivalence limits [80, 125], bioequivalence should be declared. Similarly, the confidence intervals for the other variables are indicated in the table above, with a similar conclusion of bioequivalence.

Equivalently, the *p*-value for the upper and lower Schuirmann's TOST is less than 0.0001 in all cases.

The limits of the alternative confidence intervals discussed in the previous section are also shown in the table. In all cases bioequivalence should be declared. Note that to compute the 95 % Westlake confidence interval, we first need the limits satisfying equations (14). For example, these values are $t_1 = -2.7442$ and $t_2 = 1.7845$ for AUC_{0-t}.

2.4 The power of the TOST procedure and scaling methods

The power of the TOST test, or its interval inclusion equivalent, can be computed as:

$$\beta(\phi,\sigma) = \int_0^{\nu(\theta/\sigma;\, n_1, n_2, \alpha)} \left[\Phi\left(\sqrt{\frac{n_1 n_2}{N}}\frac{(\theta-\phi)}{\sigma} - t_{(\alpha,\,N-2)}\nu\right) - \Phi\left(\sqrt{\frac{n_1 n_2}{N}}\frac{(\theta+\phi)}{\sigma} - t_{(\alpha,N-2)}\nu\right)\right] \sqrt{N-2}\; g_\chi\left(\sqrt{N-2}\nu\right) d\nu \tag{18}$$

where g_χ stands for the χ-distribution with $N-2$ degrees of freedom and

$$\nu\left(\theta/\sigma; n_1, n_2, \alpha\right) = \sqrt{\frac{n_1 n_2}{N}}\frac{\theta}{\sigma}$$

(see Wellek, 2003, p. 211).

The most obvious consequence of (18) is that, for a fixed sample size, $\beta(\phi, \sigma) \to 0$ as $\sigma \to \infty$, for any value of the formulation effect, ϕ. This means that there are alternatives with $\beta(\phi,\ \sigma) < \alpha$. This biased character of the TOST procedure has great practical importance in "high-variability" (HV) drugs or drug products. These are products containing drugs of poor pharmaceutical quality as a cause of their high variability.

A drug is assumed to be HV when the observed coefficient of variation *CV* (on the original scale) associated with the ANOVA estimate $\hat{\sigma}^2_{Res} = 2\hat{\sigma}^2_d$ exceeds 30 % (Blume and Midha, 1993). Sometimes this threshold is put at 25 %. The coefficient of variation on the original scale is related to variance on the logarithmic scale by means of the relation:

$$CV\left(\sigma^2\right) = \sqrt{\exp\left(\sigma^2\right) - 1}. \tag{19}$$

A general discussion on HV drugs analysis can be found in Shah *et al.* (1996). The main problem with HV drugs is the low power of the TOST procedure when used with the usual sample sizes in bioequivalence trials: at most, a few dozen subjects. Bioequivalence trials with hundreds or even thousands of individuals, which would be required for some HV drugs, are generally considered unfeasible.

Anderson and Hauck (1983) proposed a procedure that is more powerful than TOST, but does not adequately control Type I error probability. The test of Berger and Hsu (1996) is nearly unbiased and uniformly more powerful than TOST, but it is not widely used in practice, possibly due to its (moderate) complexity and because the rejection region includes values of $\bar{D}$ outside the limits $\pm\theta$ for large values of $\widehat{se}_{\bar{D}}$. This counter-intuitive character was pointed out by Schuirmann in the discussion accompanying Berger and Hsu (1996), and questioned in Perlman and Wu (1999) in a well-founded argument. This latter paper (which adopts a Fisherian perspective) seems to continue the debate between Fisher and Neyman (see Barnett, 1999). In fact, the approaches that are commonly taken in practice rely on widening in some way the alternative hypothesis (i.e., the bioequivalence limits), which may also seem arbitrary.

Widening the bioequivalence limits to new fixed values has been regulated by the FDA (70 %/143 % or ±0.3567 in logarithmic scale), which assumed the proposals in Shah *et al.* (1996), and by the EMEA (75 %/133 % or ±0.2877) in CPMP (2001). These proposals mainly refer to C_{max}, the bioavailability measure most frequently found to be HV. Clearly, these enlargements do not solve in a general way the power problems of the ABE testing procedures.

Boddy *et al.* (1995) suggested linearly scaling the bioequivalence limits in function of variability, jointly with deciding bioequivalence in the usual way, according to the confidence interval inclusion principle based on the classical shortest interval. Under this setting, the bioequivalence limits (*BEL*) become $\mp k\sigma_{SC}$ instead of a fixed quantity, $\pm\theta$. In 2×2 crossover studies, the most reasonable choice for the scaling variance σ^2_{SC} is the residual variance σ^2. As it is unknown, it must be replaced by an appropriate estimate. Then, the scaled bioequivalence limits become a random function of data:

$$BELsc = \mp k\hat{\sigma}_{SC}. \tag{20}$$

There are some possibly reasonable choices for the constant k (1.116 in CDER, 2003; 1.0 in Boddy *et al.*, 1995; and 0.759 in Tothfalusi and Endrenyi, 2003), but in any case the choice is somewhat arbitrary. The selection of the constant k should be drug-specific and the responsibility of regulatory agencies. Additionally, for a sufficiently large estimated variance, bioequivalence will be declared for $\bar{D}$ values far from the usual bioequivalence limits, a similar criticism to the one about the Berger and Hsu (1996) method. To try to mitigate these drawbacks, families of more flexible scaled limits were developed. Technical details, with an illustrative example, can be found in http://hdl.handle.net/2072/5456.

Apart from their possible arbitrariness, all these bioequivalence limit functions share the same problem: the size and in general the statistical properties of the decision criteria based on them are not guaranteed, as they are not based on known and constant bioequivalence limits, but on limits that are random functions of data, and no additional theoretical support is provided. As is done with individual and population bioequivalence (see below), the bootstrap method gives a possible approach, but this possibility has still not been sufficiently explored.

A more well-founded approach is to make equivalency inferences on scaled parameters for fixed limits, rather than to scale the equivalency limits. In other words, one may restate the problem as that of establishing bioequivalence from fixed limits using a scaled metric ϕ/σ_{sc}. Again, under 2×2 crossover designs, the most natural choice for scaling variability is the residual variance (7). Then $\sigma_{SC} = \sigma$ and equivalence is stated as:

$$-k < \frac{\phi}{\sigma} < +k. \tag{21}$$

An adequate criterion would be to base the final decision on an appropriate confidence interval or test procedure for this scaled parameter. A direct approach is to use the fact that, on rescaling the previous inequality as

$$-\tilde{k} < \tilde{\phi} < +\tilde{k} \tag{22}$$

with $\tilde{\phi} = (\phi/\sigma)\sqrt{2n_1n_2/N}$ and $\tilde{k} = k\sqrt{2n_1n_2/N}$, the statistic $T = \bar{D}\big/\widehat{se}_{\bar{D}}$ has a non-central Student's t distribution with $N - 2$ degrees of freedom and a non-centrality parameter $(\phi/\sigma)\sqrt{2n_1n_2/N}$. This defines the following $1 - 2\alpha$ confidence interval:

$$t_\alpha(\lambda, N-2) \leq \tilde{\phi} \leq t_{1-\alpha}(\lambda, N-2) \tag{23}$$

where $t_\alpha(\lambda, N-2)$ corresponds to the α quantile of a non-central Student's t distribution with $N - 2$ degrees of freedom and non-centrality parameter T. Bioequivalence is declared if the above confidence interval lies within the limits $\pm\tilde{k} = \pm k\sqrt{2n_1n_2/N}$. Obviously, the choice of k still remains arbitrary; as before. Some reasonable choices may be 1.116, 1 or 0.759.

The above interval inclusion procedure is equivalent to a testing procedure with rejection region of general form

$$\{c_1 < T < c_2\}$$

which is optimal (in the sense of being most powerful unbiased) for a wide class of distributions, including the normal case discussed here, as is extensively shown in Wellek (2003).

The scaled procedure, despite its optimality, is not always accepted as the adequate approach to the bioequivalence problem, which is still primarily articulated in the scale of the means and not of the scaled means.

2.5 The carry-over controversy

As has been mentioned, under model (1), $\bar{D}$ is an unbiased estimator of the true formulation effect ϕ only in absence of carry-over effect.

The analysis of the carry-over effect is straightforward. In order to estimate it, we first form the sums inside each individual, $Y_{ij\cdot} = Y_{ij1} + Y_{ij2}$. Simple computations from model (1) lead to the following expressions:

$$\begin{aligned} &var\,(Y_{ij\cdot}) = 4\sigma_S^2 + \sigma_R^2 + \sigma_T^2 = \sigma_+^2 \\ &E\,(Y_{i1\cdot}) - E\,(Y_{i2\cdot}) = \kappa. \end{aligned} \tag{24}$$

Then, the difference:

$$\hat{\kappa} = \bar{Y}_{\cdot 1\cdot} - \bar{Y}_{\cdot 2\cdot} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{i1\cdot} - \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{i2\cdot} \tag{25}$$

is an unbiased estimator of the carry-over effect with standard error estimated by:

$$\widehat{se}_{\hat{\kappa}} = \sqrt{\frac{\sum_{i=1}^{n_1}\left(Y_{i1\cdot} - \bar{Y}_{\cdot 1\cdot}\right)^2 + \sum_{i=1}^{n_2}\left(Y_{i2\cdot} - \bar{Y}_{\cdot 2\cdot}\right)^2}{n_1 n_2 \left(N-2\right)/N}}. \tag{26}$$

According to standard results (e.g. Chow and Liu, 2000, pp. 60-61) the statistic:

$$\frac{\hat{\kappa} - \kappa}{\widehat{se}_{\hat{\kappa}}} \tag{27}$$

follows an Student's central t distribution with $N - 2$ degrees of freedom if $\kappa = 0$. Grizzle (1965) proposed testing this null hypothesis of non-existence of carry-over at a significance level of $\alpha = 0.1$, or even 0.15, in order to have enough power. In case of no rejection of the null hypothesis, he recommended proceeding with the standard analysis under no carry-over. Otherwise, the recommendation was to use only the data of the first period, like data obtained in a fully randomised parallel trial. This strategy is recommended by the FDA (CDER, 2001).

This two-stage procedure is widely used in practice, despite the criticisms of Brown (1980) in terms of cost and those of Senn (1988) and Freeman (1989) in terms of its inadequate test size and power. Additional arguments against the two-stage procedure are given in Senn (1996) and Senn *et al.* (2004).

When higher-order crossover studies are performed, another possibility is to adjust for the presence of carry-over, assuming specific models for it, as in Laird *et al.* (1992) and Putt and Chinchilli (1999). This approach is also discouraged in Senn (1992), Senn and Lambrou (1998) and Senn *et al.* (2004) in terms of the implausibility of assumptions on carry-over and the analysis' lack of robustness.

Both detractors to the two-stage procedure and opponents to adjusting for carry-over state that the best policy is to not previously test for carry-over (or not use this test to take any further decision on the analysis course) and to proceed as if it was absent. In well-performed experiments, carry-over will usually be absent as 'washout' will normally succeed in eliminating it. This opinion seems to be confirmed in D'Angelo et al. (2001) in their review of 324 two-way and 96 three-way crossover studies. Only a small proportion of these studies, compatible with the common significance level at which they were performed, resulted in a significant carry-over. Moreover, for the subset of these studies reporting the p-value, its empirical distribution was very close to the uniform. With these data, this distributional null hypothesis is never rejected by the Kolmogorov-Smirnov (KS) test (Senn *et al.*, 2004). These results are contested in Putt (2005, 2006) with simulations that suggest the lack of power of these KS tests. Senn *et al.* (2005) rebut the arguments of Putt, arguing the irrelevance of power calculations to interpretation of their observational data.

Putt's analysis is not an observed power study (in the sense of our introductory section). However, as the author herself recognises, it does not prove that carry-over existed – as in Senn *et al.* (2004), results are not a proof of non-existence of carry-over.

These latter authors note the difference between significant and important (sufficient to distort the subsequent analysis) carry-over and the appropriateness of an equivalence approach. This last suggestion refers to their analyses of the possible uniformity of *p*-value data lists, but it would also be a more suitable approach to directly discarding the possible presence of sufficiently distorting carry-over in each particular crossover study.

To perform an equivalence test for *scaled* carry-over:

$$H_0 : \frac{|\kappa|}{\sigma_+} \geq \varepsilon \qquad \text{vs.} \qquad H_1 : \frac{|\kappa|}{\sigma_+} < \varepsilon \tag{28}$$

Wellek (2003, pp. 196-203) uses the fact that the statistic:

$$T = \frac{\hat{\kappa}}{\widehat{se}_{\hat{\kappa}}} \tag{29}$$

follows a non-central *t*-distribution with $N - 2$ degrees of freedom and non-centrality parameter $\varepsilon\sqrt{n_1 n_2/N}$ in the boundaries of the equivalence region. The *p*-value associated with an observed sample value, T_{obs}, may be computed as $\Pr\{F < T^2_{\text{obs}}\}$, where F stands for a random variable with non-central F distribution, with non-centrality parameter $\varepsilon^2 (n_1 n_2/N)$ and 1 and $N - 2$ degrees of freedom. Alternatively, if one wishes to test the absolute carry-over, κ, an approach like Schuirmann's test for bioequivalence can be performed with the "inside each subject sum" data $Y_{ij\cdot}$ and using results from (24) to (27). As the variance in (24) may be high, this last approach may result in a test with low power.

