ISSN: 1696-2281
SORT 30 (1) January-June 1-122 (2006)

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

SORT

Volume 30 (1), January-June 2006

Formerly Qüestiió

Contents

SORT

Volume 30 (1), January-June 2006

Formerly Qüestiió

Contents

SORT 30 (1) January-June 2006, 3-28

On the frequentist and Bayesian approaches to hypothesis testing

Elías Moreno and F. Javier Girón

University of Granada and University of Málaga

Abstract

Hypothesis testing is a model selection problem for which the solution proposed by the two main statistical streams of thought, frequentists and Bayesians, substantially differ. One may think that this fact might be due to the prior chosen in the Bayesian analysis and that a convenient prior selection may reconcile both approaches. However, the Bayesian robustness viewpoint has shown that, in general, this is not so and hence a profound disagreement between both approaches exists.
In this paper we briefly revise the basic aspects of hypothesis testing for both the frequentist and Bayesian procedures and discuss the variable selection problem in normal linear regression for which the discrepancies are more apparent. Illustrations on simulated and real data are given.

MSC: 62A01; 62F03; 62F15

Keywords: Bayes factor, consistency, intrinsic priors, loss function, model posterior probability, *p*-values.

1 Introduction

In parametric statistical inference estimating parameters and hypothesis testing are two basic but different problems, although some Bayesian estimation devices have gained some popularity as hypothesis testing tools. For instance, high posterior probability regions have been used as acceptance regions for the null hypothesis (see, for instance, the analysis of variance solution proposed by Lindley 1970, Box and Tiao 1992). This is misleading as far as the hypotheses to be tested do not play any role in the construction of such acceptance regions.

Address for correspondence: Elías Moreno and F. Javier Girón. Department of Statistics and O.R. University of Granada and University of Málaga.
Received: November 2005

To distinguish between estimation and testing was strongly recommended by Jeffreys (1961, pp. 245-249), mainly because the methods commonly used for estimation are typically not suitable for hypothesis testing. Thus, we think it is timely to devote a separate paper to discuss the setting and tools devised for hypothesis testing from the frequentist and Bayesian perspectives. We want to thank the Editor of SORT for inviting us to contribute on this topic.

In this paper we deal with frequentist and Bayesian parametric hypothesis testing procedures. A third approach, based solely on the likelihood function, which we do not discuss here, is to be found in Royall (1997) and Pawitan (2001). As Pawitan (2001, p. 15) states: *The distinguishing view is that inference is possible directly from the likelihood function; this is neither Bayesian nor frequentist, and in fact both schools would reject such a view as they allow only probability-based inference*.

From the frequentist viewpoint two closely related methods have been developed. One is the Neyman-Pearson theory of significance tests and the other one is based on Fisher's notion of *p*-values. Here, we shall give arguments that make the *p*-values to be preferable to significance tests.

On the Bayesian side, robustness with respect to the prior showed that there is a strong discrepancy between the frequentist and Bayesian solutions to parametric hypothesis testing (Berger 1985, 1994, Berger and Delampady 1987, Berger and Sellke 1987, Berger and Mortera 1999, Casella and Berger 1997, Moreno and Cano 1998, Moreno 2005, among others). This means that the discrepancy is not due to the prior chosen for the Bayesian analysis but it is of a more fundamental nature which is inherent to the procedures. In particular, there is a marked difference on the way frequentist and Bayesian methods account for the sample size, and the dimensions of the null and the alternative parametric spaces.

Since subjective prior distributions for the parameters of the models involved in hypothesis testing are not generally available, and their use is perceived as the weak point in the Bayesian implementation, objective prior distributions will be employed in this paper. By objective priors we mean priors that only depend on the sampling model and theoretical training samples. This priors are called intrinsic priors (Berger and Pericchi 1996, Moreno 1997, Moreno *et al.* 1998) and their merits can be judged for each specific application. We remark that they have been proved to behave extremely well in a wide variety of problems (Casella and Moreno 2004, 2005, Girón and Moreno 2004, 2005, Moreno *et al.* 1999, 2000, 2003, 2005, Moreno and Liseo 2003).

The rest of the paper is organized as follows. The second section is devoted to briefly describing significance tests and *p*-values. Section 3 reviews the Bayesian testing machinery and justifies the need for objective priors. The challenging problem of testing whether the means of two normal distributions with unknown variances are equal is considered in Section 4. A comparison of the frequentist and Bayesian testing procedures for the normal linear regression model, including a discussion on some fundamental issues of the variable selection problem, is given in Section 5, and some conclusions and recommendations are given in Section 6.

2 Significance tests and p-values

Let X denote an observable random variable and $\mathbf{x} = (x_1, \ldots, x_n)$ an available sample from either the model $P_0(x)$ or $P_1(x)$ with probability densities $f_0(x)$ or $f_1(x)$, respectively. Suppose that we want to choose between either the null hypothesis $H_0 : f_0(x)$ or the alternative $H_1 : f_1(x)$. This is the simplest hypothesis testing formulation.

For this problem the well-known Neyman-Pearson theory of significance tests proposes a subset of the sample space $\mathcal{R}^n$, the so-called critical or rejection region,

$$W_\alpha = \left\{ \mathbf{y} : \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} \geq c_\alpha \right\}, \tag{1}$$

as the region containing evidence against the null hypothesis. The threshold c_α is determined so that the probability of the critical region under the null is α, that is

$$P_0(W_\alpha) = \int_{W_\alpha} f_0(\mathbf{y})\, d\mathbf{y} = \alpha.$$

Given the data $\mathbf{x}$, the null H_0 is rejected at the significance level α if $\mathbf{x} \in W_\alpha$, and accepted otherwise. The value α is usually taken to be small so that we have a small probability of rejecting the null when it is true. Typically α is chosen to be 0.05 or 0.01.

Fisher criticized the notion of significance of the Neyman-Pearson theory and proposed replacing c_α in (1) with the observed likelihood ratio $\lambda_n(\mathbf{x}) = f_1(\mathbf{x})/f_0(\mathbf{x})$, so that the above critical region becomes

$$W_n(\mathbf{x}) = \left\{ \mathbf{y} : \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} \geq \lambda_n(\mathbf{x}) \right\},$$

that now depends on the observed data $\mathbf{x}$. The probability of this region under the null, $p = P_0(W_n(\mathbf{x}))$ say, is called the p-value of the data $\mathbf{x}$, and it is interpreted in such a way that a small enough p-value implies evidence against H_0. The smaller the p-value the stronger the evidence against the null.

Significance tests and p-values substantially differ. While the p-value depends on the data thus giving a measure of how strongly the observed data reject the null, the significance test does not provide such a measure. This simple fact implies that, for the dicotomous testing problem, the p-value is consistent under the null and the alternative, while the significance test is consistent under the alternative but it is not under the null.

Indeed, when sampling from the null and the sample size grows to infinity the likelihood ratio statistic $\lambda_n(\mathbf{x})$ tends to zero under P_0, and the corresponding sequence of p-values tends to 1. This implies that, asymptotically, there is no evidence at all to reject the null. Alternatively, when the sample comes from the alternative hypothesis, the statistic $\lambda_n(\mathbf{x})$ tends to ∞ under P_0, and hence the corresponding sequence of p-

values tends to 0 showing an increasing evidence to reject the null. On the other hand, when sampling from the null, the significance test will reject H_0 with probability α even when the sample size increases to infinity. This means that the significance test is not a consistent procedure under the null, although it is consistent when sampling from the alternative (Casella and Berger 1990, pp. 381-382, Wilks 1962, pp. 411).

Under some regularity conditions on the likelihood, this theory can be extended to the more general situation of considering a family of probability densities $\{f(x|\theta), \theta \in \Theta\}$, where Θ might be a multidimensional parametric space, and the null $H_0 : \theta \in \Theta_0$ and the alternative $H_1 : \theta \in \Theta_1$ may contain more than one density. To preserve the good properties of the mentioned simplest hypothesis testing procedures, the hypotheses must be nested, that is $\Theta_0 \subset \Theta_1$, and also we must have $k_0 = \dim(\Theta_0)$ strictly smaller than $k_1 = \dim(\Theta_1)$. The critical region now is

$$W_n(\mathbf{x}) = \left\{ \mathbf{y} : \frac{f(\mathbf{y}|\hat{\theta}_1(\mathbf{y}))}{f(\mathbf{y}|\hat{\theta}_0(\mathbf{y}))} \geq \hat{\lambda}_n(\mathbf{x}) \right\},$$

where $\hat{\theta}_1(\mathbf{y})$, $\hat{\theta}_0(\mathbf{y})$ are the MLE's of θ in the spaces Θ_1 and Θ_0, respectively, and $\hat{\lambda}_n(\mathbf{x}) = f(\mathbf{x}|\hat{\theta}_1(\mathbf{x}))/f(\mathbf{x}|\hat{\theta}_0(\mathbf{x}))$.

Two important difficulties now arise with the *p*-values. First, it is not clear how the probability of $W_n(\mathbf{x})$ under the null can be computed when either the null is composite or the distribution induced by $\hat{\lambda}_n(\mathbf{y})$ depends on some (nuisance) parameter, as in the Behrens-Fisher problem which we briefly deal with in Section 4.

Second, while a small *p*-value might contain evidence against the null for a dataset, the same *p*-value for a different sample size and/or for a null parameter space of different dimension might not contain the same evidence against the null. In fact, when we consider multiple tests, i.e. when we want to test that some subsets of regression coefficients are zero, as in variable selection, the frequentist literature (see, for instance Miller 2002) recognizes that the meaning of the *p*-value should depend on the dimensions of the null and the alternative parameter spaces and, consequently, provides a variety of methods to correct the *p*-value to account for this. It has also been recognized that the bigger the sample size the smaller the evidence of a given *p*-value against the null so that it is also necessary to correct the *p*-value to account for the sample size. These considerations have prompted the need to introduce frequentist criteria based on statistics that, in some way, adjust for all the varying parameters in the model such as the sample size and the dimensions of the null and the alternative hypothesis, such as the adjusted R^2, Mallow's C_p or the AIC criteria.

In summary, the meaning of the *p*-value is unfortunately unclear. Its interpretation should depend on the dimension of the null space, the dimension of the alternative space, and the sample size in an unknown, and probably complex and non-trivial, way; as a consequence, the calibration of a *p*-value is deemed to be a very difficult task, although some attempts for calibrating the *p*-values can be found in Sellke *et al.* (2001) and Girón *et al.* (2004).

Furthermore, although a p-value is derived as a probabilistic statement and, consequently, lies between zero and one, it cannot be interpreted as the posterior probability that the null is true – this is an instance of the well known *transposed conditional or prosecutor's fallacy*. However, practitioners of the p-values have very often this *wrong* probability interpretation in mind, maybe because this provides them with some sort of a (wrong) measurement devise for calibration.

3 Bayesian hypothesis testing

From a Bayesian viewpoint the testing problem is treated as follows. Consider the simplest testing problem where we have to choose either the model $M_0 : f_0(x)$ or $M_1 : f_1(x)$ based on the observations $\mathbf{x} = (x_1, \ldots, x_n)$. Let d_i denote the decision of choosing M_i and let P be the prior probability defined on the model space $\{M_0 , M_1\}$. Assume that a loss $L(d_i, M_j) = c_{ij}$, $i, j = 0, 1$, is incurred when we make the decision d_i and the true model is M_j (for other loss functions in model selection, see San Martini and Spezzaferri 1984, and Bernardo and Smith 1994).

Assuming that the loss for a correct decision is zero, that is $c_{ii} = 0$, and $c_{ij} > 0$ otherwise, d_1 is the optimal decision when the posterior risks satisfy $R(d_0|\mathbf{x}) > R(d_1|\mathbf{x})$. This implies the following inequality

$$P(M_0|\mathbf{x}) < \frac{c_{01}}{c_{01} + c_{10}}.$$

By Bayes theorem, this is equivalent to the inequality

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > \frac{c_{10}}{c_{01}} \frac{P(M_0)}{P(M_1)}.$$

Notice that the value of $P(M_0|\mathbf{x})$ is a measure, in a probability scale, of the strength we have in accepting the model M_0. We now observe a first important difference between the p-value and the Bayesian report. While the former is obtained by integration over a region of the sample space, namely the rejection region, the latter is obtained directly from the loss function and the prior probabilities assigned to the models.

The extension to the more realistic case of a parametric families of densities is straightforward. Consider the sampling models $\{f(x|\theta_0), \theta_0 \in \Theta_0\}$ and $\{f(x|\theta_1), \theta_1 \in \Theta_1\}$. A complete Bayesian specification of the models needs prior distributions for the parameter θ_0 and θ_1, that is

$$M_0 : \{f(x|\theta_0), \pi_0(\theta_0)\},$$

and

$$M_1 : \{f(x|\theta_1), \pi_1(\theta_1)\}.$$

Then under the loss function given above, assuming $c_{01} = c_{10}$ and $P(M_0) = P(M_1) = 1/2$, the model M_0 is to be rejected if

$$P(M_0|\mathbf{x}) = \frac{1}{1 + B_{10}(\mathbf{x})} < 1/2, \tag{2}$$

where $B_{10}(\mathbf{x})$, the Bayes factor for models $\{M_1, M_0\}$, is the ratio of the marginal density, sometimes called the integrated or marginal likelihood, of the data under the two models, that is

$$B_{10}(\mathbf{x}) = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})} = \frac{\int f(x|\theta_1)\pi_1(\theta_1)\, d\theta_1}{\int f(x|\theta_0)\pi_0(\theta_0)\, d\theta_0}.$$

We note that a second important difference between the *p*-values and the Bayesian method is that while in the *p*-value approach the parameters of the null and the alternative hypothesis are estimated using the maximum likelihood method, in the Bayesian approach they are integrated out using prior distributions.

For nested models it can be shown that, under mild conditions, the Bayesian procedure chooses the correct model with probability that tends to one as the sample size increases, so that it is a consistent procedure under both the null and the alternative hypothesis (see, for instance, O'Hagan and Forster 2004, p.182).

This approach needs the specification of the losses c_{ij}, $i, j = 0, 1$, the prior distributions for parameters $\pi_0(\theta_0)$ and $\pi_1(\theta_1)$, and the model prior $(P(M_0), P(M_1))$. While the specification of the loss function and the model prior seems to be a plausible task in real applications, the specification of subjective priors for parameters is admittedly a hard task. For instance, when θ_1 represents the vector of regression coefficients and the variance error of a regression model, to specify the prior is far from trivial. More so when the null parameter θ_0 is a subvector of the set of regression coefficients of θ_1, thus indicating that a submodel is being considered plausible and a testing problem is called for.

This problem is an important example in which the use of objective priors is fully justified. Unfortunately, the priors considered for estimation, as the Jeffreys or the reference priors by Berger and Bernardo (1992), are typically improper so that they depend on arbitrary multiplicative constants that leave the Bayes factor ill-defined as the following simple example shows.

Example 1 Suppose that X is a random variable with distribution $N(x|\mu, \sigma^2)$, with both parameters unknown, and we want to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. This is equivalent to choosing between models $M_0 : \{x|\sigma_0 \sim N(x|0, \sigma_0), \pi_0(\sigma_0)\}$ and

$\{x|\mu,\sigma_1 \sim N(x|\mu,\sigma_1),\pi_1(\mu,\sigma_1)\}$. The reference prior for the parameter of the null model is $\pi_0(\sigma_0) = c_0/\sigma_0$, where c_0 is an arbitrary positive constant that cannot be specified because π_0 is improper. Likewise, the reference prior for the parameter of the alternative model is $\pi_1(\mu,\sigma_1) = c_1/\sigma_1$, where again c_1 is an arbitrary positive constant. Therefore, the Bayes factor $B_{10}(\mathbf{x})$ is defined up to the multiplicative constant c_1/c_0, whatever the data $\mathbf{x}$ might be.

3.1 Intrinsic priors

Consider the Bayesian models

$$M_0 : \{f(x|\theta_0),\pi_0^N(\theta_0)\},$$

and

$$M_1 : \{f(x|\theta_1),\pi_1^N(\theta_1)\}$$

where π_0^N and π_1^N are objective, or default, improper priors.

Lempers (1971, section 5.3) overcomes the difficulty that the Bayes factor for improper priors is not well defined by considering a partial Bayes factor. This is a Bayes factor constructed as follows. A part of the sample $\mathbf{x}$, the training part, is devoted to converting the reference improper priors for the parameters of the null and alternative models into proper posteriors. Then, with the rest of the sample the Bayes factor is computed using these proper posteriors as priors. That is, the whole sample $\mathbf{x}$ is split into $(\mathbf{x} = \mathbf{x}(t),\mathbf{x}(n-t))$, where $\mathbf{x}(t)$ is the training sample of the vector and $\mathbf{x}(n-t)$ the remaining one. Then, the posterior distributions for the above priors are

$$\pi_i(\theta_i|\mathbf{x}(t)) = \frac{f(\mathbf{x}(t)|\theta_i)\pi_i^N(\theta_i)}{\int f(\mathbf{x}(t)|\theta_i)\pi_i^N(\theta_i)\,d\theta_i}, \quad i = 0,1,$$

that are now proper. Now, using the above posteriors as priors, the Bayes factor for the rest of the data $\mathbf{x}(n-t)$ turns out to be

$$B_{10}^P(\mathbf{x}) = \frac{\int f(\mathbf{x}|\theta_1)\pi_1^N(\theta_1)\,d\theta_1}{\int f(\mathbf{x}|\theta_0)\pi_0^N(\theta_0)\,d\theta_0}\frac{\int f(\mathbf{x}(t)|\theta_0)\pi_0^N(\theta_0)\,d\theta_0}{\int f(\mathbf{x}(t)|\theta_1)\pi_1^N(\theta_1)\,d\theta_1} = B_{10}^N(\mathbf{x})B_{01}(\mathbf{x}(t)).$$

Note that the partial Bayes factor is a well defined Bayes factor that uses each of the components of the sample only once. However, it does depend on the specific training sample $\mathbf{x}(t)$ we choose.

To avoid the arbitrariness in choosing the training sample Berger and Pericchi (1996) suggested computing the partial Bayes factors for all possible training samples with minimal size $\mathbf{x}(\ell)$, and then computing the mean of those partial Bayes factors. The number ℓ is chosen so that the marginals of $\mathbf{x}(\ell)$ with respect to the improper priors are positive and finite so that the factor $B_{01}(\mathbf{x}(\ell))$ is well-defined up to the multiplicative constant c_0/c_1.

The resulting value was called the arithmetic intrinsic Bayes factor, which does not depend on any arbitrary constant nor the particular training sample. Then, the Bayes factor appearing in (2) is replaced with the arithmetic intrinsic Bayes factor for computing the null posterior probability "as if" it were a Bayes factor.

We observe that the arithmetic intrinsic Bayes factor is a mean of (partial) Bayes factors and hence it reuses the sample observations. We also observe that for some subsamples of minimal size it might be the case that the marginal $m_i(\mathbf{x}(\ell))$ could be zero or infinite. In that case, the Bayes factor is not well defined and we adopt the convention of not considering those subsamples. This implies that the arithmetic intrinsic Bayes factor might be quite unstable depending on the nature of the sample at hand. Some samples can have very many nice subsamples of minimal size but others may not have so many.

However, to use the arithmetic intrinsic Bayes factor "as if" it were a Bayes factor is, in our opinion, not the best use we can give to the arithmetic intrinsic Bayes factor. It can be better employed as a tool for constructing priors. In fact, the arithmetic intrinsic Bayes factor is not a Bayes factor although as the sample size increases it becomes more and more stable and tends to be a Bayes factor for the so called intrinsic priors. Thus, if we use theoretical training samples instead of actual samples along with a limiting procedure we end up with intrinsic priors (Moreno *et al.* 1998).

Given the model

$$M_0 : \{f(x|\theta_0), \pi_0^N(\theta_0)\}$$

and

$$M_1 : \{f(x|\theta_1), \pi_1^N(\theta_1)\},$$

where $f(x|\theta_0)$ is nested into $f(x|\theta_1)$ and π_1^N is improper, the following statements can be proven.

(i) The intrinsic prior for θ_1 conditional on an arbitrary but fixed point θ_0 is given by

$$\pi^I(\theta_1|\theta_0) = \pi_1^N(\theta_1) E_{X(\ell)|\theta_1} \frac{f(X(\ell)|\theta_0)}{\int f(X(\ell)|\theta_1)\pi_1^N(\theta_1)\, d\theta_1},$$

where $X(\ell)$ is a vector of dimension ℓ with i.i.d components and distribution $f(x|\theta_1)$, such that

$$0 < \int f(X(\ell)|\theta_1)\pi_1^N(\theta_1)\,d\theta_1 < \infty,$$

ℓ being the smallest natural number satisfying the above inequality. Roughly speaking, ℓ coincides with the dimension of θ_1 .

(ii) $\pi^I(\theta_1|\theta_0)$ is a probability density for θ_1, for any fixed θ_0.

(iii) If the default prior $\pi_0^N(\theta_0)$ is also improper, the ratio

$$B_{10}^I(\mathbf{x}) = \frac{\int f(\mathbf{x}|\theta_1)\pi^I(\theta_1|\theta_0)\pi_0^N(\theta_0)\,d\theta_0\,d\theta_1}{\int f(x|\theta_0)\pi_0^N(\theta_0)\,d\theta_0} \tag{3}$$

is the limit of the sequence of Bayes factors given by

$$B_i = \frac{\int_{C_i} f(\mathbf{x}|\theta_1)\pi^I(\theta_1|\theta_0)\pi_0^N(\theta_0|C_i)\,d\theta_0\,d\theta_1}{\int_{C_i} f(\mathbf{x}|\theta_0)\pi_0^N(\theta_0|C_i)\,d\theta_0},$$

where $\{C_i, i \geq 1\}$ is a covering monotone increasing sequence of sets in Θ_0. Of course it can be shown that the limiting value (3) does not depend on the chosen sequence $\{C_i, i \geq 1\}$.

In summary, intrinsic priors are well defined priors for testing problem involving nested models. For some particular non-nested models intrinsic priors can also be defined (Cano *et al*. 2004). The Bayes factor for intrinsic priors can be seen as the stabilized version of the arithmetic intrinsic Bayes factor. Further, as the sample size n tends to infinity the sequence of intrinsic posterior probabilities of model M_0

$$P(M_0|x_1,\ldots,x_n) = \frac{1}{1 + B_{10}^I(x_1,\ldots,x_n)}$$

tends to one when sampling from the null and tends to zero when sampling from the alternative, so that the intrinsic Bayesian procedure is consistent; for a result in this direction see Moreno and Girón (2005a) for the case of the general normal linear model.

4 The two sample problem

A p-value does not always exist, so that some sort of "approximation" is in that case necessary. A classical example in which this situation ocurrs is that of comparing the means of two normal distributions with unknown variances. Let $N(x_1|\mu_1,\sigma_1^2)$, $N(x_2|\mu_2,\sigma_2^2)$ be two normal distributions where the means μ_1, μ_2 and variances σ_1^2, σ_2^2 are unknown. Suppose that samples $\mathbf{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n_1})$ and $\mathbf{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n_2})$ have been drawn independently for each of the distributions, and we are interested in testing the null $H_0 : \mu_1 = \mu_2$ versus the alternative $H_1 : \mu_1 \neq \mu_2$.

Under the frequentist point of view this problem is easily solved when $\sigma_1 = \sigma_2$ or when $\sigma_1 = k\sigma_2$ and k is known. In fact, an exact p-value can be computed by using a test statistic which follows a t-distribution.

However, standard normal theory cannot be applied when the quotient betweeen the variances σ_1^2 and σ_2^2 is unknown. This is the well-known Behrens-Fisher problem and we emphasize the fact that an exact p-value does not exist. This is a theoretically relevant result that demonstrates that frequentist testing procedures cannot be applied to some important problems. Certainly this theoretical gap is not such a serious problem from the applications viewpoint since "good" approximations – to a nonexistent solution! – were given by Fisher (1936), Wald (1955), and Welch (1947).

Under the Bayesian viewpoint the problem was "solved" by regarding it as a problem of interval estimation of the parameter $\lambda = \mu_1 - \mu_2$. From the posterior distribution of λ a $(1 - \alpha)$ highest posterior interval was computed and *the result was declared to be significant if this interval did not contain the origin* (Lindley 1970, pp. 92-93).

Notice that the posterior distribution of λ on which the inference is based – a location-scale transformation of the standard Behrens-Fisher distribution (Girón *et al.* 1999) – is obtained under the condition that $\mu_1 \neq \mu_2$; otherwise, if $\mu_1 = \mu_2$ is assumed, the posterior distribution of λ would be a point mass on zero. Therefore, using this procedure the key function for testing the null $H_0 : \mu_1 = \mu_2$ is the posterior distribution of λ conditional on the alternative hypothesis which otherwise has a posterior probability equal to zero. Of course, this cannot be the solution to the Behrens-Fisher testing problem.

In Moreno *et al.* (1999) it was shown that the Behrens-Fisher problem can be formulated as a model selection problem for nested models for which an intrinsic Bayesian solution exists. Indeed, under the null, the Bayesian default sampling model is

$$M_0 : f_0(x_1, x_2|\theta_0) = N(x_1|\mu, \tau_1^2)N(x_2|\mu, \tau_2^2),\ \pi_0^N(\theta_0) = \frac{c_0}{\tau_1\tau_2},$$

and under the alternative is

$$M_1 : f_1(x_1, x_2|\theta_1) = N(x_1|\mu_1, \sigma_1^2)N(x_2|\mu_2, \sigma_2^2),\ \pi_1^N(\theta_1) = \frac{c_1}{\sigma_1\sigma_2},$$

where $\theta_0 = (\mu, \tau_1, \tau_2)$, $\theta_1 = (\mu_1, \mu_2, \sigma_1, \sigma_2)$, π_i^N is the reference prior, and c_0, c_1 are arbitrary positive constants.

Applying the standard intrinsic methodology to these models the intrinsic prior for the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ conditional on a null point μ, τ_1, τ_2, is shown to be

$$\pi^I(\theta_1|\theta_0) = \prod_{i=1}^{2} N\left(\mu_i|\mu, \frac{\tau_i^2 + \sigma_i^2}{2}\right) HC^+(\sigma_i|0, \tau_i),$$

where $HC^+(\sigma_i|0,\tau_i)$ denotes the half-Cauchy distribution on the positive part of the real line located at 0 and with scale parameter τ_i. Under the conditional intrinsic prior the μ_i's are independent and centered at the null parameter μ and the σ_i's are also independent. Hence, the unconditional intrinsic prior distribution for μ_i is a mixture of normal distributions which has no moments. A nice property to be expected from an objective prior.

For the samples $\mathbf{x}_1$, $\mathbf{x}_2$ having size, mean and variance $(n_1, \bar{x}_1, s_1^2)$, $(n_2, \bar{x}_2, s_2^2)$ respectively, the Bayes factor for the intrinsic priors $(\pi_0^N(\theta_0), \pi^I(\theta_1))$ is given by

$$B_{10}^I(\mathbf{x}_1,\mathbf{x}_2) = \frac{2}{\pi^{5/2}} \frac{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}{\Gamma(\frac{n_1+n_2-1}{2})} \frac{B}{A},$$

where $B = \int_{-\infty}^{\infty} \left[\prod_{i=1}^{2} I_i(\mu)\right] d\mu$,

$$I_i(\mu) = \int_0^{\pi/2} \frac{d\varphi_i}{b_i(\mu,\varphi_i)},$$

$$b_i(\mu,\varphi_i) = \frac{(\sin\varphi_i)^{n_i-1}}{\left(\frac{1}{2}+\frac{\sin^2\varphi_i}{n_i}\right)^{-1/2}} \left(\frac{n_i s_i^2}{\sin^2\varphi_i} + \frac{(\bar{x}_i-\mu)^2}{\left(\frac{1}{2}+\frac{\sin^2\varphi_i}{n_i}\right)^{n_i/2}} \right)^{n_i/2},$$

and

$$A = \int_0^{\pi/2} \frac{d\varphi}{a(\varphi)},$$

$$a(\varphi) = \frac{\sin^{n_1}\varphi \cos^{n_2}\varphi}{\left(\frac{\sin^2\varphi}{n_1}+\frac{\cos^2\varphi}{n_2}\right)^{-1/2}} \left(\frac{n_1 s_1^2}{\sin^2\varphi} + \frac{n_2 s_2^2}{\cos^2\varphi} + \frac{(\bar{x}_1-\bar{x}_2)^2}{\frac{\sin^2\varphi}{n_1}+\frac{\cos^2\varphi}{n_2}} \right)^{(n_1+n_2-1)/2}.$$

It is easy to see that the Bayes factor for intrinsic priors B_{10}^I depends on the sample $(\mathbf{x}_1,\mathbf{x}_2)$ through the statistic $(s_1^2, s_2^2, |\bar{x}_1-\bar{x}_2|, n_1, n_2)$. Since the p-value in the Welch's approximation also depends on the the sample through this statistic it follows that there is a one-to-one relationship between the p-values and the null model posterior probabilities.

As an illustration of this relationship, in Table 1 we display p-values and null model posterior probabilities for sample observations with

$$n_1 = 200,\ s_1^2 = 12,\ n_2 = 120,\ s_2^2 = 40,$$

Table 1: *Comparison of p-values and null posterior probabilities for* $n_1 = 200$ *and* $n_2 = 120$.