On the other hand, what seems to be the most interesting measurement of the possible disturbing effect of carry-over is its size in relation with the formulation effect, and not its absolute or relative value in relation with variability.

An interesting alternative to the two-stage procedure, although for the moment not sufficiently developed for practical use in bioequivalence testing, is the synthetic estimators of Longford (2001), a way to combine the with-carry-over and without-carry-over formulation effect estimators. Similarly, the Bayesian approach taken in Grieve (1985) avoids taking an all-or-nothing approach for possible carry-over. The usefulness and/or correctness of this approach is contested by some authors, but it is successfully used with real data, as shown by Racine *et al.* (1986).

2.6 Other approaches to BE

An obvious extension of the preceding techniques is to make them multivariate, i.e. to test simultaneously for all available bioavailability measures and not separately for each one. The basic approach to ABE, based on confidence interval inclusion, was generalised to the multivariate case by Wang *et al.* (1999). Ghosh and Gonen (2008) provide a semi-parametric Bayesian solution to ABE. Using Montecarlo Markov Chain

Methods (MCMC), these authors assume a realistic multivariate prior, with dependent parameters.

Pharmacokinetic measures like C_{max} or *AUC* are statistical summaries computed from the concentration-by-time curves, which are the true raw data in bioavailability experiments. Thus, another possible approach to bioequivalence is to directly analyse these curves, either as multivariate data or by any modelling approach that adequately describes the curves. In this context, a standard tool are mixed models, and an adequate approach would be to establish the equivalence of their parameters or, perhaps better, to compute confidence regions for the mean curves. However, up to the authors knowledge, equivalence testing in mixed models is a still unexplored field, with the exception of Rashid (2003).

Even though all these approaches to BE may be very interesting, the regulatory approach exclusively recommends testing pharmacokinetic parameters individually, which makes the multivariate approach rare in theoretical papers and nearly absent in practical work. Next we describe two (univariate) BE approaches that have merited regulatory consideration.

Complete or nearly complete similarity between the means does not imply equivalence between both formulations. If for example the bioavailability of the test formulation is much more variable than the bioavailability of the reference formulation ($\sigma_T^2 \gg \sigma_R^2$), replacing R by T will probably imply some user risks. The concept of "population bioequivalence" (PBE) refers to equivalence both in mean and in variability; and more generally, to equivalence in the general form of the distribution of the bioavailability variable. This concept tries to express the idea that a generic form is fully *prescribable* to a patient who initiates its treatment. However, even when the distributions under T or R are marginally equivalent, it is not guaranteed that R is *exchangeable* with T in a patient who started treatment with R. The concept of "individual bioequivalence" (IBE) tries to reflect this last concept of exchangeability within the same individual.

These concepts were introduced by Anderson and Hauck (1990) and formalised in Schall and Luus (1993). These authors suggested the following aggregate scaled measure of global dissimilarity to define PBE:

$$\frac{\phi^2 + \sigma^2_{totT} - \sigma^2_{totR}}{\sigma^2_{totR}} \tag{30}$$

where

$$\begin{aligned} \sigma^2_{totR} &= var\,(Y_{i11}) = var\,(Y_{i22}) = \sigma^2_S + \sigma^2_R \\ \sigma^2_{totT} &= var\,(Y_{i12}) = var\,(Y_{i21}) = \sigma^2_S + \sigma^2_T \end{aligned} \tag{31}$$

are the "total" variances (that is, including both the between-subject variance, σ^2_S, and the residual variance) of the response under each treatment. The corresponding

moment-based measurement of individual bioequivalence uses the concept of subject-by-formulation interaction that requires higher-order crossover designs and will not be dealt with here.

Measure (30) combines, rather arbitrarily, a squared Euclidean distance, ϕ^2, with a difference of variances. The natural scaling factor for the first summand is residual variance (7) and not the total variance under the reference formulation. Given these and other difficulties with the above-mentioned concept of population bioequivalence, Schall (1995) proposed a criterion based on the probabilities of discrepancy between the responses under the test and the reference formulation, in relation to the same probability when both individuals receive the reference formulation. In a completely different approach, Wellek (2000) proposed a "disaggregate" test in the sense of separately testing for ϕ/σ and for $\sigma^2_{totT}/\sigma^2_{totR}$ and then combining both tests by means of the intersection-union principle.

In the FDA guidance CDER (2001), there are precise instructions for individual and population bioequivalence. But CDER (2003) seems to abandon the requirement of individual and population bioequivalence and to return to average bioequivalence exclusively, perhaps due to the difficulties in these concepts and in their implementation. Moreover, Senn (2001) points out that the concept of exchangeability of drugs is meaningless in clinical terms and only prescribability is useful when the clinician has to decide whether to prescribe a formulation.

3 Equivalence testing: a more general perspective

3.1 Some selected equivalence problems and applications

Bioequivalence is just one of the potential applications of the equivalence testing concept. There are many applications and potential applications of the equivalence concept, focusing either on statistical methodology or on specific fields of application.

Wellek (2003) reviews some common statistical problems that may be treated more adequately under an equivalence approach. These include comparing binomial variables, goodness of fit to a distribution, testing for homoscedasticity and testing for non-importance of interactions, i.e. for additivity in a linear model. Barker *et al.* (2001) perform an extensive (though not complete, as is pointed out by Martín Andrés and Herranz Tejedor, 2002) review of equivalence tests for binomial variables. Bayesian alternatives to some of these tests are discussed in Williamson (2007).

The following are some discussions of the applicability and/or concrete applications of equivalence testing in diverse areas: Stegner *et al.* (1996) in social sciences, Burns and Elswick (2001) in dental clinical trials, Barker *et al.* (2002) in epidemiology and Mecklin (2003) in educational research. Van Steen *et al.* (2005) propose an equivalence

procedure in DNA sequence comparison. This is an example of the distance-based approach to equivalence, to be treated in more detail in the next section.

A problem of central practical importance is simulation model validation. According to Sargent (2005), operational validation is "determining whether the simulation model's output behaviour has the accuracy required for the model's intended purpose over the domain of the model's intended applicability". If the modelled system is observable, the objective methods for validation are, essentially, two (or more) sample comparison methods: the data observed in the real system vs. the generated data experimenting with the model (i.e., simulating). Reynolds and Deaton (1982) and Kleijnen (1999) review hypothesis test methods for validation.

In contradiction with the preceding quoted definition (which in our opinion reflects pretty well the concept that simulation practitioners have in mind), the common approach to model validation states a null hypothesis of *exact* model validity. This strategy leads to severe methodological problems, illustrated by common recommendations (e.g. Sargent, 2005) of not using too large sample sizes (especially from the simulated data side), in order to avoid rejecting adequate (to the goals of the study) models. It is obvious that an equivalence approach would be much more dependable. The authors are not aware of any equivalence approach to simulation model validation, except Robinson and Froese (2004) and the ideas outlined in Warner (2002).

Most likely, in model validation a difficult problem will be to establish the equivalence limits, which may be very application area- and model-dependent. There are some regulations on how to construct and validate simulation models (e.g. http://cdds.ucsf.edu/research/sddgpreport.php is a best-practice document on simulation in drug development), but none of them considers the equivalence approach in depth.

3.2 Equivalence testing and distance-based Statistics

The great majority of the equivalence problems commented on above admit a distance-based representation, with general form:

$$H_0 : d(A, B) \geq d_0 \text{ vs. } H_1 : d(A, B) < d_0 \tag{32}$$

where d is a distance or dissimilarity index, A and B are two objects (distributions, models...) to be compared and d_0 is an equivalence limit. For example, admitting model (1) and in absence of carry-over and period effects, the distributions under T and R are, respectively,

$$A \equiv N\left(\mu_T = \mu + C_T; \sigma^2_{totT}\right) \text{ and } B \equiv N\left(\mu_R = \mu + C_R; \sigma^2_{totR}\right). \tag{33}$$

The ABE distance and criterion are $d(A, B) = |\mu_T - \mu_R| = |\phi| < d_0 (= \theta)$. The index d is a true distance measure under $\sigma^2_{totT} = \sigma^2_{totR}$.

PBE is based on the index:

$$d(A, B) = \frac{(\mu_T - \mu_R)^2 + \sigma^2_{totT} - \sigma^2_{totR}}{\sigma^2_{totR}}. \tag{34}$$

Note that (34) is not a metric distance, nor a reasonable dissimilarity measurement. For example, it is possible that $d(A, B) = 0$, when $\mu_T \neq \mu_R$ and $\sigma^2_{totT} \neq \sigma^2_{totR}$. Index (34) rewards a generic product with less variability than the brand product.

This distance-based approach is explicitly taken in Munk and Czado (1998), using a trimmed version of the pth Mallows distance between distributions:

$$\Gamma_{\alpha,p}(F, G) = (1 - 2\alpha)^{-1} \left\{ \int_{\alpha}^{1-\alpha} \left| F^{-1}(u) - G^{-1}(u) \right|^p du \right\}^{1/p} \quad \alpha \in \left[0, \tfrac{1}{2}\right), p \geq 1 \tag{35}$$

Their asymptotic results allow non-parametric goodness of fit testing, and average and population bioequivalence testing, in a unified way. One drawback of these tests is that their true size exceeds the nominal size, unless large sample sizes (much larger than is usual in bioequivalence testing) are employed.

Dragalin *et al.* (2003) use the squared Kullback-Leibler divergence, $d(f, g) = \Delta_1^2(f, g)$, where

$$\Delta_1(f, g) = \sqrt{I(f, g) + I(g, f)} \tag{36}$$

is the Jeffreys J-divergence based on the Kullback-Leibler information:

$$I(f, g) = E_f \left\{ \log \frac{f(X)}{g(X)} \right\} \tag{37}$$

for densities f and g. (36) is not a distance index, but has reasonable dissimilarity properties.

If f_T and f_R are the densities of (33), PBE is associated with the index:

$$d(f_T, f_R) = \frac{1}{2} \left\{ (\mu_T - \mu_R)^2 + \sigma^2_{totT} + \sigma^2_{totR} \right\} \left(\frac{1}{\sigma^2_{totT}} + \frac{1}{\sigma^2_{totR}} \right) - 2. \tag{38}$$

Equivalent results are also obtained for the exponential family of distributions and for the multivariate normal case. An obvious advantage of the distance approach is that the generalisation to the multivariate case is much more straightforward.

In the univariate normal case, when $\sigma^2_{totT} = \sigma^2_{totR} = \sigma^2_{tot}$, the preceding index defines a scaled BE criterion in relation to total variance, not in relation to residual variance, as in (21).

Approximate inference with the preceding indices is based on the interval inclusion approach. Bioequivalence is declared if the upper limit of the one-sided bootstrap

percentile interval for d falls below d_0. An advantage of using the Kullback-Leibler metric is that the FDA bioequivalence limits can be easily adapted to the corresponding d_0 values, because both are simple functions of the same moments.

3.3 Combining studies based on Gene Ontology

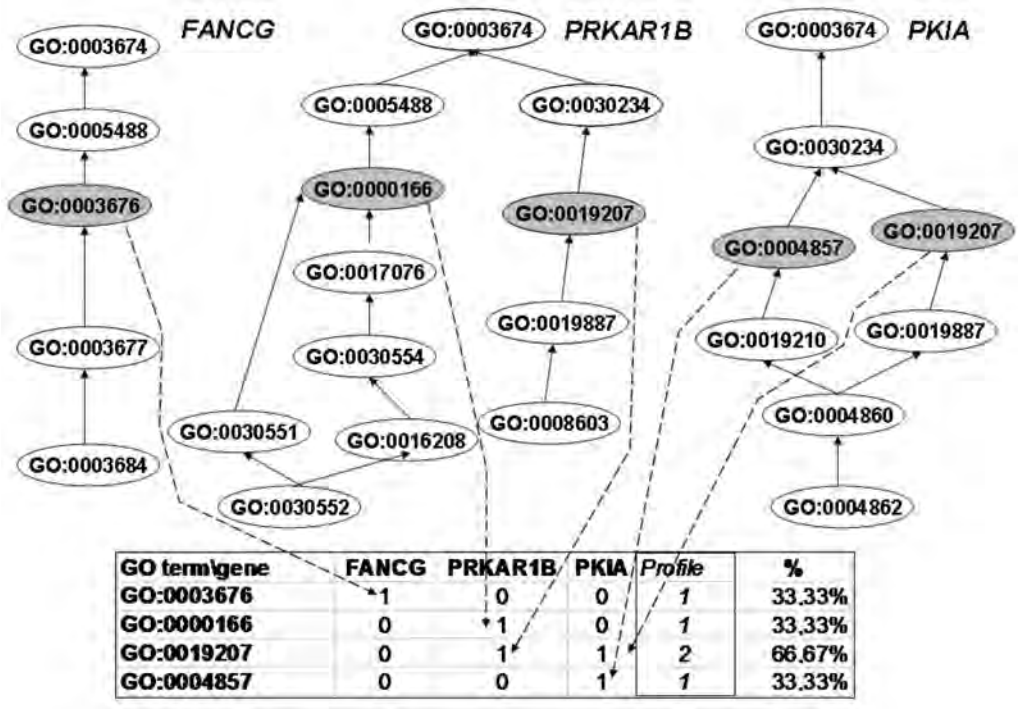
With the help of recently developed technologies like DNA microarrays, it is now possible to analyse the behaviour of thousands of genes in a single experiment. Gene Ontology (GO, www.geneontology.org) is an annotation database created and maintained by the Gene Ontology Consortium in order to systematise these huge amounts of quickly growing information. GO is organised in three basic ontologies: molecular function (MF), biological processes (BP) and cellular components (CC). Each of them can be viewed as directed acyclical graphs (DAG). The nodes in the DAG represent concepts that may help to characterise a gene (e.g. the biological processes in which it participates). The known information on a given gene is expressed as *annotations* or *hits* on one or more nodes in the GO. A way to summarise a given list of genes (e.g. those over-expressed in individuals suffering from a specific disease) is to determine its GO *profile* for a given level in one of the three GO ontologies. A level is the set of all nodes at the same distance from the origin of the ontology; it is like a cross-section in the rich DAG structure. A profile is the vector of annotation counts (or percentages or relative frequencies) in the s nodes of the chosen level, for all genes on the list: $\hat{P} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_s)$. To compute the relative frequency $\hat{p}_i$, one counts the number of annotations in node i. As a given gene may be annotated in two (or more) nodes, these relative frequencies may add more than one.

Figure 1 illustrates these concepts for a list of three human genes (FANCG, PRKAR1B and PKIA) annotated in several nodes (in grey) at level three in the MF ontology. GO nodes are solely identified with a node code, GO:nnnnnnn. Note that the sum of annotation percentages is greater than 100 %, because one of the genes is annotated in two nodes. Thus, direct use of chi-squared tests or related techniques is not adequate for overall comparison of profiles.

In Sánchez *et al.* (2007), a statistical model for GO profiles is provided. It allows a distance-based analysis, using squared Euclidean distance:

$$d\left(\hat{P}, \hat{Q}\right) = \sum_{i=1}^{s} (\hat{p}_i - \hat{q}_i)^2. \tag{39}$$

The comparison of two sample lists of genes, in terms of the squared Euclidean distance over their GO profiles, is studied in Salicrú *et al.* (2008). The comparisons can be made in terms of either a difference problem (i.e., $H_0 : d(P, Q) = 0$ vs. $H_1 : d(P, Q) > 0$) or an equivalence problem (i.e., $H_0 : d(P, Q) \geq d_0$ vs. $H_1 : d(P, Q) < d_0$).



GO term\gene	FANCG	PRKAR1B	PKIA	*Profile*	%
GO:0003676	1	0	0	*1*	33.33%
GO:0000166	0	1	0	*1*	33.33%
GO:0019207	0	1	1	*2*	66.67%
GO:0004857	0	0	1	*1*	33.33%

Figure 1: *Functional profile at level 3 of MF Ontology associated with a list of three genes*

In the most general case, one wishes to compare a sample of n genes with another sample of m genes, with $n = n_1 + n_0$ and $m = m_1 + n_0$. The quantity n_0 corresponds with the number of genes shared by both lists. If $n_0 = 0$, the two lists of genes are mutually excluding: e.g., a set of recessive vs. a set of dominant genes. If $m_1 = 0$, the first list completely includes the second one: e.g., all genes analysed in a microarray vs. those differentially expressed in a given pathology.

Designating as $\hat{P}_1$, $\hat{Q}_1$ and $\hat{P}_0$ the profiles associated with the n_1, m_1 and n_0 genes, respectively, and the sample profiles to be compared as:

$$\hat{P} = \frac{n_0}{n}\hat{P}_0 + \frac{n_1}{n}\hat{P}_1 \quad \text{and} \quad \hat{Q} = \frac{n_0}{m}\hat{P}_0 + \frac{m_1}{m}\hat{Q}_1. \tag{40}$$

If P and Q are the population profiles, the asymptotic distribution of the sample profiles is multivariate normal:

$$\left(\frac{n\,m}{n+m}\right)^{1/2} (\hat{P} - P, \hat{Q} - Q) \xrightarrow{d} Y \sim N(0,\ \Sigma_{PQ}) \approx N(0,\ \Sigma_{\hat{P}\hat{Q}}) \tag{41}$$

with covariance matrix of form:

$$\Sigma_{\hat{P}\hat{Q}} = \begin{pmatrix} \frac{m}{n+m}\left[\frac{n_0}{n}\Sigma_{\hat{P}_0} + \frac{n-n_0}{n}\Sigma_{\hat{P}_1}\right] & \frac{n_0}{n+m}\Sigma_{\hat{P}_0} \\ \frac{n_0}{n+m}\Sigma_{\hat{P}_0} & \frac{n}{n+m}\left[\frac{n_0}{m}\Sigma_{\hat{P}_0} + \frac{m-n_0}{m}\Sigma_{\hat{Q}_1}\right] \end{pmatrix}. \tag{42}$$

The $s \times s$ covariance matrices $\Sigma_{\hat{P}_0}$, $\Sigma_{\hat{P}_1}$ and $\Sigma_{\hat{Q}_1}$have the general form (Sánchez *et al.*, 2007) $\Sigma_{\hat{P}} = (\hat{\sigma}_{ij})$, with

$$\hat{\sigma}_{ij} = \begin{cases} \hat{p}_i(1-\hat{p}_i) & \text{if } i = j \\ \hat{p}_{ij} - \hat{p}_i\hat{p}_j & \text{if } i \neq j \end{cases} \tag{43}$$

where $\hat{p}_{ij}$ designates the relative frequency of genes simultaneously annotated in nodes i and j and possibly also annotated in other nodes.

An asymptotic solution to the equivalence problem (i.e., to test whether both GO profiles are not very dissimilar) may be obtained from:

$$\left(\frac{nm}{n+m}\right)^{1/2}\left\{d\left(\hat{P},\hat{Q}\right) - d\left(P,Q\right)\right\} \xrightarrow{d} Y \sim N\left(0;\omega^2\right) \tag{44}$$

where ω^2can be estimated by:

$$\hat{\omega}^2 = 4\begin{pmatrix} \hat{P}-\hat{Q} \\ -(\hat{P}-\hat{Q}) \end{pmatrix}^{\mathsf{T}} \Sigma_{\hat{P}\hat{Q}} \begin{pmatrix} \hat{P}-\hat{Q} \\ -(\hat{P}-\hat{Q}) \end{pmatrix}. \tag{45}$$

Thus, for a given equivalence limit d_0 and according to squared Euclidean distance, we may conclude equivalence of GO profiles if

$$d\left(\hat{P},\hat{Q}\right) - z_\alpha\hat{\omega}\sqrt{\frac{1}{n}+\frac{1}{m}} < d_0 \tag{46}$$

where z_α corresponds to the α quantile of standard normal distribution. For example, if $\alpha = 0.05$, then we have $z_\alpha = -1.64$.

A possible criterion to establish the equivalence limit d_0 is to fix a maximum allowed discrepancy in each GO node, $|p_i - q_i| < \varepsilon$. Then $d_0 = s\varepsilon^2$, where s is the number of compared nodes.

To illustrate the above ideas, a comparison between two microarray experiments performed by Welsh *et al.* (2001) and Singh *et al.* (2002) to study prostate tumors based on gene expression data is put forward. Although the studies were performed independently, they had similar characteristics in type of tumors, microarray platforms and sample size (see table).

Study	**Platform**	**Sample**
Welsh *et al.*, 2001	HGU95A	32: normal 8, tumor 24
Singh *et al.*, 2002	HGU95Av2	102: normal 50, tumor 52

The comparability of these studies has been exploited by various authors, such as Manoli *et al.* (2006), who used them to compare different microarray data analysis methods, or Moradi *et al.* (2006), who combined them in a predictive analysis (one

data set was used as training set and the other as test set). In either situation the study combination was justified simply on the basis of their common topic, but no quantitative argument was given.

The example below shows the results of the equivalence test performed on the second level of Gene Ontology. The lists of differentially expressed genes were selected using a p-value cutoff of 0.05. The analysis was performed with R package *goProfiles* (Sánchez *et al.*, 2008) available at Bioconductor 2.2 (www.bioconductor.org).

Applying the equivalence tests to the resulting profiles for each of the ontologies gives the following results:

	MF	**BP**	**CC**
Squared Euclidean distance	0.000619	0.001768	0.004081
d_0 threshold for equivalence test (computed as $d_0 = s\varepsilon^2$ with $\varepsilon = 0.05$)	0.037500	0.050000	0.032500
Upper confidence interval limit	0.001329	0.003548	0.006386
Reject null hypothesis of inequivalence	Yes	Yes	Yes

This suggests that it is appropriate to combine the two datasets, as Moradi *et al.* (2006) did.

4 Discussion

Equivalence testing is the most adequate way to address situations where the primary aim is to prove similarity. As is shown with some detail in the case of bioequivalence testing, it is not free from difficulties or controversy, but it does seem to be the most dependable approach to bioequivalence and to many other important problems.

As many of the difficulties with the equivalence approach are essentially technical in nature, solutions to them are likely to be found or, in the worst case, the non-existence of a solution proven. In practice, other questions are more problematic, such as, in our view, the adequate determination of the equivalence limits. Wellek (2003, pp. 11-13) makes some reasonable suggestions regarding the parameters and statistical problems under consideration. However, this problem still depends, to a great extent, on specific areas of application and even on specific problems.

The distance-based approach may be a natural way to include many equivalence problems under the same paradigm, and to permit a smooth path from a univariate to a multivariate approach. There are many distance or dissimilarity indexes that may be adequate. To some extent the decision as to which index to use is arbitrary. Some are adequate due to their simplicity, ease of interpretation or easy mathematical handling. This is the case with Euclidean distance. Other indexes have nice or natural statistical properties, unfortunately sometimes associated with some handling difficulties. This is the case of measurements associated with intrinsic criteria, like those discussed in

García and Oller (2006). A natural intrinsic distance is the distance based on Fisher's information metric and proposed in Rao (1945). In this setting some concepts may have a more natural treatment; for example, the determination of the equivalence limits, possibly related to concepts like the curvature of the parametric Riemannian manifold.

A final consideration: equivalence problems generally admit either a frequentist or a Bayesian approach, but frequentist solutions are more common in the literature and much more often used in practice, despite the nice properties of many Bayesian solutions. This may be due in part to the weight of regulatory agencies in bioequivalence testing, the most significant application area. There may well be a regulatory bias towards the frequentist approach, but it is also likely to be based on criteria of clarity and ease of use for the potential users of the methods.

Acknowledgements

We thank the reviewers and the editor for constructive suggestions in the previous version of this paper.

References

Al-Mohizea A. M., Kadi, A. A., Al-Bekairi, A. M., Al-Balla, S. A., Al-Yamani, M. J, Al-Khamis, K. I, Niazy, E. M., El-Sayed, Y. M. (2007). Bioequivalence evaluation of 320 mg gemifloxacin tablets in healthy volunteers. *International Journal of Clinical Pharmacology and Therapeutics*, 45, 617-622.

Anderson, S. and Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics-Theory and Methods*, 12, 2663-2692.

Anderson, S. and Hauck, W. W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, 18, 259-273.

Barker, L., Luman, E. T., Mc Cauley, M. M. and Chu, S. Y. (2002). Assessing equivalence: an alternative to the use of difference tests for measuring disparities in vaccination coverage. *American Journal of Epidemiology*, 156, 1056-1061.

Barker, L., Rolka, H., Rolka, D. and Brown, C. (2001). Equivalence testing for binomial random variables: Which test to use? *The American Statistician*, 55, 279-287.

Barnett, V. (1999). *Comparative Statistical Inference*. New York: John Wiley.

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

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11, 283-319.

Blume, H. H. and Midha, K. K. (1993). Bio-International '92, Conference on Bioavailability, Bioequivalence and Pharmacokinetic Studies. *Pharmaceutical Research*, 10, 1806-1811.

Boddy, A. W., Snikeris, F. C., Kringle, R. O., Wei, G. C. G., Oppermann J. A. and Midha, K. K. (1995). An approach for widening the bioequivalence acceptance limits in the case of highly variable drugs. *Pharmaceutical Research*, 12, 1865-1868.

Brown, B. W. (1980). The crossover experiment for clinical trials. *Biometrics*, 36, 69-79.

Burns, D. R. and Elswick Jr., R. K. (2001). Equivalence testing with dental clinical trials. *Journal of Dental Research*, 80, 1513-1517.

CDER, Center for Drug Evaluation and Research (2001). *Statistical Approaches to Establishing Bioequivalence*. Rockville, MD: Food and Drug Administration. www.fda.gov/cder/guidance/3616fnl.htm#P353_2704.

CDER, Center for Drug Evaluation and Research (2003). *Guidance for Industry. Bioavailability and Bioequivalence Studies for Orally Administered Drug Products-General Considerations*. Rockville, MD: Food and Drug Administration. http://www.fda.gov/cder/guidance/index.htm.

Chow, S-C. and Liu, J-P. (2000). *Design and Analysis of Bioavailability and Bioequivalence studies*. New York: Marcel Dekker.

Chow, S-C. and Shao, J. (2002). A note on statistical methods for assessing therapeutic equivalence. *Controlled Clinical Trials*, 23, 515-520.

Committee for Proprietary Medicinal Products (CPMP) (2001). *Note for Guidance on the Investigation of Bioavailability and Bioequivalence*. London: EMEA.

D'Angelo, G., Potvin, D. and Turgeon, J. (2001). Carry-over effects in bioequivalence studies. *Journal of Biopharmaceutical Statistics*, 11, 35-43.