$\lvert\bar{x}_1 - \bar{x}_2\rvert$	t	p-value	$P(M_0\lvert\mathbf{x}_1, \mathbf{x}_2)$
0.0	0.00	1.00	0.94
0.5	0.79	0.43	0.92
1.2	1.91	0.06	0.72
1.3	2.06	0.04	0.65
1.4	2.22	0.03	0.57
1.5	2.38	0.02	0.49
2.0	3.17	0.001	0.10

and several values of the difference $|\bar{x}_1 - \bar{x}_2|$ and the corresponding t statistic defined as

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

From the numbers in Table 1 we conclude that when the values of $|\bar{x}_1 - \bar{x}_2|$ are close to zero or large enough both procedures make the correct decision. However, when the empirical evidence is not conclusive, a situation far beyond intuition for which statistical methods are unavoidable, there is a strong disagreement between the report provided by Welch's p-values and that of the intrinsic null posterior probabilities. For instance, for values of the t statistic between 2.00 and 2.35, the p-values show evidence against the null hypothesis, stronger as t increases, while the null posterior probabilities show the opposite, they still favour the null hypothesis.

Such a disagreement heavily depends on the sample size. If we reduce the above sample sizes to $n_1 = 20$ and $n_2 = 12$, while maintaining the values of s_1^2 and s_2^2, the frequentist and Bayesian reports are not so strongly contradictory, as seen from Table 2, and it may happen that a p-value accepts the null but the corresponding posterior probability of the null may be less than 0.5 and then the Bayesian test rejects it. For instance, for $|\bar{x}_1 - \bar{x}_2| = 4.22$ or $t = 2.04$, the p-value is 0.06 while the null posterior probability is smaller than 0.5.

Table 2: *Comparison of p-values and null posterior probabilities for* $n_1 = 20$ *and* $n_2 = 12$.

$\lvert\bar{x}_1 - \bar{x}_2\rvert$	t	p-value	$P(M_0\lvert\mathbf{x}_1, \mathbf{x}_2)$
0.00	0.00	1.00	0.83
2.20	1.06	0.30	0.75
4.22	2.04	0.06	0.46
5.00	2.42	0.03	0.32
10.00	4.80	0.002	0.008

In passing, we note that when $|\bar{x}_1 - \bar{x}_2| = 0$ we expect both the p-value and null posterior probability to be large. Since the sample size is finite we should not expect

the p-value to attain its maximum value of one, but it does. However, the null posterior probability is always strictly smaller than one.

5 Testing hypotheses in linear regression

A scenario where the discrepancies between the p-values and the objective Bayesian test are apparent is that of testing that some regression coefficients of a linear regression model are equal to zero. Suppose that the observable random variable y follows the normal linear model

$$y = \sum_{i=1}^{k} \alpha_i x_i + \varepsilon,$$

where the random error term $\varepsilon \sim N(\varepsilon|0, \sigma^2)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^t$ is the vector of regression coefficients, and $(x_1, \ldots, x_k)$ is a set of potential explanatory variables. Given n independent observations $\mathbf{y} = (y_1, \ldots, y_n)^t$ from the model and denoting the design matrix by

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1k} \\ x_{21} & \ldots & x_{2k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nk} \end{pmatrix},$$

the likelihood function of $(\boldsymbol{\alpha}, \sigma)$ is given by the density of a $N_n(\mathbf{y}|\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)$, where it is assumed that $\mathbf{X}$ is of full rank k, $(k < n)$.

Consider the partition of $\boldsymbol{\alpha}$ as $\boldsymbol{\alpha}^t = (\boldsymbol{\alpha}_0^t, \boldsymbol{\alpha}_1^t)$ and the corresponding partition of the columns of $\mathbf{X} = (\mathbf{X}_0|\mathbf{X}_1)$, so that $\mathbf{X}_0$ is of dimensions $n \times k_0$ and $\mathbf{X}_1$ is $n \times k_1$, where $k_1 = k - k_0$.

In this setting, an important problem consists in testing that some of the covariates have no influence on the variable y. That is, we are interested in testing the null $H_0 : \boldsymbol{\alpha}_0 = 0$ versus the alternative $H_1 : \boldsymbol{\alpha}_0 \neq 0$. This is the natural way of reducing the complexity of the original linear model proposed. If the null is accepted the implication is that the covariates $x_1, \ldots, x_{k_0}$ will not be considered as explanatory variables.

5.1 The uniformly most powerful test

The frequentist testing procedure, derived from the likelihood ratio test, is based on the distribution of the ratio $\mathcal{B}_n = \mathrm{SS}/\mathrm{SS}_1$ of the quadratic forms

$$\mathrm{SS} = \mathbf{y}^t(\mathbf{I}_n - \mathbf{H})\mathbf{y}, \quad \mathrm{SS}_1 = \mathbf{y}^t(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ and $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t$ are the hat matrices of the full and the reduced models, respectively. It is well known, from standard linear model theory, that the sampling distribution of the $\mathcal{B}_n$ statistic under the null $H_0 : \boldsymbol{\alpha}_0 = 0$ is

$$\mathcal{B}_n|H_0 \sim Be\left(\cdot|\frac{n-k}{2}, \frac{k_0}{2}\right),$$

where $Be(\cdot|\alpha, \beta)$ denotes the beta distribution with parameters α and β.

When sampling from the alternative $H_1 : \boldsymbol{\alpha}_0 \neq 0$, the corresponding distribution is

$$1 - \mathcal{B}_n|H_1 \sim Be'\Big(\cdot|\frac{k_0}{2}, \frac{n-k}{2}; \delta\Big),$$

where

$$\delta = \boldsymbol{\alpha}_0^t\mathbf{X}_0^t(\mathbf{I}_n - \mathbf{H}_1)\mathbf{X}_0\boldsymbol{\alpha}_0$$

and $Be'(\cdot|\alpha, \beta; \delta)$ denotes the noncentral beta distribution with parameters α and β and noncentrality parameter δ. If $H_0 : \boldsymbol{\alpha}_0 = 0$ is true, then $\delta = 0$, and the noncentral distribution reduces to the central one.

The UMP test of size α (Lehmann 1986, theorem 5, pp. 300 and pp. 369) has the following critical region

$$\text{Reject } H_0 \text{ when } \mathcal{B}_n \leq I_\alpha^{-1}\Big(\frac{n-k}{2}, \frac{k_0}{2}\Big),$$

where $I_\alpha^{-1}((n-k)/2, k_0/2)$ denotes the α fractile of the Beta distribution $Be(\cdot|((n-k)/2, k_0/2)$.

Likewise, for a given sampling value $\mathcal{B}_n$, the p-value is given by

$$p = \int_0^{\mathcal{B}_n} be\left(z|\frac{n-k}{2}, \frac{k_0}{2}\right) dz.$$

where $be(z|(n-k)/2, k_0/2)$ denotes the density of the corresponding beta distribution.

We remark that the p-value is an increasing function of $\mathcal{B}_n$ so that small values of $\mathcal{B}_n$ contain evidence against the null hypothesis.

This test is usually written in terms of the F-statistic, which is related to $\mathcal{B}_n$, by

$$F = \frac{n-k}{k_0}\frac{1-\mathcal{B}_n}{\mathcal{B}_n},$$

although for numerical illustrations it is more convenient to use the bounded $\mathcal{B}_n$ statistic instead of the unbounded F.

5.2 The objective Bayesian test

The default Bayesian formulation of this testing problem would be that of choosing between the Bayesian models

$$M_0 : N_n(\mathbf{y}|\mathbf{X}_1\gamma_1, \sigma_0^2\mathbf{I}_n),\ \pi_0^N(\gamma_1, \sigma_0) = \frac{c_0}{\sigma_0},$$

and

$$M_1 : N_n(\mathbf{y}|\mathbf{X}\alpha, \sigma_1^2\mathbf{I}_n),\ \pi_1^N(\boldsymbol{\alpha}, \sigma_1) = \frac{c_1}{\sigma_1},$$

where π^N represents the usual improper reference prior (Berger and Bernardo, 1992) for estimating the regression coefficients and the standard error. Unfortunately these priors are improper and hence cannot be used for solving the above testing problem.

Application of the standard intrinsic methodology (Moreno, Girón and Torres 2003, Girón *et al*. 2004) renders the intrinsic priors of $(\boldsymbol{\alpha}, \sigma_1)$ conditional on (γ_1, σ_0) as

$$\pi^I(\boldsymbol{\alpha}, \sigma_1|\gamma_1, \sigma_0) = \frac{2}{\pi\sigma_0(1 + \sigma_1^2/\sigma_0^2)} N_k(\boldsymbol{\alpha}|\tilde{\boldsymbol{\gamma}}_1, (\sigma_0^2 + \sigma_1^2)\mathbf{W}^{-1}),$$

where $\tilde{\gamma}_1^t = (\mathbf{0}^t, \gamma_1^t)$ and $\mathbf{W}^{-1}$ is

$$\mathbf{W}^{-1} = \frac{n}{k+1}(\mathbf{X}^t\mathbf{X})^{-1}.$$

We note that the conditional intrinsic prior for the parameter of the alternative $\boldsymbol{\alpha}$ is centered at the null parameter $\tilde{\gamma}_1$. Further, the conditional intrinsic prior for σ_1 is a half Cauchy located at zero and with scale parameter σ_0. This implies that the conditional intrinsic prior has no moments, a desirable property for a default prior. The unconditional intrinsic prior for $(\boldsymbol{\alpha}, \sigma_1)$ is given by

$$\pi^I(\boldsymbol{\alpha}, \sigma_1) = \int \pi^I(\boldsymbol{\alpha}, \sigma_1|\gamma_1, \sigma_0)\pi_0^N(\gamma_1, \sigma_0)d\gamma_1 d\sigma_0.$$

Of course, this prior is fully automatic, i.e. does not depend on any tuning parameters nor processes any subjective prior information.

Using the so called pair of intrinsic priors $\pi_0^N(\gamma_1, \sigma_0)$ and $\pi^I(\boldsymbol{\alpha}, \sigma_1)$, the intrinsic posterior probability of model M_0 is given by

$$P(M_0|\mathbf{y}, \mathbf{X}) = \frac{1}{1 + B_{10}}$$

where

$$B_{10} = \frac{2(k+1)^{k_0/2}}{\pi} \int_0^{\pi/2} \frac{\sin^{k_0}\varphi\,(n+(k+1)\sin^2\varphi)^{(n-k)/2}}{(n\mathcal{B}_n+(k+1)\sin^2\varphi)^{(n-k_1)/2}} d\varphi. \tag{4}$$

From this expression, and also from the frequentist analysis of the testing problem in subsection 5.1, it follows that for fixed values of the sample size n, the number of covariates k and the dimension of the null hypothesis k_1, the statistic $\mathcal{B}_n$ is a sufficient statistic for the testing problem, as the Bayes factor for the intrinsic priors does not depend on other ancillary statistics such as happens with other Bayes factors for linear models found in the literature, which depend on the quotient of the determinants $|\mathbf{X}^t\mathbf{X}|$ and $|\mathbf{X}_1^t\mathbf{X}_1|$.

Bayesian testing procedures different from the above one have been given by Berger and Pericchi (1996), O'Hagan (1995) and Zellner (1986) who proposed the use of the arithmetic intrinsic Bayes factor, the fractional Bayes factor and the Bayes factor derived from the g-priors, respectively. Except for the arithmetic intrinsic Bayes factor, the other two proposals depend on some tuning parameters which have to be adjusted.

Let us mention that for normal linear models the O'Hagan fractional Bayes factor provides sensible *fractional priors* for testing problems in a similar asymptotic way as the arithmetic intrinsic Bayes factor provides *intrinsic priors* (Moreno 1997). For so doing, the tuning parameter in the fractional Bayes factor is fixed as the quotient m/n, where m is the minimal training sample size. The results obtained when using Bayes factors for fractional priors are very close to those provided by Bayes factors for intrinsic priors, and hence only intrinsic priors are being considered here.

5.3 Comparing the frequentist and Bayesian tests

The dependence of the p-value and the posterior probability of the null on the sufficient statistic $(\mathcal{B}_n, n, k, k_1)$, where n, k, and k_1 are ancillary makes possible the comparison of the frequentist and objective Bayesian test.

For fixed values of the ancillaries n, k and k_1, the p-value and the intrinsic posterior probability of the null model $P(M_0|\mathcal{B}_n, n, k, k_1)$ are monotone increasing functions of the $\mathcal{B}_n$ statistic. This permits us to establish a one-to-one relation between both measures of evidence through the parametric equations

$$\begin{aligned} y &= P(M_0|b, n, k, k_1) \\ p &= I_b\left(\frac{n-k}{2}, \frac{k-k_1}{2}\right), \end{aligned} \tag{5}$$

where the parameter b, the sufficient statistic, ranges in the interval $[0, 1]$.

The separate behaviour of y and p as the sufficient statistic b goes to zero or one is as follows. The null posterior probability and the p-value go to zero as b tends to zero, whatever the values of the ancillaries, as

$$\lim_{b\to 0} P(M_0|b,n,k,k_1) = 0 \quad \text{and} \quad \lim_{b\to 0} I_b\Big(\frac{n-k}{2}, \frac{k-k_1}{2}\Big) = 0.$$

If $\mathcal{B}_n = 0$, then the residual sum of squares of the full model SS is also 0; this means that there is no uncertainty in the full model, i.e. it is deterministic; thus, the reduced model M_0 has zero posterior probability and the p-value is also zero, so that the full model is obviously accepted.

When $\mathcal{B}_n$ tends to one, then the p-value tends to one whatever the values of the ancillaries, but the null posterior probability tends to a number strictly smaller that one (Theorem 2.2 in Girón *et al* 2004) which, on the other hand, tends to one as n goes to infinity.

In this case, as $\mathcal{B}_n = 1$, the residual sum of squares of the full SS and the reduced null model SS_1 are the same so that the data favour either model equally; however, the frequentist evidence in favour of the reduced model M_0 is one as if there were no uncertainty about what model to choose but, on the other hand, the Bayesian test accounts for the uncertainty inherent in the data rendering a posterior probability of M_0 greater than 1/2 but strictly less than 1; hence, the Bayesian test chooses the simpler model, which is a consequence of the built-in Occam's razor implicit in the objective Bayesian test.

From the above equations (5) we can eliminate the parameter b to obtain a explicit equation of the null posterior probability as a function of the p-value, n, k and k_1,

$$y = P(M_0|I_p^{-1}(n-k/2, k-k_1/2), n, k, k_1).$$

From this equation it follows that for fixed values of n, k and k_1, the null posterior probability is an increasing function of the p-value. Therefore, the difference between the frequentist and Bayesian measures of evidence is a simple calibration problem. For this reason this curve is given the name of *calibration curve* in Girón *et al* (2004).