Dragalin, V., Fedorov, V., Patterson, S. and Jones, B. (2003). Kullback-Leibler divergence for evaluating bioequivalence. *Statistics in Medicine*, 22, 913-930.

Freeman, P. R. (1989). The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Statistics in Medicine*, 8, 1421-1432.

García, G. and Oller, J. M. (2006). What does intrinsic mean in statistical estimation? *SORT*, 30, 125-170.

Girón, F. J., Martínez, M. L., Moreno, E. and Torres, F. (2006). Objective testing procedures in linear models: calibration of *p*-values. *Scandinavian Journal of Statistics*, 33, 765-784.

Ghosh, P. and Gonen, M. (2008). Bayesian modeling of multivariate average bioequivalence. *Statistics in Medicine*, 27, 2402-2419.

Grieve, A. P. (1985). A Bayesian Analysis of the Two-Period Crossover Design for Clinical Trials. *Biometrics*, 41, 979-990.

Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics*, 21, 467-480.

Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55, 19-24.

Hsu, J. C., Hwang, J. T. G., Liu, H-K. and Ruberg, S. J. (1994). Confidence intervals associated with test for bioequivalence. *Biometrika*, 81,103-114.

Kleijnen, J. P. C. (1999). Validation of models: statistical techniques and data availability. *Proceedings of the 1999 Winter Simulation Conference, P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds.*, 647-654.

Laird, N. M., Skinner, J. and Kenward, M. (1992). An analysis of two-period crossover designs with carry-over effects. *Statistics in Medicine*, 11, 1967-1979.

Lehmann, E. L. (1959). *Testing Statistical Hypothesis*. New York: Wiley.

Longford, N. T. (2001). Synthetic estimators with moderating influence: the carry-over in cross-over trials revisited. *Statistics in Medicine*, 20, 3189-3203.

Manoli, T., Gretz, N., Grone, H-J., Kenzelmann, M., Eils, R. and Brors, B. (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, 22, 2500-2506.

Martín Andrés, A. and Herranz Tejedor, I. (2002). Equivalence testing for binomial random variables: which test to use? (Letter to the Editor). *The American Statistician*, 56, 253-254.

Mecklin, C. J. (2003). The use of equivalence testing in conjunction with standard hypothesis testing and effect sizes. *Journal of Modern Applied Statistical Methods*, 2, 329-340.

Metzler, C. M. (1974). Bioavailability: a problem of equivalence. *Biometrics*, 30, 309-317.

Midha K. K., Rawson M. J. and Hubbard J. W. (2005). The bioequivalence of highly variable drugs and drug products. *International Journal of Clinical Pharmacology and Therapeutics*, 43,485–498.

Moradi, M., Mousavi, P. and Abolmaesoumi, P. (2006). Pathological distinction of prostate cancer tumors based on DNA microarray data. http://cscbc2006.cs.queensu.ca/assets/documents/Papers/paper127.pdf.

Moreno, E. and Girón, F. J. (2006). On the frequentist and Bayesian approaches to hypothesis testing. *SORT*, 30, 3–28.

Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60, 223-241.

Munk, A. and Pflüger, R. (1999). $1 - \alpha$ equivariant confidence rules for convex alternatives are $\alpha/2$–level tests –with applications to the multivariate assessment of Bioequivalence. *Journal of the American Statistical Association*, 94, 1311-1319.

Perlman, M. D. and Wu, L. (1999). The emperor's new tests. *Statistical Science*, 14, 355-381.

Putt, M. and Chinchilli, V. M. (1999). A mixed effects model for the analysis of repeated measures cross-over studies. *Statistics in Medicine*, 19, 3037-3058.

Putt, M. (2005). Comment on 'Carry-over in cross-over trials in bioequivalence: theoretical concerns and empirical evidence'. Senn S, D'Angelo G, Potvin D. *Pharmaceutical Statistics*, 4, 215-216.

Putt, M. (2006). Power to detect clinically relevant carry-over in a series of cross-over studies. *Statistics in Medicine*, 25, 2567-2586.

Racine, A., Grieve, P., Fluhler, H. A. and Smith, F. M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry. *Applied Statistics*, 35, 93-150.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 81-91.

Rashid, M. M. (2003). Rank-based tests for non-inferiority and equivalence hypotheses in multi-centre clinical trials using mixed models. *Statistics in Medicine*, 22, 291-311.

Reynolds, M. R. and Deaton, M. L. (1982). Comparisons of some tests for validation of stochastic simulation models. *Communications in Statistics – Simulation and Computation*, 11, 769-799.

Robinson, A. P. and Froese, R. E. (2004). Model validation using equivalence tests. *Ecological Modelling*, 176, 349-358.

Rodda, B. E. and Davis, R. L. (1980). Determining the probability of an important difference in bioavailability. *Clinical Pharmacology and Therapeutics*, 28, 247-252.

Salicrú, M., Ocaña, J. and Sánchez, A. (2008). Comparison of lists of genes based on functional profiles. *Submitted*.

Sánchez, A., Salicrú, M. and Ocaña, J. (2007). Statistical methods for the analysis of high-throughput data based on functional profiles derived from the Gene Ontology. *Journal of Statistical Planning and Inference*, 137, 3975-3989.

Sánchez, A., Ocaña, J. and Salicrú, M. (2008). *goProfiles: an R package for the Statistical Analysis of Functional Profiles*. http://estbioinfo.stat.ub.es/pubs/goProfiles-Usersguide.pdf

Sargent, R. G. (2005). Verification and validation of simulation models. *Proceedings of the 2005 Winter Simulation Conference, M. E. Kuhl, N. M. Steiger, F. B. Armstrong and J. A. Joines, eds.*, 130-143.

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

Schall, R. and Luus, H. G. (1993). On population and individual bioequivalence. *Statistics in Medicine*, 12, 1109-1124.

Schuirmann D. J. (1987). A comparison of the Two One-sided Test procedure and the Power Approach for assessing the equivalence of Average Bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657-680.

Sellke, T., Bayarri, M. J. and Berger, O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, 55, 62-71.

Senn, S. (1988). Cross-over trials, carry-over effects and the art of self-delusion. *Statistics in Medicine*, 7, 1099-1101.

Senn, S. (1992). Is the 'simple carry-over' model useful? *Statistics in Medicine*, 11, 715-726.

Senn, S. (1996). The AB/BA Cross-over: How to perform the two-stage analysis if you can't be persuaded that you shouldn't. *Hansen, B and De Ridder, M. eds. Liber Amicorum Roel van Strik*, 93-100. Rotterdam: Erasmus University.

Senn, S. (2001). Statistical issues in bioequivalence. *Statistics in Medicine*, 20, 2785–2799.

Senn, S., D'Angelo, G. and Potvin, D. (2004). Carry-over in cross-over trials in bioequivalence: theoretical concerns and empirical evidence. *Pharmaceutical Statistics*, 3, 133-142.

Senn, S., D'Angelo, G. and Potvin, D. (2005). Rejoinder: Dr. Putt's analysis. *Pharmaceutical Statistics*, 4, 217-219.

Shah, V. P., Yacobi, A., Barr, W. H., Benet, L. Z., Breimer, D., Dobrinska, M. R., Endrenyi, L., Fairweather, W., Gillespie, W., Gonzalez, M. A., Hooper, J., Jackson, A., Lesko, L., Midha, K. K., Noonan, P. K., Patnaik R. and Williams R. L. (1996). Evaluation of Orally Administered Highly Variable Drugs and Drug Formulations. *Pharmaceutical Research*, 13, 1590-1594.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1, 203–209.

Stegner, A. L., Bostrom, A. G. and Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning*, 19, 193-198.

Tothfalusi, L. and Endrenyi, L. (2003). Limits for the scaled average bioequivalence of highly variable drugs and drug products. *Pharmaceutical Research*, 20, 382-389.

Van Steen, K., Raby, B. A., Molenberghs, G., Thijs, H., De Wit, M. and Peeters M. (2005). An equivalence test for comparing DNA sequences. *Pharmaceutical statistics*, 4, 203-214.

Wang, W. W., Hwang, J. T. G. and Dasgupta, A. (1999). Statistical tests for multivariate bioequivalence. *Biometrika*, 86, 395-402.

Warner, B. (2002). Equivalence testing. *MORS Workshop "Test & Evaluation, Modeling & Simulation and VV&A: Quantifying the Relationship Between Testing and Simulation", Kirtland AFB, Albuquerque, NM*. http://www.mors.org/meetings/test_eval/presentations/C_Warner.pdf.

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

Wellek, S. (2001). On a reasonable disaggregate criterion of population bioequivalence admitting of resampling-free testing procedures. *Statistics in Medicine*, 19, 2755-2767.

Wellek, S. (2003). *Testing Statistical Hypotheses of Equivalence*. Boca Raton: Chapman & Hall/CRC.

Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson Jr, H. F. and Hampton, G. M. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Research*, 61, 5974–5978.

Williamson, P. P. (2007). Bayesian equivalence testing for binomial random variables. *Journal of Statistical Computation and Simulation*, 77, 739 - 755

Zapater, P., Horga, J. F. (1999). Bioequivalencia y Genéricos. Los estudios de Bioequivalencia. I. Una aproximación a sus bases teóricas, diseño y realización. *Revista de Neurología*, 29:1235-1246.

SORT 32 (2) July-December 2008, 177-198

Sampling design variance estimation of small area estimators in the Spanish Labour Force Survey*

M. Herrador[a], D. Morales[b], M. D. Esteban[b], A. Sánchez[b],
L. Santamaría[b], Y. Marhuenda[b] and A. Pérez[b]

Abstract

The main goal of this paper is to investigate how to estimate sampling design variances of model-based and model-assisted small area estimators in a complex survey sampling setup. For this purpose the Spanish Labour Force Survey is considered. Sample and aggregated data are taken from the Canary Islands in the second trimester of 2003 in order to obtain some small area estimators of ILO unemployment totals. Several problems arising from the application of standard small area estimation procedures to the survey are described. It is shown that standard variance estimators based on explicit formulas are not applicable in the strict sense, since the assumptions under which they are derived do not hold. In addition two resampling techniques, bootstrap and jackknife, are considered. These methods treat all the considered estimators in the same manner and therefore they can be used as performance measures to compare them. From the analysis of the obtained results, some recommendations are given.

MSC: 62D05, 62J05.

Keywords: Labour Force Survey, small area estimation, linear models, mean squared error, bootstrap, jackknife, unemployment totals, calibrated weights.

1 Introduction

Small area estimation is an increasingly important part of survey sample inference with applications to social and economic statistics. Almost all the methodological developments up to date in this context has been carried out under the assumption that the assumed small area model is true, and that the appropriate measure of accuracy of

* Supported by the grant MTM2006-05693.

[a] Instituto Nacional de Estadística, Spain.

[b] Centro de Investigación Operativa, Universidad Miguel Hernández de Elche, Spain.

Received: October 2007
Accepted: July 2008

the small area estimator is its repeated sampling variability under random realizations of the population assuming the small area model holds. The fact that the assumed model only approximates reality, and that the measures that capture sampling variability relative to the actual population values are often of primary interest, is often ignored. This paper attempts to redress this imbalance by focusing on the repeated sampling properties of the most commonly used model-based methods of small area estimation.

The paper describes investigations on several issues arising from the application of standard small area estimation techniques, as they have been typically developed to be used under simple random sampling and they do not take into account problems derived from data coming up from surveys with complex sampling design, non response, reliability of population sizes, selection of auxiliary variables, consistency with the officially published data at a higher level of aggregation, estimation of mean squared error in a complex setup and many others. For this sake, some model-assisted and model-based estimators are adapted to the Spanish Labour Force Survey (SLFS) in order to estimate totals of unemployed people by sex and small areas in the Canary Islands. The paper has thus an applied-oriented character, attempting to diminish the gap between theory and practice.

The rest of the paper is organized as follows. Section 2 introduces some standard small area estimators and the corresponding explicit formulas to estimate their variances or mean squared errors. It also describes the auxiliary variables employed to estimate totals of unemployed people in the SLFS. Section 3 discusses two resampling approaches for estimating design-based variances. Section 4 describes technical details of the SLFS, with special emphasis on the sampling design, the separated ratio estimator of totals and the calibration of sampling weights. Section 5 proposes a two-stage bootstrap and a delete-one-cluster jackknife method to estimate sampling variances of small area estimators in the SLFS. These resampling methods produce performance measures to compare estimators of totals. Section 6 gives a discussion on the performance of the small area estimators and on the three methods to estimate their variances. The paper has two appendices. Appendix A presents estimated totals of unemployed people. Appendix B gives figures with estimated coefficients of variation and presents dispersion graphs to illustrate the behaviour of the small area estimators with respect to the basic estimator of the SLFS.

2 Estimators of small areas totals in complex surveys

Let Ω be a population with N units and let $s \subset \Omega$ be a sample of size n selected with a given sampling design. Let $\pi_i = P(i \in s)$ and $w_i = 1/\pi_i$ be the inclusion probability of unit $i \in \Omega$ and its sampling weight. Let y_i and $\mathbf{x}_i$ be the target variable and the vector of auxiliary variables defined for each $i \in \Omega$. Let $\mathbf{y}$ and $\mathbf{X}$ be the vector and the matrix containing the values of y_i and $\mathbf{x}_i$ for all units in the population. The three basic

inferential frameworks in survey sampling are the design-based, the model-based and the model-assisted approaches. In the design-based framework **y** and **X** are regarded as constants and the only source of randomness is the selection of the sample. In the model-based framework a model is assumed for **y** conditioned on **X**. In the model-assisted framework, both probability sampling design and model have a role (see Särndal, et al. 1992, pp. 227, 238-239). The model is used to propose an estimator with the restriction of being approximately unbiased in the sampling distribution.

We are interested in estimating the total Y_d of a target variable y in a domain d of size N_d. Let $s_d = \Omega_d \cap s$ be the subsample of units in domain d. In this section we introduce some standard small area estimators of Y_d. We also give explicit formulas to estimate the sampling variances of design-based and model-assisted estimators and to estimate the mean squared errors of model-based estimators. As the main goal of this paper is to investigate how to estimate the design-based variance of different types of small area estimators, we consider four of them: a design-based, a model-assisted and two model-based ones. At the end of this section we describe the auxiliary variables employed in the SLFS setup.

2.1 Direct estimator

The direct estimator is the design-based estimator (10.3.6) appearing in Särndal et al. (1992), p. 391, when N_d is known. Its expression is

$$\hat{Y}_d^{dir} = N_d \hat{\bar{Y}}_d^{dir}, \text{ where } \hat{\bar{Y}}_d^{dir} = \frac{1}{\hat{N}_d} \sum_{j \in s_d} w_j y_j \text{ and } \hat{N}_d = \sum_{j \in s_d} w_j.$$

An explicit-formula estimator of its sampling variance is

$$var(\hat{Y}_d^{dir}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{j \in s_d} w_j(w_j - 1)(y_j - \hat{\bar{Y}}_d^{dir})^2.$$

2.2 GREG estimator

GREG estimator is a model-assisted estimator. The one presented here is assisted by a linear model. Consider p explanatory variables measured at N population units; i.e. $\mathbf{x}_j = (x_{j,1}, \ldots, x_{j,p})$, $j = 1, ., N$. Let

$$\bar{\mathbf{X}}_d = \frac{1}{N_d} \sum_{j \in \Omega_d} \mathbf{x}_j \quad \text{and} \quad \hat{\bar{\mathbf{X}}}_d^{dir} = \frac{1}{\hat{N}_d} \sum_{j \in s_d} w_j \mathbf{x}_j$$

be the domain means of the auxiliary variables and their direct estimators. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{X}$ is an $n \times p$ matrix with rows $\mathbf{x}_j$, $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{W}^{-1})$ and $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$. The weighted least square estimator of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{y} = (\sum_{j \in s} w_j \mathbf{x}_j^t \mathbf{x}_j)^{-1}(\sum_{j \in s} w_j \mathbf{x}_j^t y_j).$$

Observe that the set of p explanatory variables can include artifical variables. Here the first variable is such that $x_{j,1} = 1$, $j = 1, ..., n$; i.e. we assume a linear model with intercept term. In this way, estimation of $\boldsymbol{\beta}$ does not depend on the type of selected small area in territories with hierarchical structure. In this paper the GREG estimator of a total is a slight modification of the model-assisted estimator (2.4.8) appearing in Särndal et al. (1992, p. 410). Its expression is

$$\hat{Y}_d^{greg} = N_d \hat{\bar{Y}}_d^{dir} + N_d(\bar{\mathbf{X}}_d - \hat{\bar{\mathbf{X}}}_d^{dir})\hat{\boldsymbol{\beta}}.$$

Observe that

$$\hat{Y}_d^{greg} = \sum_{j \in s} g_{dj} w_j y_j \quad \text{and} \quad N_d \bar{\mathbf{X}}_d = \sum_{j \in s} g_{dj} w_j \mathbf{x}_j$$

where

$$g_{dj} = \frac{N_d}{\hat{N}_d} I_{\Omega_d}(j) + N_d(\bar{\mathbf{X}}_d - \hat{\bar{\mathbf{X}}}_d^{dir})(\sum_{j \in s} w_j \mathbf{x}_j^t \mathbf{x}_j)^{-1}\mathbf{x}_j^t,$$

and I_{Ω_d} is the indicator function of subset Ω_d. An explicit-formula estimator of its sampling variance is

$$var(\hat{Y}_d^{greg}) = \sum_{j \in s_d} w_j(w_j - 1) g_{dj}^2 (y_j - \mathbf{x}_j \hat{\boldsymbol{\beta}})^2.$$

2.3 EBLUPA estimator

The EBLUPA estimator is a composite estimator based on the 2-level linear mixed model (model A)

$$y_{dj} = \mathbf{x}_{dj}\boldsymbol{\beta} + u_d + v_{dj}^{-1/2} e_{dj},$$

where $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ and $e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$ are independent. The model is fitted by calculating maximum likelihood estimators of the regression and variance component parameters with a Fisher-scoring algorithm (see e.g. Rao, 2003, ch. 5-6). The EBLUPA estimator of a total is $\hat{Y}_d^{eblupa} = N_d \hat{\bar{Y}}_d^{eblupa}$, where

$$\hat{\bar{Y}}_d^{eblupa} = \hat{\gamma}_d(\hat{\bar{Y}}_d^{dir} - \hat{\bar{\mathbf{X}}}_d^{dir}\hat{\boldsymbol{\beta}}) + \bar{\mathbf{X}}_d\hat{\boldsymbol{\beta}}, \quad \text{with } \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + (\hat{\sigma}_e^2/\nu_d)},\ \nu_d = \sum_{j \in s_d} \nu_{dj}.$$

The EBLUPA estimator is in fact a pseudo-eblup estimator studied in work package 4 of the EURAREA project (`http://www.statistics.gov.uk/eurarea/`) and related to the ones proposed by Prasad and Rao (1999) and You and Rao (2002). Mean squared error is estimated by using $g_1 - g_4$ explicit formulas given by Prasad and Rao (1990) and later extended by Das, Jiang and Rao (2001) to more general linear mixed models. Recent results are reviewed by Jiang and Lahiri (2006).

2.4 EBLUPB estimator

The EPLUPB estimator is a composite estimator based on the area-level model (model B)

$$\bar{Y} = \bar{\mathbf{X}}_d\boldsymbol{\beta} + u_d \quad \text{and} \quad \hat{\bar{Y}}_d^{direct} = \bar{Y}_d + \varepsilon_d,$$

where $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ and $\varepsilon_d \overset{iid}{\sim} N(0, \sigma_d^2)$ are independent. This model was introduced by Fay and Herriot (1979) to estimate average per capita income for small areas in USA. The model is fitted by the same method as model A. Under model B, EBLUP estimator of total is

$$\hat{\bar{Y}}_d^{eblupb} = \hat{\gamma}_d\hat{\bar{Y}}_d^{dir} + (1 - \hat{\gamma}_d)\bar{\mathbf{X}}_d\hat{\boldsymbol{\beta}}, \quad \text{with} \quad \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_d^2}.$$

Mean squared error is estimated by using $g_1 - g_3$ explicit formulas given by Prasad and Rao (1990).

2.5 Auxiliary variables to estimate totals of unemployed people in the SLFS

To obtain models with high predictive properties, the selection of adequate explanatory variables is very important. In the case of individual level models, auxiliary variables are needed at both individual and domain level. At the individual level auxiliary variables are obtained from the survey sample and, except for the cases of non-response, their values are available. However it is much more difficult to evaluate auxiliary variables at the domain level, because their values come from external sources which sometimes are not available, have not sufficiently good quality or may even present definition differences with their sample counterparts. Because of these reasons, the number of available auxiliary variables for individual-level models describing unemployment is in general very small. In the real data application of this paper the domains of interest are small areas (provisional geographical divisions for statistical purposes) crossed with

sex. There are $2 \times 27 = 54$ domains in the considered universe (Canary Islands). The following auxiliary variables have been used to estimate totals of unemployed people in the SLFS:

1. Auxiliary variables at aggregated and unit level are:
   - GSAC: groups of sex (1-2), age (1-3) and employment claimant (1-2) with 12 values. Three age groups have been considered: 16-24, 25-54 and ≥ 55.
   - CLUSTER: groups of province and population size of the municipality with 4 values.
2. An auxiliary variable aggregated at the domain level without sample counterpart has been used. This variable, GSAU, has 12 categories representing the groups of sex (1-2), age (1-3) and registered as unemployed in the administrative register of employment claimants (1-2).
3. To estimate totals of unemployed people we use the following auxiliary variables:
   - CLUSTER and GSAC for estimators GREG and EBLUPA.
   - CLUSTER and GSAU for estimators EBLUPB.

3 Design-based variance estimation

The most commonly used methods for design-based variance estimation with complex survey data are linearization and resampling methods. Krewski and Rao (1981) showed the asymptotic consistency of the variance estimates for nonlinear functions of design-unbiased mean estimators based on linearization or on some of the existing resampling methods applied to multistage designs in which the primary sampling units are selected with replacement. The linearization method requires theoretical calculation and subsequent programming of derivatives, which can make it cumbersome to implement. For this reason resampling methods are becoming each time more popular. In this section we review, without being exhaustive, some resampling methods that can be adapted to complex survey sampling designs.

3.1 Bootstrap with replacement

Efron (1979) proposed a bootstrap method that involves generation of independent resamples, each drawn from the original with replacement. For each such resample the statistic of interest is calculated and the obtained values form the basis of inference. The properties of the bootstrap method have been extensively studied for the i.i.d. case. In the framework of survey sampling Efron's original bootstrap requires modifications to handle issues like finiteness of population, without replacement sampling, complexity

of survey designs, weighting schemes and nonlinearity of population parameters and estimators. Under random sampling without replacement the finite population correction factor (f.p.c.), $f = n/N$, plays an important role. If f is not negligible the with bootstrap with replacement (BWR) method tends to overestimate the variance of linear estimators. To overcome this difficulty McCarthy and Snowden (1985) suggested to use a bootstrap sample size $n' = (1 - f)^{-1}(n - 1)$. Rao and Wu (1988) proposed a BWR method which rescales the bootstrap samples so as to recover the f.p.c. factor in the usual simple random sampling without replacement (SRSWOR) variance formula for the design unbiased estimators of the population mean. For interesting related papers dealing with the impact of BRW methods in survey sampling, see Rust and Rao (1996), Sitter (1992), Shao (2003) and Lahiri (2003).

All the mentioned references treat the problem of estimating parameters of the global population. However, in the small area estimation (SAE) setup the application of the BRW method has extra difficulties because the parameters of interest are from non-designed domains with small expected sample sizes. Further in the SAE framework is quite common to use nonlinear model-based estimators (like the EBLUP). For these estimators, there exist a variety of methods to estimate their model-based mean squared error, but not to estimate its design-based variance. For these reasons the bootstrap proposals mentioned above need adaptation to be applied in a SAE setup with complex survey data, so that the naive BWR method becomes a simpler and worthwhile approach to be considered for complex sampling designs like the one of the SLFS. In this paper, the naive BWR method involves the following basic steps:

1. Using a suitable probability sampling scheme, generate resamples from the original sample.
2. From each resample calculate the estimator $\hat{\theta}$. Denote them by $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.
3. Bootstrap estimator of variance is $var_B(\hat{\theta}) = (B - 1)^{-1} \sum_{b=1}^{B} (\hat{\theta}^*_b - \hat{\theta}^*)^2$ with $\hat{\theta}^* = B^{-1} \sum_{b=1}^{B} \hat{\theta}^*_b$.

3.2 Jackknife

Quenouille (1949) introduced the jackknife method to estimate the bias of an estimator by deleting one datum each time from the original data set and recalculating the estimator based on the rest of the data. The jackknife has become a more valuable tool since Tukey (1958) found that the jackknife can also be used to construct variance estimators. The first theorem concerning the jackknife variance estimator was given by Miller (1964). Since then jackknife theory has been widely developed (see e.g. Shao and Tu, 1995), although not much work has been done on its adaptation to complex survey designs. One exception is the paper by Rao and Tasui (2004) where jackknife variance estimators are introduced under stratified multistage sampling. Rao and Tasui (2004) consider a population stratified in L strata and from each stratum h, $m_h \geq 2$ clusters are

selected, independently across the strata. They further assume that subsampling within the sampled clusters is performed to ensure unbiased estimation of cluster totals, Y_{hi}, $i = 1, \ldots, m_h$, $h = 1, \ldots, L$. An unbiased estimator of the cluster total Y is given by

$$\hat{Y} = \sum_{s} w_{hik} y_{hik},$$

where w_{hik} is the inverse of the first order inclusion probability of unit k in cluster i and stratum h. To obtain the delete-one-cluster jackknife estimator of the variance of $\hat{Y}$ the jackknife weights, when the (g, j)-th cluster is deleted or equivalently in the jackknife sample $s^*_{(g,j)}$, are $w^*_{hik(g,j)} = w_{hik} b_{hi(g,j)}$ with $b_{hi(g,j)} = m_g/(m_g - 1)$ if $h = g$, $i \neq j$, and $b_{hi(g,j)} = 1$ if $h \neq g$. The variance of $\hat{Y}$ can be approximated by

$$var_J(\hat{Y}) = \sum_{g=1}^{L} \frac{m_g - 1}{m_g} \sum_{j=1}^{m_g} (\hat{Y}^*_{(g,j)} - \hat{Y})^2, \quad \text{where} \quad \hat{Y}^*_{(g,j)} = \sum_{s_{(g,j)}} w^*_{hik(g,j)} y_{hik}.$$

4 The Spanish Labour Force Survey

The SLFS is a good example of complex survey design where a lot of challenging statistical issues takes place. The research in this paper is motivated by them. This section summarizes the key points of its sampling design as well as some details on how the collected survey data is handled in practice. Additional information can be downloaded from web site of the Spanish Statistical Office (Instituto Nacional de Estadística-INE)