Unfortunately, the null posterior probability also depends on n, k, k_1 so that the properties of the calibration curve have to be established in a case-by-case basis. When simultaneous hypothesis testing are considered we have to jump among different calibration curves hence losing the monotonicity between the null posterior probabilities and the p-values. In this latter case calibration is no longer possible.

In the remainder of this section the number of possible regressors k will be keep fixed. In Figures 1 and 2 we display the typical behavior of the calibration curves, first, for different values of the sample size n and fixed k_1, and second, for different values of k_1 and fixed n.

The calibration curves in Figure 1 correspond to $k = 10$ and $k_1 = 9$ and $n = 20$, 50 and 90; they indicate that for a given p-value the null posterior probability increases as the sample size increases. Further, from the consistency of the Bayes factor for intrinsic priors it follows that the slope of the calibration curve at the origin tends to infinity as n increases. This shows that the evidence against the null conveyed by the p-values should be diminished as the sample size n increases in order to reconcile the frequentist and Bayesian test. Otherwise, we would reject a null hypothesis that has a very high posterior probability.

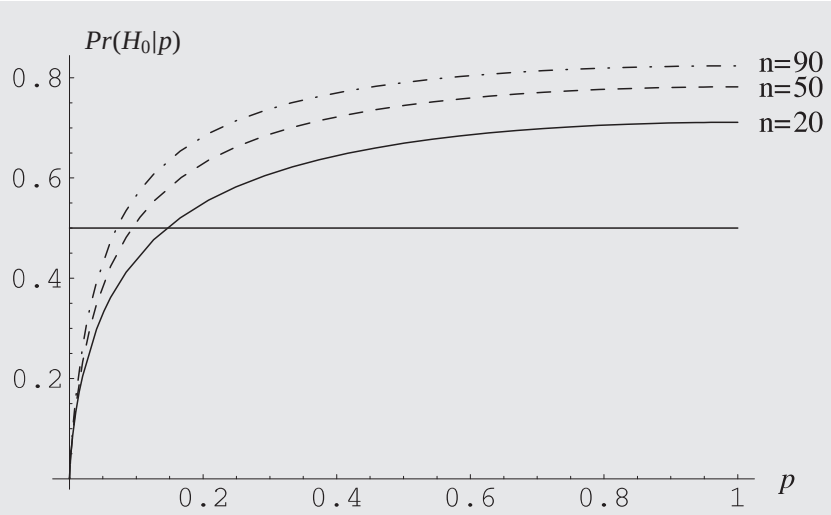

Figure 1: *Three calibration curves for different sample sizes* $n = 20$, $n = 50$ *and* $n = 90$, *when the number of regressors is* $k = 10$, *and* $k_1 = 9$.

The curves in Figure 2, for a fixed value of the sample size n, indicate that small p-values correspond to small posterior probabilities when k_1 is large, and also that the posterior probability increases as k_1 decreases. This implies that for small values of k_1 a p-value would reject a null hypothesis that has a large posterior probability, a fact which is generally acknowledged in the literature. But, on the other hand, for large values of k_1 a p-value would accept a null hypothesis that has a small posterior probability. Notice, in Figure 2, that for the curve with $k_1 = 9$ there is an interval of p-values larger than 0.05 whose corresponding posterior probabilities of the null are smaller than 1/2. This important fact, which is generally overlooked by the feeling that p-values tend to reject the null hypothesis more often than the Bayesian tests, also reveals that the opposite may happen sometimes; namely, that a null hypothesis may be not rejected by the frequentist p-value and rejected by the Bayesian test for the same data.

Figure 2 also indicates that, when comparing several nulls of different dimension k_1 that convey the same frequentist evidence, the Bayesian test chooses the simplest

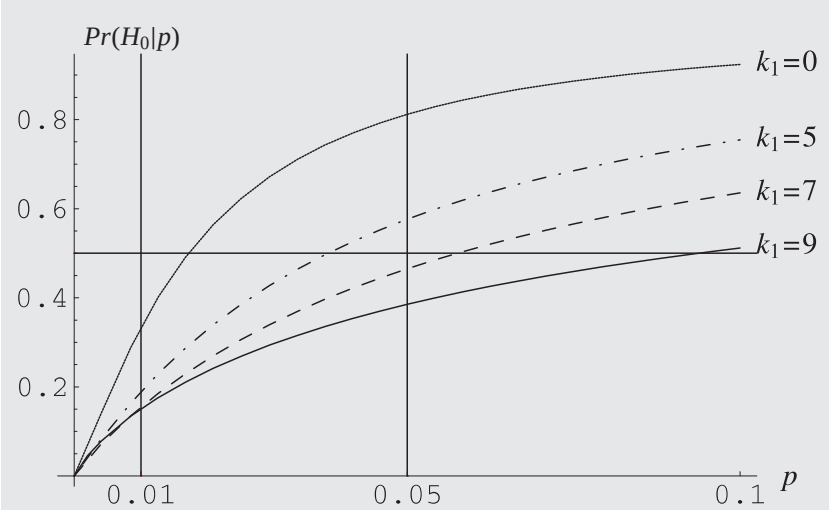

Figure 2: *Four calibration curves, plotted for values of p-values in the interval* $(0, 0.10)$*, for different choices of* $k_1 = 0, 5, 7, 9$*, when the sample size is* $n = 50$*, and the number of regressors is* $k = 10$*.*

hypothesis or model. As before, this is another instance of the automatic Occam's razor property implicit in the objective Bayesian test.

5.4 Variable selection in regression

Although the purpose of this paper is to critically revise the similarities and, above all, the enormous differences between the two approaches to hypothesis testing, we want to devote this last section to the important problem of variable selection in normal linear regression, where we have to consider simultaneously a large number of multiple tests of different dimensions k_1, and then provide an ordering of the plausibility of the different models considered. A recent contribution to the subject, mainly from a frequentist perspective, is the revised version of the classical monograph by Miller (2002), where a chapter on Bayesian variable selection has been added.

We do not discuss here sequential, or non-exhaustive, frequentist variable selection procedures such as forward, backward and stepwise selection methods nor the more recent ones such as the lasso or shrinkage methods, in order to concentrate our discussion on exhaustive search methods from the frequentist and Bayesian perspectives. Neither do we discuss here the well known asymptotic *BIC* criterion for model selection, because we can dispense with it as we have the non-asymptotically based *from above* and *from below* Bayesian criteria.

The *all subsets selection criterion* is based on the idea of classifying the set of all models, say $\mathcal{M}$, in k disjoint classes, where k is the number of covariates of the full model, according to the number of covariates j, i.e. $\mathcal{M} = \bigcup_{j=1}^{k} \mathcal{M}_j$. Within each class,

the model with minimum residual sum of squares SS_j is chosen, so that we end up with *k maximal* models, one for each class. Note that *p*-values or the equivalent *F* statistic provide the same ordering within the class $\mathcal{M}_j$ and, consequently, the same sets of maximals, but these criteria turn to be useless when comparing models with different number of covariates, a fact which is well recognized by frequentist statisticians.

Hence, the problem of choosing the *best* model within the maximals, is far from trivial and a large number of procedures have been proposed in the literature. The underlying idea is to correct the SS_j, or a simple function of it and the ancillaries to account for the different number of covariates, as do the well known adjusted R^2, Mallows C_p, and the AIC criteria. Unfortunately, these corrections do not always work properly.

From the objective Bayesian viewpoint, which is mainly based on the results of Section 5.2, two procedures for model selection have been proposed. The main difference between these procedures relies on the form of encompassing – that is, of nesting – the class of all possible submodels. The so called *from above* procedure (Casella and Moreno 2005, Girón *et al.* 2004) is based on comparing all the submodels with the full model, and ordering them according to the posterior probabilities of all submodels using the formulae of Section 5.2. We denote these posterior probabilities by $P_{fa}(M_i|\mathcal{S}_n)$ for any submodel M_i, where $\mathcal{S}_n = (\mathcal{B}_n, n, k, k_1)$. The interpretation of these probabilities is that the model with highest posterior probability represents the most plausible reduction in complexity from the full model, the second highest the second most plausible model, and so on.

As these posterior probabilities are monote increasing functions of the $\mathcal{B}_n$ statistic for fixed ancillaries, this means that within each class of models $\mathcal{M}_j$ the ordering provided is the same as the one based on the residual sum of squares SS_j. Thus, the Bayesian solution *from above* is the maximal model having the largest posterior probability of its corresponding model. No need for extra adjustment!

The so called *from below* procedure (Girón *et al.* 2005b and Moreno and Girón 2005), based on the simple fact that the intercept only model is nested into any other possible model as far as it includes the intercept, produces a possibly different ordering of all the submodels of the full model. The ordering is now based on the Bayes factors resulting from comparing the current submodel with the intercept only model. Further, it turns out that this procedure provides a coherent set of model posterior probabilities on the set of all possible submodels denoted by $P_{fb}(M_i|\mathcal{S}_n)$, and these coherent probabilities are monote increasing functions of the R^2 statistic as now $\mathcal{B}_n = 1 - R^2$, which in turn, is also a monote decreasing function of the residual sum of squares SS_j of the corresponding submodel. This means, as with the *from above* Bayesian criterion, that the model chosen by the from below criterion is also the maximal model having the largest Bayes factor or, equivalently, the highest model posterior probability $P_{fb}(M_i|\mathcal{S}_n)$.

The main conclusion derived from these comparisons is, first, that the two Bayesian criteria always choose a maximal model, i.e. they are compatible with the best subsets

partial ordering and, second, that they are fully automatic in the sense that no tuning of extra parameters, neither the use of outside information nor additional criteria, is needed.

Table 3: *Comparison of different variable selection criteria for Hald's data*

Models	**From below** $P_{fb}(M_i\|\mathcal{S}_n)$	**From above** $P_{fa}(M_i\|\mathcal{S}_n)$	R^2	**Adjusted** R^2	**Mallows** C_p
$\{x_1, x_2\}$	0.5466	0.7407	0.9787	0.9744	2.6782
$\{x_1, x_4\}$	0.1766	0.5364	0.9725	0.9670	5.4958
$\{x_1, x_2, x_4\}$	0.0889	0.7231	0.9823	0.9764	3.0182
$\{x_1, x_2, x_3\}$	0.0879	0.7211	0.9823	0.9764	3.0413
$\{x_1, x_3, x_4\}$	0.0708	0.6809	0.9813	0.9750	3.4968
$\{x_2, x_3, x_4\}$	0.0165	0.3780	0.9728	0.9638	7.3375

Table 3 compares the results of several model selection criteria for the famous Hald's data on the composition of cement. The analysis illustrates the Occam's razor property of the Bayesian criteria. Note also that, for these data, the adjusted R^2 does not adjust the ordering provided by the original R^2 for the most plausible models.

A large simulation study, see Moreno and Girón (2005b) for the description and extent of the study, has shown that the adjusted R^2 performs very poorly in almost all situations, a well known fact. Mallow's C_p and the *AIC* criteria perform in a very similar way –another well known fact– but they show a poorer behaviour when compared with either Bayesian criteria in most circumstantes. This suggests that the Bayesian criteria account for the difference in dimensionality in some automatic way, hidden in the formulae of their corresponding Bayes factors, in the same manner they automatically obey Occam's razor principle.

Table 4 illustrates these comments for a medium size linear model, $k = 6$, i.e. with five covariates excluding the intercept, sample size $n = 40$ and values of k_1 ranging from 2 to 6.

Table 4: *Comparison of different variable selection criteria for the simulated data.*

	N.º of covariates				
Criterion	**1**	**2**	**3**	**4**	**5**
From below	0.901	0.910	0.962	0.977	1.000
From above	0.573	0.657	0.793	0.927	1.000
Mallows C_p	0.500	0.563	0.692	0.850	1.000
Adjusted R^2	0.234	0.292	0.452	0.716	1.000

The model considered for simulation is

$$\mathbf{y} = \mathbf{X}\alpha + \varepsilon$$

where $\mathbf{y}$ is a vector of length 40, $\mathbf{X}$ is a 40×6 matrix whose entries were obtained by simulation from a standard normal distribution $N(0, 1)$, except the entries in the first

column which were set equal to 1 to include the intercept, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_6)^t$ is a vector of length 6. The error terms in $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ are i.i.d. $\varepsilon_i \sim N(0, 1)$.

After fixing $\mathbf{X}$, samples of size 5000 were simulated from the model for five different settings of the vector of regression coefficients $\boldsymbol{\alpha}$ including 1, 2 3, 4 and 5 non zero coefficients. In particular, we set

$$\begin{aligned}
\boldsymbol{\alpha}_1 &= (-1, -2, 0, 0, 0, 0) \\
\boldsymbol{\alpha}_2 &= (-1, -2, 2, 0, 0, 0) \\
\boldsymbol{\alpha}_3 &= (-1, -2, 2, -3, 0, 0) \\
\boldsymbol{\alpha}_4 &= (-1, -2, 3/2, -2, 2, 0) \\
\boldsymbol{\alpha}_5 &= (-1, -2, 3/2, -2, 2, -1).
\end{aligned}$$

The entries in Table 4 represent the proportion of times that the true model is selected in the first place in the 5000 simulations according to the four criteria and to the number of nonzero regression coefficients in the model.

The relation between the p-values, or any equivalent model selection procedure, and the Bayesian model posterior probabilities in the variable selection problem can be summarized as follows. For a fixed sample size n, models with the same number of regressors k_1 are ordered in the same manner by all criteria: p-values, R^2 and adjusted R^2, Mallows C_p, AIC and the two Bayesian procedures. However, when comparing models with different number of regressors all frequentist and Bayesian criteria generally provide different orderings of the models. But, as we have learned from the simulations, the frequentist behaviour of the Bayesian criteria generally outperforms that of the frequentist ones. Comparisons between the *from below* and *from below* Bayesian criteria are discussed in Moreno and Girón (2005b). The conclusion in that paper is that for models with a small or medium number of relevant covariates, as the one illustrated in Table 3, the *from below* criterion performs better that the *from above* one, but for models with a large number of influential covariates, the opposite may happen.

6 Discussion

Two measures of evidence for hypothesis testing, frequentist and Bayesian, have been considered and compared in this paper for some important testing problems. The case of the standard normal model is not dealt with in the paper as it is a particular case of the normal linear model with no covariates; in this case the sample size is the only ancillary, and the frequentist-Bayesian comparison or calibration just depends on the sample size.

We have first recalled that standard normal-theory does not apply to the Behrens-Fisher problem of testing the equality of the means of two normal populations under heteroscedasticity, as there is no clear-cut p-value and, consequently, frequentist theory has to resort to computing an approximate p-value by adjusting the degrees of freedom of a t-distribution on which the solution of the homoscedastic case is based.

For this problem, we have illustrated the fact that *p*-values reject the null hypothesis for data for which the Bayesian inference accepts, more markedly as both sample sizes increase. To make *p*-values more unsatisfactory we have also seen that for small sample sizes *p*-values would accept the null for data for which the Bayesian rejects.

Therefore, we have to admit that the usual interpretation of a *p*-value as a measure of evidence against the null regardless the sample size, though it entails a notable simplification, may produce wrong answers. It is clear that the enormous success of the *p*-values in the realm of applications is partially due to their simplicity for scientific communication, but the bad news is that such a simplicity may be misleading.

The study of the relation between the two measures of evidence has been extended to normal data in the presence of covariates, that is to the normal linear model. Here two new ancillaries, in addition to the sample size, arise: the number of covariates and the dimension of the null. We have illustrated that the dimension of the null hypothesis is another fundamental ancillary to be taken into account when interpreting *p*-values. The disregard of this ancillary may produce the rejection of null hypotheses that have high posterior probabilities or the acceptation of nulls that have low posterior probabilities.

Finally, we have considered the variable selection problem, a challenging multiple testing problem, because for this problem the sample size and the dimension of the null play a very important role. For selecting variables we necessarily have to jump among models whose null parameter spaces have different dimensions. To overcome this difficulty, the frequentist approach has to adjust some statistic, usually the one based on the ratio of sums of the residuals under the full and null models, in several ways to account for the different number of covariates involved. This accommodation, however, is not very convincing. The objective Bayesian approach seems to deal with this problem in a more appropiate way than the frequentist counterpart because all the ancillaries in the problem are properly taken into account. Furthermore, the objective Bayesian solution works in a fully automatic way in the sense that there is no need for adjusting any tuning parameters.

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berger, J.O. (1994). An overview of robust Bayesian analysis (with discussion). *Test*, 3, 5-124.

Berger, J.O. and Bernardo, J.M. (1992). On the development of the reference prior method. In *Bayesian Statistics 4*. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F:M. Smith, eds. Oxford University Press: Oxford, pp. 35-60.

Berger, J.O. and Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, 94, 542-554.

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91, 109-122.

Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of *p*-values and evidence (with discussion). *Journal of the American Statistical Association*, 82, 112-139.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*, New York: Wiley.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.

Cano, J.A., Kessler, M. and Moreno, E. (2004). On intrinsic priors for nonnested models. Test, 13, 445-463.

Casella, G. and Berger, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82, 106-111.

Casella, G. and Berger, R.L. (1990). *Statistical Inference*. Belmont: Wadsworth.

Casella, G. and Moreno, E. (2005). Intrinsic meta analysis of contingency tables, *Statistic in Medicine*, 24, 583-604.

Casella, G. and Moreno, E. (2005). Objective Bayesian variable selection. *Journal of the American Statistical Association* (to appear).

Fisher, R.A. (1936). The fiducial arguments in statistical inference. *Annals of Eugenics*, 6, 391-422.

Girón, F.J., Martínez, L. and Imlahi, L. (1999). A characterization of the Behrens-Fisher distribution with applications to Bayesian inference. *Comptes rendus de l'Académie des sciences de Paris, Ser I*, 701-706.

Girón, F., Martínez, L., Moreno, E. and Torres, F. (2003). Bayesian analysis of matched pairs in the presence of covariates. In *Bayesian Statistics 7*. J.M. Bernardo *et al*. (eds.), 553-563, Oxford: Oxford University Press.

Girón, F.J., Martínez, L., and Moreno, E. (2003). Bayesian analysis of matched pairs. *Journal of Statistical Planning and Inference*, 113, 49-66.

Girón, F.J., Martínez, M.L., Moreno, E. and Torres, F. (2004). Objective testing procedures in linear models. Calibration of the *p*-values. Submitted.

Girón, F.J., Moreno, E. and Martínez, L. (2005). An objective Bayesian procedure for variable selection in regression. In *Advances on distribution theory, order statistics and inference*. Eds. N. Balakrishnan *et al.*, Birkhauser Boston, (to appear).

Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press.

Lempers, F.B. (1971). *Posterior Probabilities of Alternative Linear Models.* Rotterdam: Rotterdam University Press.

Lehman, E. (1986). *Testing Statistical Hypotheses*, (second edition). New York: Wiley.

Lindley, D.V. (1970). *An Introduction to Probability and Statistics from a Bayesian Viewpoint* (Vol. 2). Cambridge: Cambridge University Press.

Miller, A.J. (2002). *Subset Selection in Regression. 2nd edition*. London: Chapman and Hall.

Moreno, E. (1997). Bayes Factor for Intrinsic and Fractional Priors in Nested Models: Bayesian Robustness. *IMS Lectures Notes-Monograph Series*, 31, 257-270.

Moreno, E. (2005). Objective Bayesian analysis for one-sided testing, *Test*, 14, 181-198.

Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypotheses testing. *Journal of the American Statistical Association*, 93, 1451-1460.

Moreno, E., Bertolino, F. and Racugno, W. (1999). Default Bayesian analysis of the Behrens-Fisher problem. *Journal of Statistical Planning and Inference*, 81, 323-333.

Moreno, E., Bertolino, F. and Racugno, W. (2000). Bayesian model selection approach to analysis of variance under heterocedasticity, *Journal of the Royal Statistical Society, Series D (The Statistician)*, 49, 1-15.

Moreno, E., Bertolino, F. and Racugno, W. (2003). Bayesian inference under partial prior information, *Scandinavian Journal of Statistics*, 30, 565-580.

Moreno, E. and Cano J.A. (1989). Testing a point null hypothesis: asymptotic robust Bayesian analysis with respect to the priors given on a subsigma field. *International Statistical Review*, 57, 221-232.

Moreno, E. and Girón, F.J. (2005*a*). Consistency of Bayes factors for linear models, *Comptes rendus de l'Académie des sciences de Paris, Ser I*, 911-914.

Moreno, E. and Girón, F.J. (2005*b*). Comparison of Bayesian objective procedures for variable selection in regression. Submitted.

Moreno, E., Girón, F.J., and Torres, F. (2003). Intrinsic priors for hypothesis testing in normal regression models. *Revista de la Real Academia de Ciencias Serie A,. Mat.*, 97, 53-61.

Moreno, E. and Liseo, B. (2003). A default Bayesian test for the number of components of a mixture. *Journal of Statistical Planning and Inference*, 111, 129-142.

Moreno, E., Torres, F., and Casella, G. (2005). Testing the equality of regression coefficients in heteroscedastic normal regression models. *Journal of Statistical Planning and Inference*, 131, 117-134.

O'Hagan, A. (1995). Fractional Bayes factor for model comparison (with discussion). *Journal of the Royal Statistical Society Series B*, 57, 99-138.

O'Hagan, A. and Forster, J. (2004). *Bayesian Inference.* Kendall's Advances Theory of Statistics (Vol. 2B). London: Arnold.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press: Oxford.

San Martini, A. and Spezaferri, F. (1984). A predictive model selection criterion. *Journal of the Royal Statistical Sociery, Series B*, 46, 296-303.

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

Wald, A. (1955). Testing the difference between the means of two normal populations with unknown standard deviations. In *Selected papers in Statistics and Probability*, T.W. Anderson *et al*. (eds.), 669-695, Stanford University Press.

Wilks, S.S. (1962). *Mathematical Statistics*. New York: Wiley

Welch, B.L. (1951). On the comparison of several means values: an alternative approach. *Biometrika*, 38, 330-336.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, P.K. Goel and A. Zellner (eds.), 233-243, Amsterdam: Elsevier.

Discussion of “On the frequentist and Bayesian approaches to hypothesis testing” by Elías Moreno and F. Javier Girón

George Casella[*]

University of Florida

1 Introduction

Professors Moreno and Girón are to be congratulated on their assessment of the current state of hypothesis testing as viewed from the two different approaches of the Bayesians and the frequentists. Indeed, the divergence of views has been long documented, written about, and fought over. It has always been the view of this author that, in the realm of hypothesis testing, the advantage always went to the frequentist. Not because the approach is inherently better, but rather because the frequentist approach was better for hypothesis testing.

More recently, I am less convinced of this, but in my mind the advantages are still problem specific. Although I believe that the frequentist's approach still has the overall advantage, there are now approaches, especially those based on the intrinsic methodology, that yield Bayesian answers that are quite satisfactory. In particular, the methodology championed by Moreno and co-authors, where the alternative prior is centered on the null, has yielded a big improvement in the performance of Bayesian tests.

2 Reconcilability

Moreno and Girón say there is "strong discrepancy" between frequentists and Bayesian solutions, but I have long believed that this discrepancy was mostly due to poor choices of Bayesian priors in the testing of point null hypotheses. This belief led to the results in Casella and Berger (1987), who showed that in one-sided testing problems it is possible to exactly reconcile *p*-values and posterior probabilities, "doing what comes naturally" according to DeGroot (1973). Thus, I am still convinced that any *vast* discrepancy between a good frequentist solution and an "objective" Bayes solution will be the

[*] Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611-8545. Supported by National Science Foundation Grant DMS-04-05543. Email: casella@stat.ufl.edu

result of an anomaly – such as point mass null priors – rather than a true divergence of philosophies.

2.1 Should we be reconciled?

Yes – because if we are reconciled then we are more sure of a good solution. If two reasonable approaches, reasonable but different, yield the same, or similar answers, then we can be more assured that we have arrived at a good solution. But note that reconciliation does not mean exact agreement – no two statistical procedures will agree exactly in all circumstances. Reconciliation means that the two procedures agree in a qualitative sense – they may disagree on borderline cases but they will agree in principle.

2.2 Are we reconciled?

There are a number of cases where reconciliation is problematic – see Section 3, but as an illustration that is straightforward we will concentrate on the example of Section 5 of Moreno and Girón, on testing in linear regression. Although there are some discrepancies between the intrinsic posterior probability and the p-value, it seems that Moreno and Girón have shown as much reconcilability as irreconcilability – it all depends how you interpret the pictures.

As a contrast to their Figures 1 and 2, here we plot the intrinsic posterior probability and the p-value against $\mathcal{B}_n$. This is to show that for most of the usual range of $\mathcal{B}_n$, the discrepancy is not that big. Moreover, all of the evaluations of Moreno and Girón are done under the null hypothesis, so we thought that looking at a power calculation would be interesting. Our results are summarized in Figures 1 and 2.

In Figure 1, for $n = 20$ and $k_1 = 5$, we show behaviour of both the intrinsic posterior probability and the p-value under the null hypothesis (left panel) and under the alternative (right panel). If we are going to compare the intrinsic posterior probability to .5 and the p-value to .05 then, qualitatively, looking at the left panel, there is only a small range of values of $\mathcal{B}_n$ where we would make different conclusions. For the right panel, we calculated the expected value of $\mathcal{B}_n$ for each value of the noncentrality parameter, and then evaluated both the intrinsic posterior probability and the p-value at that expected value. Again, there is only a small range of values for which the conclusions would be qualitatively different. The same calculations were done in Figure 2, where $n = 20$ and $k_1 = 9$.

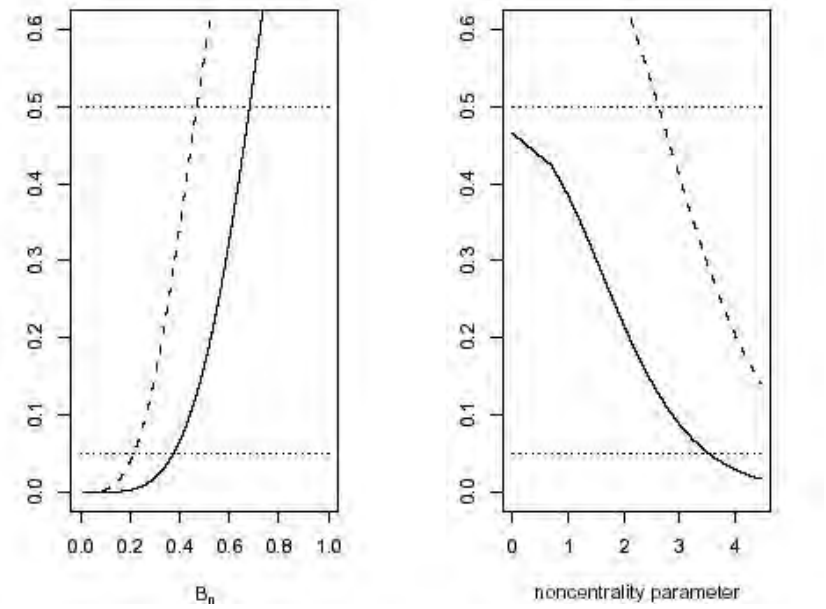

Figure 1: *For $n = 20$, $k_1 = 5$, values of intrinsic posterior probability (dashed) and p-values (solid). The left panel shows them as a function of the observed statistic $\mathcal{B}_n$. The right panel shows them as a function of the noncentrality parameter of the alternative*

It is clear that the power of the procedures increases with k_1, the size of the submodel. That is, both the intrinsic posterior probability and the *p*-value have more power to detect larger submodels ($k_1 = 9$) than smaller submodels ($k_1 = 5$). Again, the discrepancies occur on the middle values for a small range of $\mathcal{B}_n$ or the noncentrality parameter. We do not think that there is much to get excited over here – both procedures are behaving in a similar fashion.

3 Other Remarks

In a problem like that of Section 5 – testing for a submodel in linear regression –, we argue that both the intrinsic posterior probability and the *p*-value give essentially the same answer. Yes, there are discrepancies, and yes, there will be data which will lead to different conclusions. But for the most part the procedures are commensurate. This is not just good, it is very good, for it shows that in "regular" problems we have two sensible approaches that lead to similar answers. Perhaps this is telling us that we have got it right!