`http://www.ine.es/en/docutrab/epa05_disenc/epa05_disenc_en.pdf.`

SLFS is a quarterly survey following a stratified two-stage random sampling design with separate samples s_p drawn from each province p. The Primary Sampling Units (PSUs) are Census Sections (geographical areas with a maximum of 500 dwellings-approximately 3000 people) and they are grouped in strata according to the size of municipality. Within each stratum, PSUs are selected with probabilities proportional to size according to the number of dwellings. In the second stage sampling, the Secondary Sampling Units (SSUs) are dwellings and a random start systematic sampling is applied to draw a fixed number (18 in most cases) of SSUs from each selected PSU. All people aged 16 years old or more in the selected SSUs are interviewed. The probability that a dwelling v belonging to PSU a of stratum h be selected in s_p is

$$P(Dwe_{hav}) = P(PSU_{ha}) P(Dwe_{hav}) | PSU_{ha}) = m_h \frac{V_{ha}}{V_h} \frac{18}{V_{ha}} = \frac{18 m_h}{V_h},$$

where V_{ha} and V_h are the totals of dwellings in PSU a of stratum h and in stratum h respectively and m_h is the number of sections allocated in stratum h. Because all individuals in a selected dwelling are interviewed, the inclusion probabilities of individuals and dwellings coincide. Therefore, the inclusion probability of individual j in dwelling v and stratum h is

$$\pi_j = \frac{18m_h}{V_h} = \pi_h.$$

This means that all individuals within a given stratum have the same selection probability, i.e. this survey uses what is called a self-weighting design. Afterwards, at stratum level, probabilities π_j are modified to take non-response into account and their inversions produce sampling weights $w_j^{(1)}$ adjusted by non-response. Consequently the survey is still using a self-weighting design inside of each stratum. Up until year 2001 the INE used a ratio estimator, with Demographic Population Projections as auxiliary variable, to estimate the total Y_p of variables y in the province p, i.e.

$$\hat{Y}_p^{lfs,0} = \sum_{h \subset \Omega_p} \frac{N_h}{\hat{N}_h} \sum_{v \subset s_h} \sum_{j \in v} w_j^{(1)} y_j \quad \text{with} \quad \hat{N}_h = \sum_{v \subset s_h} \sum_{j \in v} w_j^{(1)} = w_j^{(1)} n_h,$$

where $\hat{N}_h$ is the projection of the population living in familiar dwellings in stratum h, with reference to the half of the quarter and n_h is the number of people living in the dwellings in the sample, in stratum h, at the time of the interview. Alternatively,

$$\hat{Y}_p^{lfs,0} = \sum_{h \subset \Omega_p} \sum_{j \in s_h} \frac{N_h w_j^{(1)}}{\hat{N}_h} y_j = \sum_{j \in s_p} w_j^{(2)} y_j,$$

with the sample dependent weights

$$w_j^{(2)} = w_j^{(2)}(s_p) = \frac{N_h w_j^{(1)}}{\hat{N}_h} = \frac{N_h}{n_h} \quad \text{if} \quad j \in s_h.$$

Since the first quarter of 2002, reweighting (or calibration) techniques are applied to estimators so as to adjust the survey estimates to some given information from external sources. The reweighting technique (see Deville and Särndal (1992)) requires the availability of K auxiliary variables appearing in the sample s_p and whose populations totals are known, i.e.

$$\sum_{j \in \Omega_p} x_{jk} = X_k, \quad k = 1, \ldots, K.$$

The target is to find a new estimator

$$\hat{Y}_p^{lfs} = \sum_{j \in s_p} w_j y_j$$

with new weights w_j satisfying the balance equations

$$\sum_{j \in s_p} w_j x_{jk} = X_k, \quad k = 1, \ldots, K,$$

and being as similar as possible to $w_j^{(2)}$. The problem aims to find values w_j minimizing

$$\sum_{j \in s_p} w_j^{(2)} G(w_j / w_j^{(2)}) \quad \text{restricted to} \quad \sum_{j \in s_p} w_j x_{jk} = X_k, \quad k = 1, \ldots, K,$$

where G is a function of distance. In the second trimester of 2003 the SLFS weights were calibrated so that their sum coincide with the population projections for individuals aged 16 years and over per groups of sex and age in autonomous communities, and per provinces. In order to obtain the practical solution for this problem, it was employed the CALMAR (CALage sur MARges) software, programmed in SAS code by the INSEE (Institut National de la Statistique et des Études Économiques) in France.

SLFS estimator of the total Y_p of variable y in province p is $\hat{Y}_p^{lfs}$. In this setup direct estimators of the total and the mean (cf. Section 2.1) of domain d are

$$\hat{Y}_d^{lfs} = \sum_{j \in s_p} w_{dj} y_{dj} \quad \text{and} \quad \hat{\bar{Y}}_d^{lfs} = \frac{\hat{Y}_d^{lfs}}{\hat{N}_d}, \quad \text{with} \quad \hat{N}_d = \sum_{j \in s_p} w_{dj}.$$

For provinces, it holds $\hat{Y}_p^{lfs} = \sum_{d \in \Omega_p} \hat{Y}_d^{lfs}$; i.e. there exists consistency between direct estimates at domain and SLFS estimate at province level.

INE publishes estimates of unemployment totals at province level. If in the near future these publications were extended to domain levels it should be necessary to force consistency between both types of data. This is to say that the sum of the estimated totals in all the domains within a province should coincide with the actual estimated total by SLFS in the province. In order to fulfil this consistency criterion, in this paper the following modification of all the considered small area estimators has been implemented.

Let $\hat{Y}_d^{lfs}$ be the SLFS estimator of total Y_p in province p. Assume that province p is partitioned in D_p domains; i.e. $\Omega_p = \cup_{d=1}^{D_p} \Omega_{pd}$ with $\Omega_{d_1} \cap \Omega_{d_2} = \emptyset$ if $d_1 \neq d_2$. Let $\hat{Y}_1, \ldots, \hat{Y}_{D_p}$ be some given estimators of totals $Y_1, \ldots, Y_{D_p}$. In general, the consistency property

$$\hat{Y}_p^{lfs} = \sum_{d=1}^{D_p} \hat{Y}_d$$

Table 4.1: *Consistency factors for the totals of unemployed men (left) and women (right) in the SLFS 2003/02 of Canary Islands.*

Province	dir	greg	eblupa	eblupb	dir	greg	eblupa	eblupb
1	0.948	0.944	0.952	0.949	0.937	0.875	0.865	0.932
2	1.005	0.963	0.947	0.981	0.995	0.868	0.879	0.998

is not satisfied. In such cases $\hat{Y}_1, \ldots, \hat{Y}_{D_p}$ can be transformed into consistent estimators by the following calculation

$$\hat{Y}_d^c = \lambda_{yp} \hat{Y}_p \,, \quad \text{where} \quad \lambda = \frac{\hat{Y}_p^{lfs}}{\sum_{d=1}^{D_p} \hat{Y}_d}$$

are *consistency factors*. For consistent estimators, it holds

$$\hat{Y}_p^{lfs} = \sum_{d=1}^{D_p} \hat{Y}_d^c.$$

Table 4.1 presents the consistency factors of direct, GREG, EBLUPA and EBLUPB estimators of totals of unemployed men (left) and women (right) in the SLFS 2003/02 of Canary Islands. One can observe that the deviations from the SLFS estimation at province level are at most of 15% for the four small area estimators.

5 Resampling methods for design-based variance estimation in the SLFS

In this section we describe a two-stage bootstrap method as well as a two-stage jackknife method to estimate variances of small area estimators of totals in the SLFS.

5.1 A naive two-stage bootstrap method

Let θ be a parameter to be estimated with $\hat{\theta}$. Bootstrap (see e.g. Efron and Tibshirani, 1998) is a resampling method which is often used to estimate variances $var(\hat{\theta})$. To implement the proposed two-stage bootstrap method, it is not necessary to construct artificial populations since the procedure generates bootstrap samples directly from the original SLFS sample as it is explained in next lines. Let s be an SLFS sample in a given province. Let $s = \cup_{h=1}^{H} s_h$, where $s_1, \ldots, s_H$ are subsamples by strata. Let $s_h = \cup_{a=1}^{m_h} s_{ha}$, where $s_{h1}, \ldots, s_{hm_h}$ are subsamples in the m_h selected PSUs from the stratum h. Finally, let $s_{ha} = \cup_{v=1}^{m_{ha}} s_{hav}$, where $s_{ha1}, \ldots, s_{ham_{ha}}$ are the subsamples in the m_{ha} visited dwellings in PSU a and stratum h. Selection of bootstrap samples in stratum h, $h = 1, \ldots, H$, is done in the following way:

1. Select a simple random sample with replacement of m_h PSUs from the set of m_h PSUs appearing in the original SLFS sample.
2. Within each selected PSU, draw a simple random sample with replacement of m_{ha} dwellings from the set of m_{ha} dwellings appearing in the given PSU of the original SLFS sample.
3. Select all the individuals aged 16 or more from the dwellings in the bootstrap sample.

Variance estimation is done as follows:

A. By using the procedure described above, use sample s to draw B bootstrap samples ($B = 500$ in this paper). For every bootstrap sample calculate $\hat{\theta}_b^*$, $b = 1, ., B$, in the same way as $\hat{\theta}$ was calculated. So, in each bootstrap sample, the weights $w_{j,b}^{*(2)} = N_h/n_{hb}^*$ (where n_{hb}^* is the number of individuals selected in bootstrap sample b and stratum h) are adjusted by a calibration procedure to obtain calibration weights w_j^* in the same way as in SLFS sample. These calibration weights w_j^* are used to calculate $\hat{\theta}_b^*$.

B. The observed distribution of $\hat{\theta}_1^*, \ldots, \hat{\theta}_B$ is expected to imitate the distribution of estimator $\hat{\theta}$ in the SLFS sampling design.

C. The variance of $\hat{\theta}$ is approximated by

$$var_B(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)^2, \quad \text{where} \quad \hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*.$$

D. A bootstrap estimator of the sampling error (*coefficient of variation*) in % of $\hat{\theta}$ is

$$cv_B(\hat{\theta}) = \frac{\sqrt{var_B(\hat{\theta})}}{\hat{\theta}}\,100.$$

An important step when estimating variances through the bootstrap method is to take into account the consistency property of estimators of totals at province level. The consistency property was not required in the bootstrap samples. To estimate variances of consistent estimators, estimated variances of non consistent estimators are multiplied by the square of the consistency factor λ (cf. Section 4). However, for the coefficient of variation this adjustment is not necessary. More concretely, if $\hat{\theta}^c = \lambda\hat{\theta}$ is the consistent version of a total estimator $\hat{\theta}$, where λ is the consistency factor calculated in the original SLFS sample, then bootstrap estimators of the variance and the coefficient of variation of $\hat{\theta}^c$ are

$$var_B(\hat{\theta}^c) = \lambda^2 var_B(\hat{\theta}) \quad \text{and} \quad cv_B(\hat{\theta}^c) = cv_B(\hat{\theta}).$$

5.2 A delete-one-cluster jackknife method

In order to apply the jackknife for variance estimation in SLFS samples, we use the delete-one-cluster jackknife method (see e.g. Rao and Tausi, 2004). To obtain the delete-one-cluster jackknife variance estimator of $\hat{\theta}$, we generate jackknife samples by deleting a PSU each time. So within each province, there are as many jackknife samples as PSUs are in the corresponding SLFS sample.

Consider the jackknife sample, $s^*_{(g,j)}$, obtained by excluding PSU j of stratum g. The jackknife weight of individual k, PSU i and stratum h in the sample $s^*_{(g,j)}$ is $w_{hik(g,j)} = w^{(2)}_{hik} b_{hi(g,j)}$, where $b_{hi(g,j)} = \frac{m_g}{m_g - 1}$ if $h = g$, $i \neq j$, $b_{hi(g,j)} = 1$ if $h = g$, and m_g is the number of PSUs in the stratum g. Note that the case $h = g$ and $i = j$ does not appear in the jackknife sample $s^*_{(g,j)}$. If H is the number of strata in the sample, the variance estimation is done as follows:

A. By using the procedure described above, use sample s to draw jackknife samples $s^*_{(g,j)}$, $g = 1, \ldots, H$, $j = 1, \ldots, m_g$. For every jackknife sample calculate $\hat{\theta}^*_{(g,j)}$ in the same way as $\hat{\theta}$ was calculated. So, in each jackknife sample, the weights $w_{hik(g,j)}$ are adjusted by a calibration procedure to obtain calibrated weights $w^*_{hik(g,j)}$ in the same way as it was done with the SLFS sample. Theses calibrated weights $w^*_{hik(g,j)}$ are used to calculate $\hat{\theta}^*_{(g,j)}$.

B. The observed distribution of $\{\hat{\theta}^*_{(g,j)} : g = 1, \ldots, H, j = 1, \ldots, m_g\}$ is expected to imitate the distribution of estimator $\hat{\theta}$ in the SLFS sampling design.

C. The variance of $\hat{\theta}$ can be approximated by

$$var_J(\hat{\theta}) = \sum_{g=1}^{H} \frac{m_g - 1}{m_g} \sum_{j=1}^{m_g} (\hat{\theta}^*_{(g,j)} - \hat{\theta})^2.$$

D. A jackknife estimator of the sampling error (*coefficient of variation*) in % of $\hat{\theta}$ is

$$cv_J(\hat{\theta}) = \frac{\sqrt{var_J(\hat{\theta})}}{\hat{\theta}} 100.$$

6 Discussion

6.1 On the small area estimators

In this section a specific analysis of the behaviour of direct, GREG, EBLUPA and EBLUPB estimators of unemployment totals (men and women), in the SLFS of Canary Islands in the second trimester of 2003, is given. Conclusions are mainly based on data

from Table A.1 and in figures presented in Appendix B. In Appendix B explicit-formula, bootstrap and jackknife estimates of the variances or mean squared errors (MSE) of the estimators of totals of unemployed men are plotted. In order to analyze the degree of bias of the estimators of totals, in Figure B.5 they are plotted against the basically unbiased-design SLFS estimator in dispersion graphs. Similar figures have been plotted for the case of women. However, for the sake of brevity, they are not presented here. In relation to the different estimators tested the main conclusions are:

1. The four considered estimators tend to give the same numerical results as LFS estimator when sample size increases. See Table A.1.
2. To estimate totals of unemployed people, the four considered estimators are acceptably unbiased with respect to LFS estimator (see Figure B.5). From the figures in Appendix B we conclude that EBLUPA estimator is in general the one with the lowest MSE.

6.2 On the estimation of variances or mean squared errors

In this section advantages and disadvantages of the three considered variance or mean squared error (MSE) estimation procedures (explicit formula, bootstrap and jackknife) are analyzed.

Explicit formulas to estimate the variance or MSE of estimators of totals are easy to implement and require the same sample and auxiliary information than the one needed for the given estimators of totals. These formulas can also be extended to more general types of parameters (e.g. nonlinear) via Taylor linearization.

In the case of design-based or model-assisted estimators the formulas of variances are derived with respect to the sampling distribution with some simplifications to avoid double inclusion probabilities. What is estimated is thus a simplified version of the variance. In addition, elevation factors are treated as if they were inverted inclusion probabilities. Explicit formulas to estimate variances of design-based or model-assisted small area estimators of totals may have the following sources of error:

- They estimate simplified formulas of the variance that do not take into account second order inclusion probabilities.
- They assume that calibrated sampling weights are inverses of inclusion probabilities, when they are in fact sample dependent and therefore random.

In the case of model-based estimators MSE formulas are derived with respect to the model distribution. However, survey sampling statisticians are mainly interested in MSE with respect to the sampling distribution. If the model fits the data well, both types of MSE are usually close enough. In our application to real data, model-based and

jackknife estimators of MSE produce quite close results because the models fit the data acceptably well. Another issue is whether or not to use the sampling weights under the model-based approach, and how to use them.

Explicit formulas to estimate MSEs of model-based small area estimators of totals may have the following sources of error:

- They estimate the MSE with respect to the model distribution when we are interested in the MSE with respect to the sampling distribution.
- They are derived for simple random sampling. Under complex sampling designs, the use of sampling weights is still an unsolved problem.

As a summary we can say that explicit estimators of variances or MSEs are easy to apply, but give unreliable estimates as they are based on assumptions that do not hold in practice. Their use should have an orientate character.

The proposed two-stage bootstrap method generates resamples from the original SLFS sample. The method does not require the generation of bootstrap populations. The idea is that small area estimators in the original sample and in the bootstrap samples have very similar distributions, so that variance of estimators in the original sample could be estimated via Monte Carlo method by using the bootstrap samples. In simple random sampling (nonparametric) the bootstrap method is easy to implement and produce consistent (in an asymptotic sense) variance estimates. However in two-stage sampling this is not at all straightforward and it is quite difficult to check asymptotic properties with respect to PSUs or SSUs.

The bootstrap method needs to generate resamples in the same way that the original sample was generated. Here it is necessary to reproduce all the steps followed with the SLFS sample: extraction of the sample, calibration of weights, consistency of estimators at province level, and so on. However the naïve two-stage bootstrap method produce resamples whose distributions are not close enough to the one of the original sample. The key problem is that resamples are obtained with replacement and the original sample was obtained without replacement. Further research is thus needed to adapt BWR methods to the SLFS. By observing the obtained numerical results we conclude that this method over-estimates the variances of the small area estimators. A positive aspect of the bootstrap method is that variance estimates have a small loss of quality in domains with low sample sizes.

To estimate variances of small area estimators of totals, the naïve two-stage bootstrap method may have the following sources of error:

- Distributions of small area estimators in the original sample and in the resamples are not close enough.
- There exists a tendency to over-estimate variances.
- It is an excessively complex method, which needs a lot of delicate work.

The delete-one-cluster jackknife method generates resamples taking one PSU at a time out of the original sample and by recalibrating the sampling weights. It is a simple and easy method to implement. Main problem of jackknife method is that it works erratically in domains containing very few PSUs in the sample. For those domains this method is unreliable and should not be used.

If we compare the numerical results obtained with the three methods to estimate variances or MSEs, we obtain the following conclusions:

- In domains with large sample sizes, the three methods produce basically the same results.
- The naïve bootstrap method gives higher estimates of the variances than the explicit-formula or jackknife methods, so it seems that our implementation is positively biased.
- Assumptions required to deriving explicit formulas to estimate variances or MSEs do not hold in practice, so their use should have an orientative character.
- The delete-one-cluster jackknife method avoids the theoretical problems of the explicit-formula methods and the difficulty of implementation of the bootstrap method. It works quite well in all the domains except in those with very few sampled PSUs.

Acknowledgements

The authors would like to thank the INE household sampling design unit for their support and helpful comments.

References

Das, K., Jiang, J. and Rao, J. N. K. (2001). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818-840.

Deville J. C. and Särndal C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Society*, 87, 376-382.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Efron, B. and Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74, 269-277.

Jiang J. and Lahiri P. (2006). Mixed model prediction and small area estimation. *Test*, 15, 1-96.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, 18, 199-210.

McCarthy, P. J. and Snowden, C. B.] (1985). The bootstrap and finite population sampling. In Vital and Health Statistics 2-95. Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington.

Miller, R. G. (1964). A trust worthy jackknife. *Annals of Mathematical Statistics*, 35, 1594-1605.

Rao, J. N. K. and Wu, C.-F. J.] (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley.

Rao, J. N. K. and Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistic. Theory and methods*, 33, 2087-2095.

Rust, K. F. and Rao, J. N. K. (1996) Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag.

Shao, J. and Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer-Verlag.

Shao, J. (2003). Impact of bootstrap on sample surveys. *Statistical Science*, 18, 191-198.

Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Prasad, N. G. N. and Rao, J. N. K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, 25, 67-72.

Quenouille, M. (1949). Approximation tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11, 18-84.

Tukey, J. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29, 614.

You, Y. and Rao J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small-area estimation using survey weights. *Survey Methodology*, 30, 431-439.

Appendix A: Estimated totals of unemployed people in Canary Islands

Table A.1: *Estimated totals of unemployed men (left) and women (right) in Canary Islands with the SLFS 2003/02.*

Area	lfs	dir	greg	ebluba	eblupb	lfs	dir	greg	ebluba	eblupb
1	550	546	1172	1411	1865	3168	3022	3107	1993	2823
2	141	105	206	270	72	187	128	255	297	0
3	526	435	473	236	196	119	85	110	177	141
4	0	0	94	131	100	229	240	180	133	207
5	1131	1183	1047	765	967	467	490	866	892	638
6	458	379	347	432	260	635	464	511	499	527
7	13544	14081	13332	13275	14264	15637	16382	14044	13912	16118
8	319	401	274	724	695	289	350	470	829	107
9	1239	1161	818	1217	1185	1263	968	1260	1630	920
10	369	319	467	395	275	328	276	415	378	345
11	2451	2049	2219	1620	1162	959	760	855	1213	1237
12	2295	2364	2575	2546	1980	1889	2005	3097	3218	2107
13	343	277	527	987	506	787	798	1053	1311	863
14	2548	2006	1638	1656	1833	1791	1414	1381	1589	1515
15	9261	9928	9489	9505	10389	11802	12120	11071	11140	11422
16	507	420	514	564	681	681	614	592	665	746
17	1848	1241	1147	1079	1328	2530	1650	1409	1342	1302
18	496	985	1760	2419	1289	1253	2171	2960	2933	2321
19	966	1809	1650	1158	1022	426	717	818	1194	998
20	5502	5339	4303	3458	3848	5054	4890	4576	3788	4575
21			162	155	184			166	164	334
22	210	251	223	295	226	472	569	321	335	163
23	837	670	981	1095	911	1528	1388	1284	1190	1350
24			311	300	327			310	308	187
25	194	108	103	191	206	203	108	138	176	310
26	1599	1276	1344	1183	1244	446	353	957	1116	1056
27	0	0	159	263	316	545	726	483	269	377

Appendix B: Figures

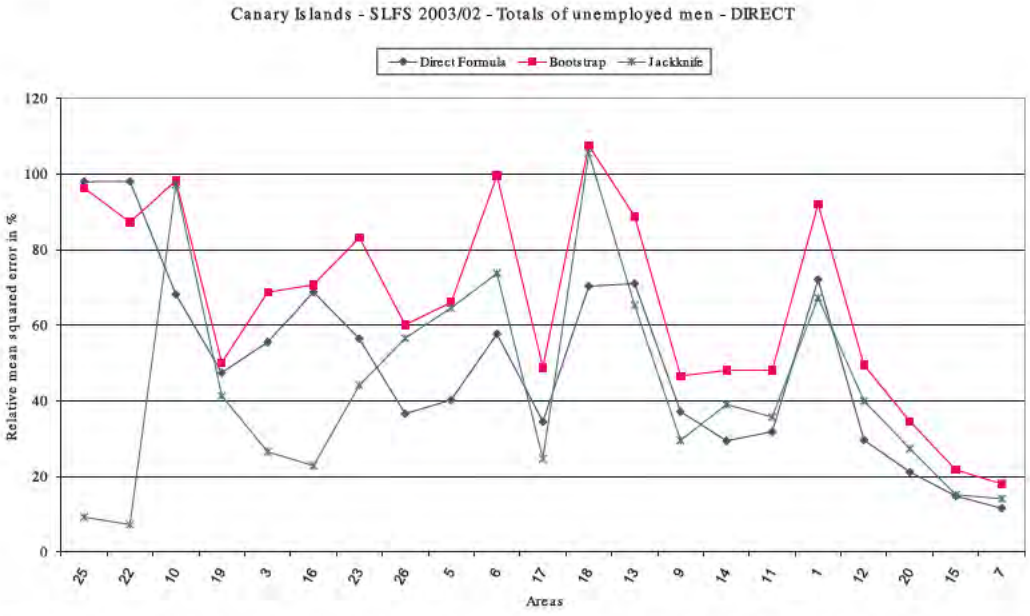


Figure B.1: Direct-formula, bootstrap and jackknife estimates of coefficients of variation in % of direct estimates of totals of unemployed men.

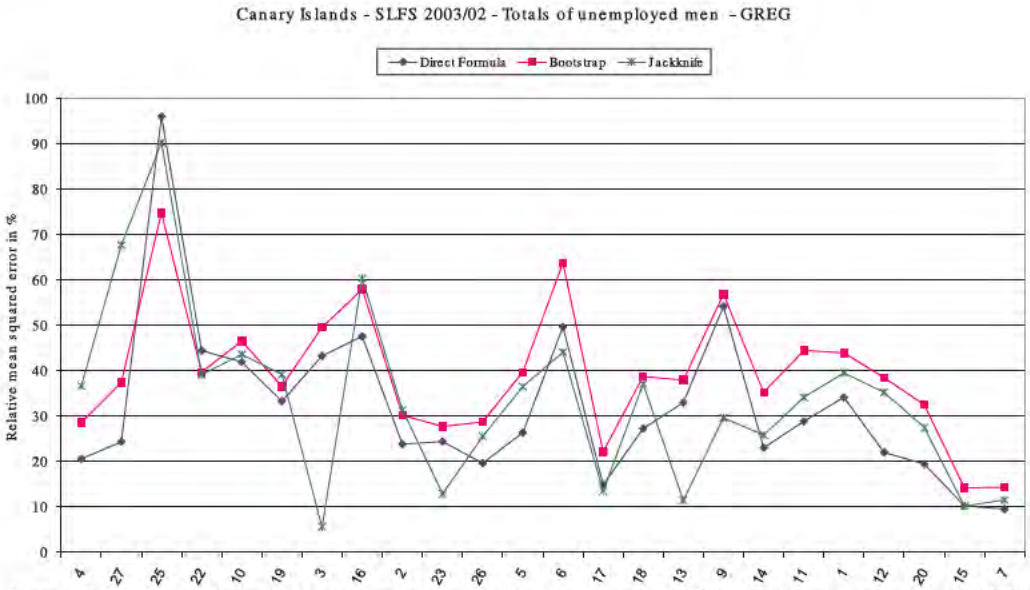


Figure B.2: Direct-formula, bootstrap and jackknife estimates of coefficients of variation in % of GREG estimates of totals of unemployed men.