There are other examples discussed by Moreno and Girón where things do not work out so well, for they are examples for which the frequentist answer is not quite

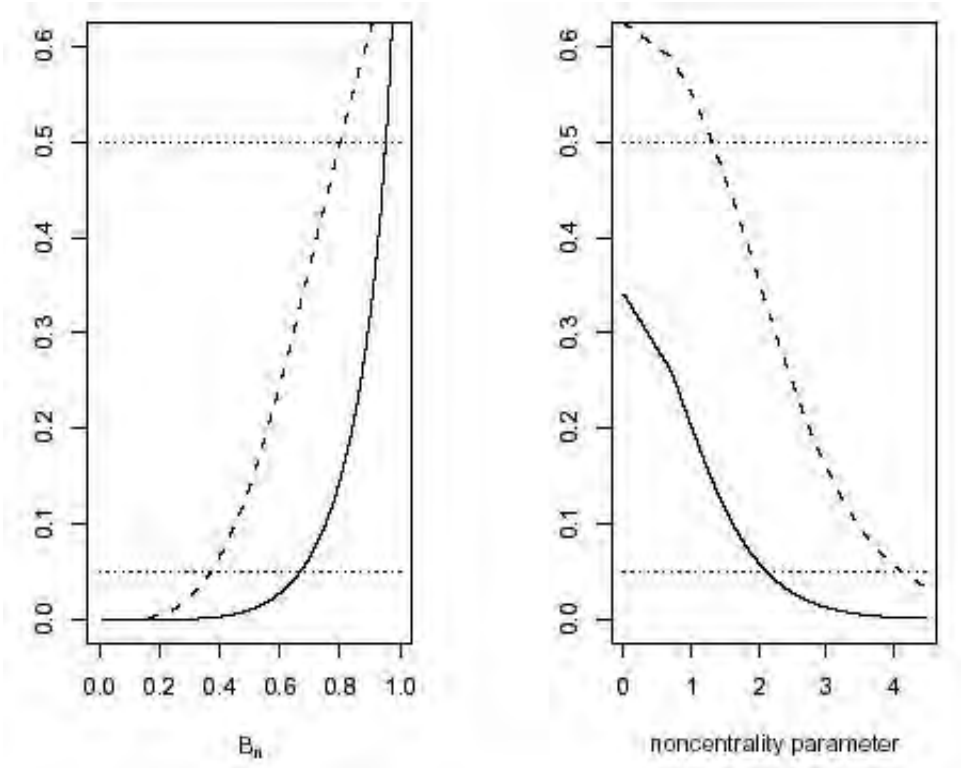

Figure 2: *For $n = 20$, $k_1 = 9$, values of intrinsic posterior probability (dashed) and p-values (solid). The left panel shows them as a function of the observed statistic $\mathcal{B}_n$. The right panel shows them as a function of the noncentrality parameter of the alternative*

satisfactory. In the beginning it was mentioned that our feeling is that the frequentist approach is better suited for hypothesis tests. While generally true, the big exception is when there are nuisance parameters. This is the case where the Bayesian shines and the frequentist sputters.

In the Behrens-Fisher problem frequentist testing suffers from the nuisance parameter in the null, and the problem of defining a *p*-value in such a circumstance. One of the best frequentist solutions to this problem is given by Berger and Boos (1994), who show how to compute legitimate *p*-values in the face of null nuisance parameters. It would be interesting to compare an intrinsic posterior probability to the Berger-Boos *p*-value. Also, where there are multiple hypotheses and multiple dimensions, the frequentist has trouble defining a coherent set of *p*-values. The approach of Goutis *et al.* (1996) attempts to define a valid way for the frequentist to approach this problem, and may provide a good comparison to an objective Bayesian approach.

4 References

Berger, R. L. and Boos, D. D. (1994). *P*-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89, 1012-1016.

DeGroot, M. H. (1973). Doing What comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, 68, 966-969.

Goutis, C., Casella, G. and Wells, M. T. (1996). Assessing evidence in multiple hypotheses. *Journal of the American Statistical Association*, 91, 1268-1277.

Daniel Peña

Universidad Carlos III de Madrid

This is an excellent paper on a fundamental topic in statistics and I want to congratulate the authors for this fine contribution. The paper illustrates very well the state of the art, is well written and clarifies many topics which have been discussed in the literature. Also, it deals with a problem to which Spanish statisticians have made important contributions, as can be seen in the list of references, which can be supplemented by Bernardo and Rueda (2002), De la Horra and Rodríguez-Bernal (2001) and Gomez-Villegas Maín and Sanz, L. (2002), among others.

Hypothesis testing has been a controversial issue in statistics since its origins with the bitter discussions between Fisher and Neyman and E. Pearson. Thus, it is not surprising that this problem concentrates the main sources of controversy between the frequentist and the Bayesian approach. Also, the relative importance of estimation versus testing has been the subject of hot debates in statistics during the 20th century. I believe that some of this controversy may be due to the use of the same word, testing, for solving two different kind of problems in statistical inference. The first one is model selection: we have a set of models, possibly of different parameter dimensions, and we want to select the one which seems most in agreement with the observed data. This problem is different from estimation, which is concerned with finding parameter values which fit the data as well as possible: if we have two models and the smaller one is nested in the larger model we cannot rely on estimation principles because the larger model will always fit better than the smaller one. Thus we need to modify the estimation principles to solve this important problem. However, hypothesis testing is also applied to decide about the value of a parameter in a given distribution. That is, we just have a model, for instance normal with mean μ and variance 1, and the problem is to decide about the value of μ. This second class of problems are often better solved by estimating the parameter (either by maximum likelihood or computing the posterior distribution if we have prior information) and it is not so clear that new principles are required to provide a satisfactory answer to the problem. Of course, if a decision is needed to be made the principles of decision analysis can be used, but still the problem can be decomposed as first estimation and then using this estimation to make a decision taking into account the consequences and the utiliy function of the decision maker. It may be argued that sometimes a particular value of one parameter may imply a different model as then we obtain a distribution which is a particular case of the first, as for instance in the power

exponential family of Diananda, see Box and Tiao (1973), but still the decision is about the value of a parameter which can be estimated by standard principles.

The problem that the authors of this interesting paper have in mind is model selection, as indicated in the first sentence of the Abstract. That is, following the notation of the paper, we have a random sample $x = (x_1, ..., x_n)$ generated by one of a set of models, $P_i(x)$,with probability densities $f_i(x)$, for $i = 0, ..., I$, and we want to select the model which is more in agreement, in a way to be defined, with the data. Then the authors present the frequentist (likelihood ratio test) approach and the Bayesian approach for this problem. As indicated in the paper the key differences between the frequentist and the Bayesian approach for testing are (1) the inclusion of prior information, and (2) the use of extreme values of the likelihood in the frequentist approach versus the use of the average values in the Bayesian approach. Although both points are important, I would like to emphasize a little bit more the key role of the second one.

As indicated in Section 3 in the paper, in the frequentist approach we use the maximum value of the likelihood in the parameter space to define both the likelihood of the null and the alternative hypothesis, whereas in the Bayesian approach we use the average value of the likelihood, where the average is computed by using the prior distribution. As using the maximum is very different from using the average it is not surprising that this choice can lead to huge differences in the conclusions. For me this is the easiest way to understand the well known Lindley-Jeffreys paradox (see Shafer, 1982 and the discussion) that I will illustrate in a simple example. Suppose that we have a random sample of size n from a N(μ, 1) distribution and we want to test $\mu = 0$ versus $\mu \neq 0$, and the mean of the sample is $\overline{x} = 1$. Then, in the frequentist approach, we compute the likelihood of the null, the maximum likelihood under the alternative, that is of course for $\widehat{\mu} = \overline{x} = 1$, and the ratio is equivalent to the t-statistic $t = \sqrt{n}$. Therefore the null is rejected even for small sample sizes and there is an overwhelming evidence against the null for medium size n, as for instance $n = 36$, which will lead to a very small p-value. From the Bayesian approach, assuming equal prior probabilities for the hypothesis, we compare the likelihood of the null to the average likelihood under the alternative, and by using a prior for μ N(0,τ^2), this Bayes factor is given by

$$\frac{P(x|H_0)}{P(x|H_1)} = \frac{e^{-n/2}\sqrt{1+n\tau^2}}{e^{-n/2(1/(1+n\tau^2))}} = B_{01}$$

and the posterior probabilities are $P(H_0|x) = B_{01}/(1 + B_{01})$ and $P(H_1|x) = 1/(1 + B_{01})$. Thus if $\tau^2 \to \infty$ then $B_{01} \to \infty$ and therefore $P(x|H_0) \to 1$. We obtain the surprising result that we may have an overwhelming evidence for rejecting the null from the frequentist point of view and at the same time a posterior probability for the null close to one from the Bayesian approach. For instance, if $n = 100$ and we assume a standard deviation for the prior $\tau = 100$, both approaches coincide and the $t = 10$ is in agreement

with a posterior probability of the null close to zero (10^{-19}). However, if we increase the variance of the prior and assume $\tau = 10^{30}$ then $P(x|H_0) \approx 1$. Thus, the average value of the likelihood under the alternative can be made as small as we wish by computing the average in a large region. This unexpected strong effect of the prior on hypothesis testing does not appear in estimation where the posterior mean will be close to the sample mean for any large value of the variance of the prior. From my point of view this analysis shows the importance of indicating correctly in Bayesian statistics what we really believe and the risk of a naive use of reference distributions or non-informative priors in Bayesian hypothesis testing.

The extreme versus average difference of both approaches also appear in all the problems of model selection and, in particular, when we compare models of different dimensions. The intrinsic Bayes factors is one of the tools introduced to compare Bayes factors in this situation and several authors have shown that it seems to work very well in selecting linear models. In particular, the result that Bayesian criteria work better than frequentists criteria in linear models have been also found in Guttman, Peña and Redondas (2005). These authors also showed that the BIC criterion seems to provide a fast, simple and effective way for computing Bayes factors which can be almost as good as the intrinsic Bayes factor. It would be interesting to know if this property, which can introduce a large simplification in many application of Bayesian model selection and Bayesian model averaging, is also found in other problems, as those discussed by the authors in this paper.

I have found the results presented in Section 5 very interesting and, in particular, the main conclusion obtained in Section 5.4 seems to me very intriguing: for models with small or medium number of relevant variables, why does the from below criterion perform better than the from above one? I would appreciate if the authors could give us some more intuition about the performance of the two procedures in different problems.

Finally, I would like to know the opinion of the authors about how to solve problems of model selection when the number of variables is as large, and even larger, than the number of observations, as often happen in image and microarrays analysis. In these problems the standard tools of statistical inference, either Bayesian or frequentist, need to be modified: for example, we cannot compute intrinsic Bayes factors in the usual way. Also, the standard criteria such as BIC need not to be consistent any more. I would be grateful to the authors for any comments on the extension of their ideas for this kind of problem.

In closing, I want to thank the authors for this fine and stimulating piece of research that I have enjoyed very much.

References

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70, 351-372.

Box, G.E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley.

De la Horra, J. and Rodríguez-Bernal, M. T. (2001). Posterior predictive p-values: What they are and what they are not. *Test*, 10, 75-86.

Gómez-Villegas M. A., Maín, P. and Sanz, L. (2002). A suitable Bayesian approach in testing point null hypothesis: Some examples, *Communications in Statistics A*, 31, 201-217.

Guttman, I, Peña, D. and Redondas, D. (2005). A Bayesian approach for predicting with polynomial regression of unknown degree, *Technometrics*, 47, 23-33.

Shafer, G. (1982). Lindley's paradox, *Journal of American Statistical Association*, 77, 325- 334.

C. P. Robert*

1 Warning

While the authors have made a great job of exposing the advantages of using Bayes factors for hypothesis testing (compared with classical solutions like UMP tests or *p*-values) and should be congratulated for a new review paper on the objective Bayes approach to testing, let me first take the oratory precaution to state that I will play here devil's advocate by arguing that objective Bayesian theory has not yet reached a satisfactory position with regards to hypothesis testing. (Obviously, I do not consider the *p*-values as valid answers either, but I will not discuss their shortcomings here since they are already discussed in Robert, 2001, Chapter 6.)

2 Bayes solutions

Even though the authors mention the standard Bayes solution against a classical $0-1$ loss at the beginning of their paper, they evacuate quite forcibly this solution in favour of alternative albeit less well-defined procedures. This is quite unfortunate in that this is the only Bayesian solution to the testing problem if one wants to come up with a "yes/no" answer. (Whether or not this is a good decisional setup is another story, as discussed below, but this is often the type of answers required from statisticians!) Indeed, a Bayes factor, a posterior probability or even a *p*-value may be used in a decision process but they are intrinsically *not* decision-theoretic procedures. This is also why I regret the early dismissal of the likelihood ratio: when both hypotheses are simple, the likelihood ratio (or rather the comparison of the likelihood ratio with a certain bound c) is a Bayes rule, no matter the value of c. When at least one hypothesis is composite, the likelihood ratio is no longer Bayes, but if we follow the asymptotic arguments of the authors, it can be taken as a first order approximation in some situations.

* Discussion written during a visit to the Department of Statistics, University of Florida, Gainesville. The author is grateful to the organisers of the 8th Winter Statistics Conference and in particular to George Casella and to Jim Hobert for their hospitality. CEREMADE, Université Paris Dauphine and CREST, INSEE, Paris. xian@ceremade.dauphine.fr

Why, thus, is there a problem with the Bayes solution, especially when considering it is *consistent*? The main point is that consistency is a very weak criterion, because a Bayes rule can choose to accept the null hypothesis when the likelihood ratio (or the ratio of marginals) is above about any bound, depending on the prior model. Is this bad or wrong? Formally speaking, this is completely acceptable as the range of Bayes rules usually covers the range of (admissible) possible answers. This only starts looking bad when one seeks a universal or common answer or, in other words, an "objective" procedure. Always from a Bayesian point of view, this universal perspective can be disputed as non-Bayesian, because of excluding all prior inputs.

I am afraid that it may be the case that the inconsistencies discussed below are simply beacons that signal the impossibility of Bayesian non-informative or objective procedures. (The fact that these procedures exist for estimation problems simply is a reflection on the smoother topological structure of estimation setups.) Or, more radically, it may be argued that the whole approach to testing as a $\{0, 1\}$ problem is flawed in that the true consequences of a decision are not properly modeled. (Hence the common confusion between scientific hypothesis testing – as in *is the age of the Universe larger than* 13 *billion years?* – and model choice – as in *how close to the family of probit models is the true model?*)

It indeed seems to me that model choice calls for a completely different decisional approach that integrates both the complexity of the models under comparison and the consequences of choosing one model rather than another. Since the authors are not adopting this perspective in variable selection, they have to resort to cascades of Bayesian tests without properly evaluating the consequences of this repetition of tests. There is, for instance, no uncertainty evaluation on the ranking and the selection of the most likely model. Nor is the sequence of decisions evaluated sequentially or conditionally. Model choice is in fact much more an estimation problem for the difference between the true model and a hypothetical collection of models than a testing model. Methods based on function divergences (Robert, 2001, Chapter 7) should thus be preferred as they ascertain the different consequences of picking the "wrong" model, even though complexity summaries like AIC, BIC or DIC are standing far away from the Bayesian paradigm. I tend to agree with the authors that a Bayesian approach is more likely to account automatically for the complexity of a model than frequentist and likelihood perspectives but I also think that a more thorough assessment of the consequences of model choice should be undertaken, rather than trusting blindly a dimension-free Bayes factor. In this regard, the recent paper by Young (2005) is quite illuminating in that it exposes the conflict between model selection and parameter estimation, establishing in particular that model averaging (Raftery, 1996) cannot reach optimal convergence rates.

Note at last that under a completely different decision-theoretic perspective, namely when losses of the type $(\delta - \mathbb{I}_{H_0}(\theta))^2$ are used, the posterior probabilities are themselves

Bayes rules (Robert, 2001) and that there exist cases where *p*-values are admissible as truncations of Bayes rules (Hwang *et al.*, 1992). Although these are rather formal results, I think they are still worth mentioning.

3 Inconsistencies

If we now turn to the alternative solutions provided in the paper, there is a lot to be said against Bayes factors, pseudo-Bayes factors and intrinsic priors.

First, as stated above, the Bayes factor is not even on the same scale as the Bayes solution given that it is dimension free. The authors often switch back and forth between Bayes factors and posterior probabilities. But doing this implies the choice of a particular prior ratio $\pi(H_0)/\pi(H_1) = 1$, a choice that is never discussed in the paper, while being paramount for the comparison with the *p*-values. In particular, there is no clear reason why $\pi(H_0) = 1/2$ should be considered as a "non-informative" solution (Robert, 1993). In some instances, intrinsic priors are associated with unbalanced prior weights $\pi(H_0)$ and $\pi(H_1)$ for example. If the complexity of the model under hypothesis H_0 is much higher than under hypothesis H_1, the prior weight of H_0 could be lowered as a consequence of Occam's razor rule. (A personal aside: I never really understood the need to call for this rule. In fact, while being interesting from an epistemological point of view, Occam's key sentence *Pluralitas non est ponenda sine neccesitate* does not constitute an operational principle and about anything can be justified on this vague sentence.)

If we now turn to pseudo-Bayes factors, they seem to cumulate the shortcomings of Bayes factors and of pseudos! While providing a workable solution to the impossibility of using improper σ-finite measures under both hypotheses, they are suffering from a high level of adhocquery that is reflected by the myriad of versions found in the literature (as discussed in Robert, 2001, Chapter 6). Pseudo-Bayes are clever mathematical constructs but they do not enjoy the same justifications as true Bayes factors. While I do not think that the (minor) criticism by the authors that "the arithmetic intrinsic Bayes factor (...) reuses the sample observations" is particularly true [many genuine Bayes procedures appear as weighted sums of averages on part of the data, think for instance of mixtures of distributions], both the lack of symmetry between H_0 and H_1 and the possible difficulty in defining acceptable subsamples and minimal sample sizes maintain the pseudo-Bayes factors at a considerable distance from genuine Bayesian inference.

The construction of the intrinsic prior proposed in the review is clever and reminiscent of Pérez and Berger (2002). Note that this prior can also be written as

$$\pi_1^I(\theta_1) = \int \pi^I(\theta_1|\theta_0)\pi_0^N(\mathrm{d}\theta_0) = \int \mathbb{E}_{\theta_0}[\pi_1^N(\theta_1|X)]\pi_0^N(\mathrm{d}\theta_0)\,,$$

a representation which somehow is more intuitive, apparently works for non-nested models, but also exposes the limitation of the device. Indeed, once θ_0 is integrated out in the above equation, we are faced with a new improper prior and the mathematical result that

$$B_{10}^I(\mathbf{x}) = \frac{\int f(\mathbf{x}|\theta_1)\pi^I(\mathrm{d}\theta_1|\theta_0)\pi_0^N(\mathrm{d}\theta_0)}{\int f(\mathbf{x}|\theta_0)\pi_0^N(\mathrm{d}\theta_0)}$$

does not depend on the normalising constant of π_0^N does not translate so easily to the ratio

$$B_{10}^I(\mathbf{x}) = \frac{\int f(\mathbf{x}|\theta_1)\pi_1^I(\mathrm{d}\theta_1)}{\int f(\mathbf{x}|\theta_0)\pi_0^N(\mathrm{d}\theta_0)}\,,$$

although they are mathematically the same. In other words, were we given π_1^I and π_0^N separately as in a regular improper prior hypothesis testing we would not know how to normalise both priors. (This is a dilemma common to missing data problems: while the density of the observables can be written as a marginal of another distribution, the dummy variables used in the marginalisation have no specific meaning.)

We also note that the proposed solution is not symmetric in (H_0, H_1), which is a drawback shared by most pseudo-Bayes procedures. In addition, when more than two hypotheses are in competition, this approach requires either a change of priors for each pair of hypotheses or the subjective choice of a *reference* hypothesis H_0 under which all other intrinsic priors are constructed.

The authors mention several times that having no moment is a "nice property to be expected from an objective prior". This is a rather peculiar remark since the lack of moments must depend on the parameterisation of the model: if we use a bounded parameterisation, the corresponding parameter will have finite moments. To continue about puzzling remarks, I do not understand the point about the one-to-one relationship between *p*-values and posterior probabilities: both quantities depend on the same (multinomial) sufficient statistic but since there is no one-to-one relationship between the *p*-value and the sufficient statistic, this does not seem possible.

4 Conclusions and perspectives

This paper provides a (partial) summary of the large literature on the comparison between frequentist (restricted to *p*-values) and Bayesian (restricted to Bayes factors), but it may constitute too restricted a vision of the challenges met by both approaches in both hypothesis testing and model comparison. To find that *p*-values do not provide the same numerical answer than posterior probabilities or Bayes factors is not a fundamental difficulty in that they are not to be treated as similar objects.

On the one hand, I definitely acknowledge the urgent need for objective Bayes procedures in testing problems and do not want to disparage the past work led by the authors and others on this topic. On the other hand, there currently exist much more pressing challenges in hypothesis testing, for instance with the emergence of massively multiple tests in Genomics (Genovese and Wasserman, 2002) and in other fields. Even in the case of model selection, the complexity (or the combinatorics) of the decision space often prevents a perfect exploration and new tools are necessary to ascertain whether or not important models and situations are not forgotten. About ten years ago (Madigan and Raftery, 1994), *model averaging* appeared as a potentially rich tool for handling multiple models simultaneously. While this is not a panacea, in that it does not directly allow for pruning in a large collection of models and may lead to sub-efficiencies (Yong, 2005), this direction should neither be abandoned altogether.

5 References

Genovese, C. and Wasserman, L. (2002). Bayesian and frequentist multiple testing. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 7*, 145-161. Oxford University Press.

Hwang, C., Casella, G., Robert, C., Wells, M., and Farrel, R. (1992). Estimation of accuracy in testing. *The Annals of Statistics*, 20, 490-509.

Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89, 1535-1546.

Pérez, J. M. and Berger, J. (2002). Expected posterior prior distributions for model selection. *Biometrika*, 89, 491-512.

Raftery, A. (1996). Hypothesis testing and model selection. In Gilks, W., Spiegelhalter, D., and Richardson, S., editors, *Markov Chain Monte Carlo in Practice*, pages 163-188. Chapman and Hall, New York, London.

Robert, C. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica*, 3, 601-608.

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, second edition.

Yong, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika*, 92(4), 937-950.

Rejoinder

It is not easy to have together three statisticians of the scientific stature of George Casella, Christian Robert and Daniel Peña discussing on the foundations of an important branch of Statistics, as is hypothesis testing and model selection, which have been prompted by our paper. We feel really very lucky in having these people to comment on the paper, so that we want to take this opportunity to thank the discussants and the Editor of SORT, Carles Cuadras, for making it possible.

Each of the discussants has put the emphasis across different points. While George Casella focuses on the context and magnitude of the disagreement between the frequentist and Bayesian approaches, Christian Robert points out that objective Bayesian analysis is still under developement and that $0-1$ loss functions and intrinsic priors are an option which could be replaced with Kullback-Leibler pseudo-distances and reference priors. Finally Daniel Peña raises the important issue of the consequences of eliminating the unknown parameters via *maximizing* in frequentist testing or via *integration with respect to a prior* in the Bayesian approach.

The discussions are wonderful, and they have made precise many aspects of the controversy that we should not forget about. Again, we are very grateful to the disccussants for the very good job they have done.

We want to answer each of them on their main points.

Response to George Casella

We thank Professor Casella for his thoughtful comments on the paper. With his usual clarity, he brings us the following general message: the frequentist and Bayesian procedures for testing hypothesis essentially agree, but there are some exceptions; namely, when there are nuisance parameters, for multiple hypotheses and multiple dimensions. For these cases the frequentist has trouble defining a coherent set of *p*-values. Of course, we agree with this diagnostic.

We also agree that the divergence of views has been long documented and that to reconcile both approaches strengthens the conclusions. Many works in the literature have been devoted to show models, priors and loss functions for which the Bayesian procedure matches the well established frequentist procedure. Books by Lindley (1970), Box and Tiao (1973) and Datta and Mukerjee (2004) are good examples of that. These sort of analyses are useful as they inform us how the frequentist solution looks like from the Bayesian viewpoint.

However, this is not the sort of reconciliation we have in mind. The point is to know whether good objective Bayesian procedures give similar answers to those of the well established frequentist procedures. Some early examples in this direction were derived by using Bayesian robustness ideas (Berger and Sellke 1987, Casella and Berger 1987)

Thus, for example, Casella completes our analysis displayed in Figures 1 and 2 illustrating how are the disagreement sampling regions between the frequentist and the intrinsic Bayesian test for normal linear models. One interesting conclusion he draws appears to be that both procedures have more power to detect large submodels than smaller submodels, and that the disagreement sampling regions are not that big. Although the disagreement regions are small we remark that they vary as the ancillaries n, k_1 and k change.

Furthermore, a p-value may correspond to a small intrinsic null posterior probability for some values of the ancillaries and to large null posterior probabilities for other values. In other words, a p-value must be interpreted regarding the values of the three ancillaries. In practice, this really is a very hard task. Many statisticians have avoided this difficulty and tried to give some magic and simple guidelines to interpret the p-values, more as a warning, when being small, that something might be wrong than as a measure of evidence in favour or against the null hipothesis.

As a measure of evidence the p-values use to be quite extreme: it may tend to one even when the evidence in favour of the null is just mild. This extreme behaviour is not shared by the intrinsic null posterior probability. To illustrate these assertions let us consider an example from Casella and Moreno (2006), Appendix C. It is related to testing independence in contingency tables. The notation $\{a, b, c; d, e, f\}$ denotes a 2×3 table with first row $\{a, b, c\}$ and second row $\{d, e, f\}$. The table total is fixed and hence the underlying sampling model is multinomial.

Table 1 shows two artificial contingency tables, with their corresponding p-values (derived from an exact test), and intrinsic null posterior probabilities.

Table 1: *Tests for independence*

Table	p-values	Post. Prob. (Intrinsic)
{2,2,2;2,2,2}	1.00	.702
{6,6,6;6,6,6}	1.00	.888

Both measures of evidence favour independence. The p-values are absolutely conclusive indicating that independence is present in both tables without any uncertainty. However, the intrinsic null posterior probabilities indicate that independence is present in both tables but there is less uncertainty in asserting it for the second table than for the first one, which is a more sensible conclusion.

When testing in the presence of nuisance paramters as in section 4, the Berger-Boos‘s p-value and the approximate p-value are quite close to each other so that the conclusions are essentially the same.

It seems pertinent to bring here what George Casella said to Elías Moreno in a relaxed conversation "*I feel very comfortable with the Bayesian approach for statistical inference and with the frequentist approach for evaluating the procedure*".

Response to Chris Robert

We are grateful to Professor Chris Robert for his thoughtful comments. In his role of devil's advocate or promoter of the faith – Bayesian faith not frequentist, we presume – Chris Robert critically examines the facts and comments dealt with in the paper and, therefore, includes in his discussion everything that is unfavourable to the paper. But our paper is neither intended for beatification nor canonization, not even amongst the Bayesian community; it is but a modest contribution or addition to the frequentist-Bayesian controversy, and it is not intended as a general review paper.

At first sight we thought we mostly disagreed with him but after a second reading of his discussion we realized that it was not so and, in fact, we agree with many of his comments. We will first present the ideas we share with him and after those on which we disagree; we anticipate that most of the disagreement is found in his Section 3 entitled "inconsistencies" which, we feel, seems disproportionate in some aspects.

We recall to the reader that we have adopted the restricted setting where several parametric sampling models are in competition. Therefore, the so-called open perspective (Bernardo and Smith 1994), or the equivalent case where the sampling distribution is itself uncertain (Robert 2001), has not been considered. The reason is that we believe it is not a well formulated problem and hence no sensible solution for it can be obtained.

We agree with Chris Robert that objective Bayesian theory for statistical inference (not only for testing) is still under development, but we are on the way. A way that might look shocking for those who think we should have a universal objective prior for the model parameters, since the main tool for estimation, the reference prior, depends on the quantity of interest (Berger and Bernardo 1992), and the intrinsic priors, our main tool for testing, also depends on the null and the alternative hypotheses.

We also agree with Chris Robert that model choice is a decision problem on a space of models $\{M_i, i = 1, \ldots, k\}$ that calls for a loss function $L(d_i, M_j)$ which should take into account not only the loss we incur when the decision d_i is made and the true model is M_j but also the complexity of the model. Assuming that this can be done, the formal

solution to this decision problem is simply an ordering of the models according to the posterior risk values

$$R(d_i|\mathbf{x}) = \sum_{j=1}^{k} L(d_i, M_j)P(M_j|\mathbf{x}),\ i = 1, \ldots, k,$$

or, equivalently, according to the proportional quantities

$$\sum_{j=1}^{k} L(d_i, M_j)m_j(\mathbf{x})P(M_j),\ i = 1, \ldots, k,$$

where $m_j(\mathbf{x}) = \int f(\mathbf{x}|\theta_j)\pi_j(d\theta_j)$ is the marginal of the data under model M_j. Therefore, a loss function, a prior for the model parameters and a prior for the models are all we need to reach the solution to the decision problem.

Of course, the ordering of the models according to their model posterior probabilities does not necessarily coincide with the ordering we obtain from the posterior risk except when the loss function is

$$L(d_i, M_j) = \begin{cases} 1, \text{ if } i \neq j, \\ 0, \text{ if } i = j. \end{cases}$$

In this case all we need is a good search technique (stochastic search for most of real problems) for finding models with large posterior probability. As a consequence, we cannot understand Chris Robert when he says at the beginning of Section 2 that we mention the standard Bayesian solution against classical $0-1$ loss but we "evacuate quite forcibly this solution in favor of alternative albeit less well-defined procedures".

This rather natural formulation has been forced in his view of the problem to using the Kullback-Leibler pseudo-distance as a loss function instead, maybe because of the difficulties that might be encountered in assessing $L(d_i, M_j)$. For instance, when $k = 2$ and the null model is $\theta_0 \in \Theta_0 \subset \Theta$, the loss function is formulated as the pseudo-distance between the sampling distribution at a point $\theta \in \Theta$ and the sampling distribution at a null point θ_0 (Goutis and Robert 1998; Bernardo and Rueda 2002). The null point θ_0 is ruled out by taking the infimum over the null space Θ_0 and then θ is integrated out with respect to the posterior reference prior which, in passing, is the only instance the data enter for the solution of the problem. It is hard to understand the different ways this procedure eliminates the parameters of the null and the full model. Once the unknown parameters have been eliminated the decision rule is: a large value of quantity thus computed from the pseudo-distance rejects the null. Thus, this approach needs to additionally specify what is the meaning of "large values computed from the pseudo-distance". Therefore, under this approach the decision problem is not fully determined

by priors plus losses, which is not sensible. Further, we wonder how the data enter into the problem of comparing the simplest hypothesis testing setting $H_0 : f(x)$ *versus* $H_1 : g(x)$, where there are no parameters to play the nonsymmetric role of the infimum and integration operations.