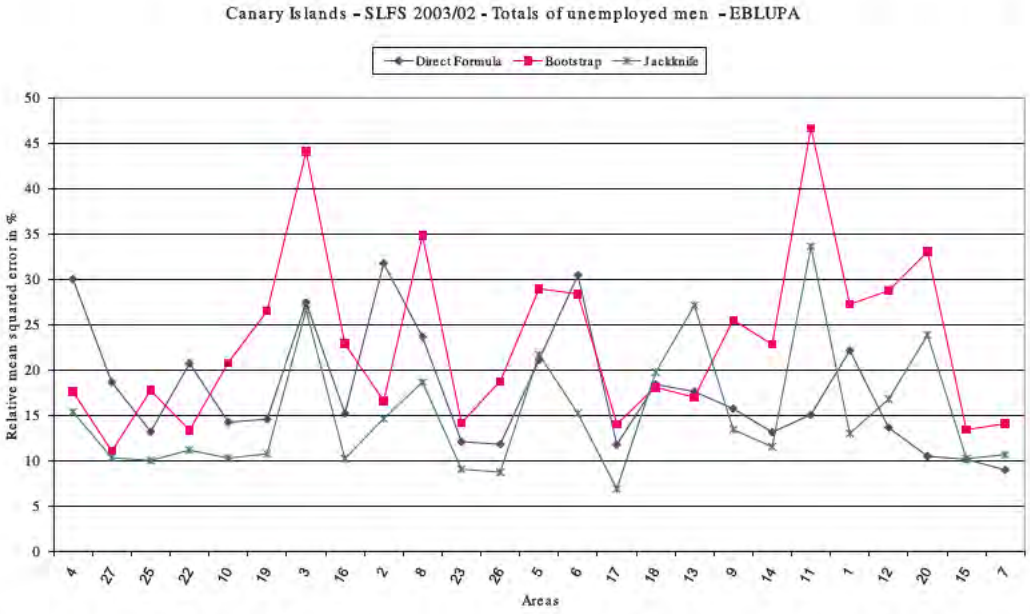


Figure B.3: Direct-formula, bootstrap and jackknife estimates of coefficients of variation in % of EBLUPA estimates of totals of unemployed men.

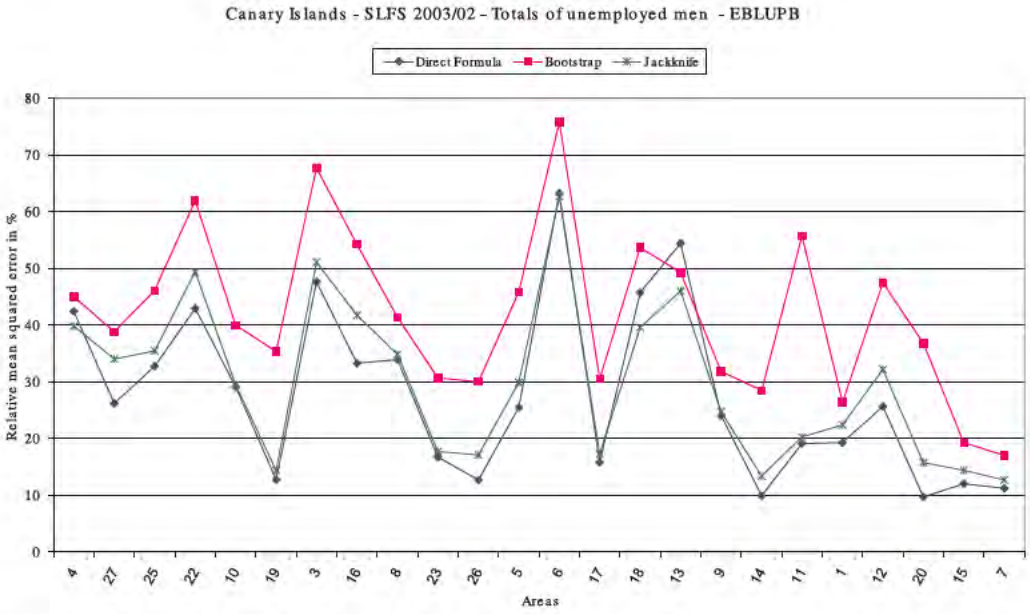


Figure B.4: Direct-formula, bootstrap and jackknife estimates of coefficients of variation in % of EBLUPB estimates of totals of unemployed men.

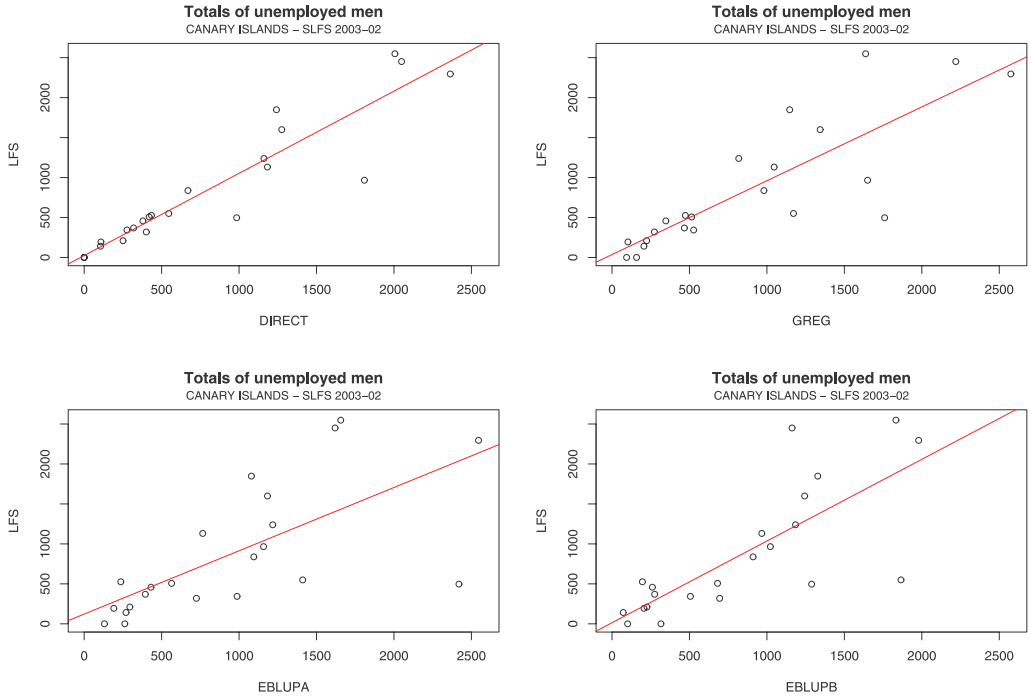


Figure B.5: *Dispersion graphs of LFS versus direct, GREG, EBLUPA and EBLUPB estimates of totals of unemployed men.*

Book review

LA PRÁCTICA DEL ANÁLISIS DE CORRESPONDENCIAS

Michael Greenacre

Fundación BBVA, Rubes Editorial, 2008

ISBN: 978-84-96515-71-0

This book is essentially the translation to Spanish of the recent second edition of the book *Correspondence Analysis in Practice* (2007, Chapman and Hall) that was first published in 1993. The book offers an extremely well structured introduction to this technique, giving special attention to practical issues and didactical aspects. In fact, the author has dedicated more than 30 years to the research, practice and divulgation of correspondence analysis and the book clearly benefits from this, providing real examples in a number of fields, such as medical sciences, linguistics, sociology or biology.

The structure of the book is quite unique in some senses. It contains 25 chapters of eight or nine pages each. This makes the book quite accessible and didactic. A final summary with a list of key concepts is provided at the end of each chapter, which helps to understand the topics presented. The theoretical concepts are always illustrated through motivating examples and a large number of tables and graphs, with very informative captions, are provided.

First chapters (1 to 5) are dedicated to introductory concepts such as plots and distances for categorical data. The basis of correspondence analysis, with its properties and applications, is presented in chapters 6 to 15. Chapters 16 to 22 provide an introduction to some extensions and variants of the correspondence analysis, such as multiple correspondence analysis. Last chapters (23 to 25) offer some technical aspects that can be of interest, especially those regarding inference. Furthermore, a mathematical and complete description of the method is provided in annex A, but which is not at all compulsory, since a general understanding of the technique is easily obtainable from previous chapters. Finally, appendix B is generously dedicated to the computation of correspondence analysis using the statistical package R, through the use of the libraries **ca** for the method itself and **rgl** for visualizing plots in three dimensions. Here, examples used along the chapters are visited again to illustrate the use of some functions implemented by the author and to obtain some of the figures and results in

the book. Indications are also given for the implementation of correspondence analysis using Excel.

First principles of correspondence analysis theory were developed at the beginning of the XXth century. However, the works by Jean Paul Bénzecri, Brigitte Escoffier and their French colleagues, during the 1960s, can be considered as the foundation of the present theory of correspondence analysis. Brigitte Escofier-Cordier thesis (Analyse des correspondances, 1963), directed by J.P. Bénzecri and the second volume of the book *Analyse des Données* (1973, Dunod, Paris, v. II *L'analyse des correspondances*) offer a complete vision of the pioneers works of the so called "French school" of data analysis. An active community of French researchers participated in the divulgation of this technique, such as Ludovic Lebart and Alain Morineau, by developing the software SPAD. This divulgation was mostly made in France but also at an international level. Furthermore, Michael Greenacre, who initiated his doctoral thesis with Benzécri during the 1970s, has also been an important contributor for the divulgation of correspondence analysis in the English-speaking world. Finally, correspondence analysis was introduced in Spain in the early 1980s and fundamental books in Spanish, such as *Tratamiento estadístico de datos* (Lebart, Morineau y Fénelon, Marcombo, 1984) and *Análisis factorials simples y multiples. Objetivos, métodos e interpretación* (Escoffier y Pagès, Servicio Editorial de la Universidad del País Vasco, 1990), allowed for a large diffusion of the technique and also for placing it in the context of the multivariate methods.

My first contact with correspondence analysis took place at the end of the 1990s when I was a student of statistics at the Polytechnic University of Catalonia. At that time, Monica Bécue introduced me to these methods, more specially their application to textual data. The book *Análisis estadístico de textos* (Lebart, Salem y Bécue, Ed Milenio, 2000) offers a vision of what is called textual statistics. M Bécue encouraged me to undertake a research stay at Paris with Alain Morineau, at the nowadays disappeared CISIA, which was a little private company, but with an important research activity and vocation. During my stay, I could go into correspondence analysis in more depth and get more conscience of the endless list of applications where this method can be applied. From my posterior own experience as statistical consultant at the Statistical Service of the Autonomous University of Barcelona, I can state that correspondence analysis can be applied in a variety of real contexts, with satisfactory, and quite often surprising, results. Moreover, it is a technique that provides a rapid interpretation and comprehension of the information, sometimes not so evident, in the data. Finally, for the last four years I have been teaching multivariate statistical analysis, including correspondence analysis, at the Universitat Autònoma de Barcelona and, this experience has shown me how motivating this technique can be for intrepid students, who can discover the pleasure of obtaining exciting results from the analysis of data.

In conclusion, I strongly recommend this book to everyone interested in the analysis of categorical data. It can be particularly useful for university students or teachers, researchers and professionals for any field involved with the analysis of categorical data.

Joan Valls Marsal
joan.valls@iconcologia.net
Catalan Cancer Registry, Catalan Institute of Oncology, Barcelona
Departament de Matemàtiques, Universitat Autònoma de Barcelona

Information for authors and subscribers

Information for authors and subscribers

Submitting articles to SORT

Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.es) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a LaTeX 2_ε.

In any case, upon request the journal secretary will provide authors with LaTeX 2_ε templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

Publishing rights and authors' opinions

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following examples:

Citations
Mahalanobis (1936), Rao (1982b)

Journal articles
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9, 73-84.

Books
Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium*. New York: John Wiley and Sons.

Parts of books
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Web files or "pages"
Nielsen, S. F. (2001). *Proper and improper multiple imputation*
http://www.stat.ku.dk/~feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel.: +34-93 412 09 24 or +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.es

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name _____

_____

Organisation _____

_____

Street Address _____

_____

Zip/Postal code _____ City _____

State/Country _____ Tel. _____

Fax _____ NIF/VAT Registration Number _____

E-mail _____

Date _____

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)*** for the year 2008 (volume 32)

Annual subscription rates:

– Spain: €22 (4 % VAT included)
– Other countries: €25 (4 % VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4 % VAT included)
– Other countries: €17/issue (4 % VAT included)

Method of payment:

☐ Bank transfer to account number 2013-0100-53-0200698577

☐ Automatic bank withdrawal from the following account number

☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Check made payable to the Institut d´Estadística de Catalunya

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Bank copy

Authorisation for automatic bank withdrawal in payment for

SORT *(Statistics and Operations Research Transactions)*

The undersigned _____

authorises Bank/Financial institution _____

located at (Street Address) _____

Zip/postal code _____ City _____

Country _____

to draft the subscription to **SORT *(Statistics and Operations Research Transactions)*** from my account

number ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Date _____

Signature

SORT *(Statistics and Operations Research Transactions)*
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Quatre modalitats de subscripció al DOGC

(Diari Oficial de la Generalitat de Catalunya)

Generalitat de Catalunya

Imprès, edició diària

Base de dades, actualització diària

DVD, edició semestral

A la carta, servei diari personalitzat

A més, per als subscriptors de l'edició impresa i del DVD, tramesa gratuïta d'un CD-ROM trimestral que conté les pàgines en format PDF (DOGC en imatges)

L'Administració més a prop

EADOP • Informació i subscripcions • Rocafort, 120 - Calàbria, 147 • 08015 Barcelona
Tel. 93.292.54.17 • Fax 93.292.54.18 • subsdogc@gencat.net • www.gencat.net/eadop