We agree with the first part of Section 3. We might not be impartial by taking the reference prior $\pi(H_0) = \pi(H_1) = 1/2$. An attempt to be impartial is that of Robert and Caron (1996), but unfortunately it does depend on the nuisance parameters so that for the moment the reference prior seems to be the only option we have.

In the second part of Section 3, the Bayes factor for intrinsic priors $(\pi^N(\theta_0), \pi^I(\theta_1))$ is found to be "inconsistent", in the sense of being ill-defined, we presume. We disagree. We have not claimed that the intrinsic priors can be normalized. What we claim is that the Bayes factor for intrinsic priors

$$B_{10}^I(\mathbf{x}) = \frac{\int f_1(x|\theta_1)\pi^I(d\theta_1)}{\int f_0(x|\theta_0)\pi^N(d\theta_0)}$$

is a well-defined notion as it is a well-defined limit of a sequence of Bayes factors for proper priors (Moreno *et al.* 1998).

Furthermore, for normal linear models, as the sample size tends to infinity the intrinsic null posterior probability tends to one when sampling from the null, and to zero when sampling from the alternative (Moreno and Girón 2005).

That the conditional intrinsic prior distribution of the regression coefficients be an elliptical distribution centered at the null with no moments seems to us a nice property of an objective distribution for the regression coefficients as the parameter space R^k is unbounded. This also implies that this prior has very heavy tails, thus allowing for large departures from the null model.

We also disagree with Chris when he questions the relationship between the p-values and the null model posterior probability when testing that some regression coefficients are zero in normal linear models. We do not understand his reasoning but we assert that the sufficient statistic $\mathfrak{B}_n$, for fixed values of the ancillary statistics n, k and k_0, does not have a multinomial distribution but instead a beta distribution when sampling from the null, and a noncentral beta distribution when sampling from the alternative. Moreover, there is a one-to-one relationship between the p-value and the sufficient statistic $\mathfrak{B}_n$ since

$$p\text{-value} = \int_0^{\mathfrak{B}_n} B(z|\frac{n-k}{2}, \frac{k_0}{2})\, dz.$$

We agree that p-values and posterior probabilities do not need to provide the same numerical values. A rather different thing is that we would like to obtain the same

decisions from the use of the *p*-values and from the posterior probabilities (see the discussion by George Casella).

We are firmly convinced that *p*-values and null model posterior probabilities are nowadays the recognized tools for hypothesis testing. The message in the paper is that the Bayesian solution based on intrinsic priors is a well-justified objective Bayesian procedure that does not exhibit the well documented drawbacks of the *p*-values. Of course, loss functions different from $0-1$ can be considered but even in this case model posterior probabilities play a basic role.

Response to Daniel Peña

We thank Daniel Peña for his thoughtful comments and for calling our attention to the additional references on the topic. We agree with his lucid and in-depth perception of the differences between estimation, testing and model comparison. As he brings forward the Lindley paradox in his discussion, we feel that it is timely to briefly discuss it here, since it has been considered by many authors as the proof that the Bayesian model posterior probability approach to hypothesis testing is wrong.

In our opinion the so-called Lindley paradox is not at all a mathematical paradox (see also Robert 1993 and Moreno and Girón 2005) but simply a logical consequence of the formulation. In the standard normal example, the prior for μ is chosen to be a $N(\mu|0,\tau^2)$. When $\tau \to \infty$, the probability mass on the real line tends to 0, and hence the Bayesian machinery can not choose a model with prior probability equal to zero. It chooses instead the null model whatever the sample, which is the right thing to do. There is nothing wrong or paradoxical with this. What is really wrong is to put probability mass zero on the space where the parameter lives and to pretend that we are considering a prior distribution on that space.

What we can learn from the paradox (if any) is that the limit of proper prior distributions is not necessarily a reasonable prior for testing. For the paradoxical case, $\lim_{\tau\to\infty} N(\mu|0,\tau^2) = 0$, which is obviously not a reasonable prior.

We also agree with the positive conclusion Daniel Peña draws from the paradox of "the strong effect of the prior on hypothesis testing which does not appear to be the case for estimation". This assertion implies that any approximation to a marginal density of the data should not avoid the influence of the prior.

It can be argued that the simple BIC criterion adjusts the classical likelihood ratio to favour the reduced model and, under some constraints, it can be interpreted as a statistic that corresponds to a proper prior (O'Hagan and Foster 2004, Kass and Wasserman 1995). Therefore, we agree with Daniel Peña that BIC could be an effective alternative

to the Bayes factor for intrinsic priors when the sample size n is large enough; otherwise, there is no way to justify it, specially in high-dimensional parameter spaces or when asymptotics arguments do not apply.

In response to the question he raises about the comparison between the from below and from above criteria our comments are mainly based on empirical findings, which have been drawn from a larger scale simulation study described in Moreno and Girón (2005b). On a more theoretical scale, one advantage of the from below criteria is that the full model need not be specified in advance as the intercept only model is always nested into every model. On the other hand, the from above criteria requires the specification of the full model so that the resulting ordering of the models it renders depends on the chosen full model. Further, the from below criterion apparently applies Occam's razor more sharply than the from above one, thus tending to favour more parsimonious models.

The last question about the problem of having more variables than observations, which not only occurs in image and microarray analysis but also when considering interactions among variables in linear models, is as Daniel Peña aptly says a very important problem which calls for new statistical tools. In this respect, we want to remark that the from below intrinsic Bayes factor enjoys the following nice property not shared for any other model choice criterion as far as we know (see Theorem 2 of Moreno and Girón, 2005b).

Stated in an informal way the theorem asserts that if the number of explanatory variables k exceeds the sample size n, then the Bayes factor for comparing this model with the intercept only model is always less that 1 or, equivalently, the posterior probability of any model with more regressors than the sample size is less than that of the intercept only model, and it is smaller as the number of variables increases. This result automatically prevents the consideration of models with too many explanatory variables and, therefore, this criterion always penalizes complex models in accordance with Occam's razor principle.

This theoretical property of the from below intrinsic Bayes factor is deeply rooted in the fact that the conditional intrinsic distribution of the regressors in Section 5.2 is now a generalized multivariate normal distribution with singular covariance matrix (see, Rao 1973, pp. 518-528) where the matrix $\mathbf{W}^{-1}$ of Section 5.2 would be now replaced by

$$\mathbf{W}^{-} = \frac{n}{k+1}(\mathbf{X}^t\mathbf{X})^{-}$$

where $(\mathbf{X}^t\mathbf{X})^{-}$ denotes the generalized inverse or pseudoinverse of the $k \times k$ matrix $\mathbf{X}^t\mathbf{X}$. Using this expression, it can be proven that formula (4) for the resulting Bayes factor still holds.

The application of this result considerably reduces the number of models to be considered as only those having less that n variables are eligible. But, even with this simplification, the number of possible candidate models is still very large; thus, we need to resort to some sort of stochastic model search.

Additional References

Bernardo, J.M. and Rueda, R. (2002). Bayesian hypothesis testing: a reference approach. *International Statistical Review*, 70, 351-372.

Casella, G. and Moreno, E. (2006). Intrinsic tests of independence of contingency tables. Submitted.

Datta, G.S. and Mukerjee, R. (2004). *Probability matching priors: higher order asymptotic*. Lectures Notes in Statistics, Vol. 178, Springer-Verlag: New York.

Goutis, C. and Robert, C.P. (1998). Model choice in generalized linear models: a Bayesian approach via Kulbacki-Leibler projections, *Biometrika*, 85, 29-37.

Kass, R.E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928-934.

Rao, C.R. (1973). *Linear Satistical Inference and its Applications*. Wiley: New York.

Robert, C.P. (1993). A note on the Jeffrey-Lindley paradox. *Statist. Sinica*, **3**, 601-608.

Robert, C.P. (2001). *The Bayesian Choice*. Springer: New York (second edition).

Robert, C.P. and Caron, N. (1996). Noninformative Bayesian testing and neutral Bayes factors, *Test*, 5, 411-437.

SORT 30 (1) January-June 2006, 55-84

The importance of being the upper bound in the bivariate family*

C. M. Cuadras

University of Barcelona

Abstract

Any bivariate cdf is bounded by the Fréchet-Hoeffding lower and upper bounds. We illustrate the importance of the upper bound in several ways. Any bivariate distribution can be written in terms of this bound, which is implicit in logit analysis and the Lorenz curve, and can be used in goodness-of-fit assesment. Any random variable can be expanded in terms of some functions related to this bound. The Bayes approach in comparing two proportions can be presented as the problem of choosing a parametric prior distribution which puts mass on the null hypothesis. Accepting this hypothesis is equivalent to reaching the upper bound. We also present some parametric families making emphasis on this bound.

MSC: 60E05, 62H17, 62H20, 62F15

Keywords: Hoeffding's lemma, Fréchet-Hoeffding bounds, given marginals, diagonal expansion, logit analysis, goodness-of-fit, Lorenz curve, Bayes test in 2×2 tables.

1 Introduction

Several concepts and equations play an important role in statistical science. We prove that the bivariate upper Fréchet bound and the maximal Hoeffding correlation are two related expressions which, directly or implicitly, are quite useful in probability and statistics.

* Dedicated to the memory of Joan Augé (1919-1993).

Address for correspondence: C.M. Cuadras. Universitat de Barcelona. Diagonal, 645. 08023 Barcelona (Spain).
E-mail: ccuadras@ub.edu

Received: February 2006
Accepted: June 2006

Let X, Y be two random variables with continuous joint cumulative distribution function (cdf) $H(x,y)$ and marginal cdf's $F(x), G(y)$. Assuming finite variances, Hoeffding (1940) proved that the covariance in terms of the cdf's is given by

$$\mathrm{Cov}(X,Y) = \int_{R^2} (H(x,y) - F(x)G(y))dxdy. \tag{1}$$

Then he proved that the correlation coefficient

$$\rho_H(X,Y) = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$$

satisfies the inequality

$$\rho^- \leq \rho_H \leq \rho^+,$$

where ρ^-, ρ^+ are the correlation coefficients for the bivariate cdf's

$$H^-(x,y) = \max\{F(x) + G(y) - 1, 0\} \quad \text{and} \quad H^+(x,y) = \min\{F(x), G(y)\},$$

respectively.

In another seminal paper Fréchet (1951) proved the inequality

$$H^-(x,y) \leq H(x,y) \leq H^+(x,y), \tag{2}$$

where H^- and H^+ are the so-called lower and upper Fréchet-Hoeffding bounds. If H reaches these bounds then the following functional relations hold between the random variables:

$$\begin{array}{ll} F(X) = 1 - G(Y), & \text{(a.s.) if} \quad H = H^-, \\ F(X) = G(Y), & \text{(a.s.) if} \quad H = H^+. \end{array}$$

The distributions H^-, H^+ and $H = FG$ (stochastic independence) are examples of cdf's with marginals F, G. The construction of distributions when the marginals are given is a topic of increasing interest – see, for example, the proceedings edited by Cuadras, Fortiana and Rodriguez-Lallena (2002).

Note that H^- and H^+ are related by

$$H^+(x,y) = F(x) - H^-(x, G^{-1}(1 - G(y)),$$

and that the $p-$dimensional generalization of (2) is

$$H^-(x_1, \ldots, x_p) \leq H(x_1, \ldots, x_p) \leq H^+(x_1, \ldots, x_p),$$

where $H(x_1, \ldots, x_p)$ is a cdf with univariate marginals $F_1, \ldots, F_p$ and

$$\begin{aligned} H^-(x_1, \ldots, x_p) &= \max\{F_1(x_1) + \ldots + F_p(x_p) - (p-1), 0\}, \\ H^+(x_1, \ldots, x_p) &= \min\{F_1(x_1), \ldots, F_p(x_p)\}. \end{aligned}$$

However, if $p > 2$, in general only H^+ is a cdf, see Joe (1997). Thus we may focus our study on the Fréchet-Hoeffding upper bound.

The aim of this paper is to present some relevant aspects of H^+, which may generate any bivariate cdf and is implicit in some statistical problems.

2 Distributions in terms of upper bounds

Hoeffding's formula (1) was extended by Cuadras (2002a) as follows. Let us suppose that the ranges of X, Y are the intervals $[a, b], [c, d] \subset \overline{R}$, respectively. Thus $F(a) = G(c) = 0$, $F(b) = G(d) = 1$. Let $\alpha(x), \beta(y)$ be two real functions of bounded variation defined on $[a, b], [c, d]$, respectively. If $\alpha(a)F(a) = \beta(c)G(c) = 0$ and the covariance between $\alpha(X), \beta(Y)$ exists, it can be obtained from

$$\mathrm{Cov}(\alpha(X), \beta(Y)) = \int_a^b \int_c^d (H(x, y) - F(x)G(y)) d\alpha(x) d\beta(y). \tag{3}$$

Suppose that the measure $dH(x, y)$ is absolutely continuous with respect to $dF(x)dG(y)$ and that

$$\int_a^b \int_c^d (dH(x, y))^2 / dF(x)dG(y) < \infty.$$

Then the following diagonal expansion

$$dH(x, y) - dF(x)dG(y) = \sum_{k \geq 1} \rho_k a_k(x) b_k(y) dF(x)dG(y) \tag{4}$$

exists, where ρ_k, $a_k(X)$, $b_k(Y)$ are the canonical correlations and variables, respectively (see Hutchinson and Lai, 1991).

Let us consider the upper bounds

$$F^+(x, y) = \min\{F(x), F(y)\}, \quad G^+(x, y) = \min\{G(x), G(y)\},$$

and the symmetric kernels

$$K(s, t) = F^+(s, t) - F(s)F(t), \quad L(s, t) = G^+(s, t) - G(s)G(t).$$

Then using (3) and integrating (4), we can obtain the following expansion

$$H(x,y) = F(x)G(y) + \sum_{k\geq 1} \rho_k \int_a^b K(x,s)da_k(s) \int_c^d L(t,y)db_k(t),$$

which shows the generating power of the upper bounds (see Cuadras, 2002b, 2002c). Thus we can consider the nested family

$$H_n(x,y) = F(x)G(y) + \sum_{k=1}^{n} \rho_k \int_a^b K(x,s)da_k(s) \int_c^d L(t,y)db_k(t),$$

by taking generalized orthonormal sets of functions (a_k) and (b_k) with respect to F and G. It is worth noting that it can exist a non-countable class of canonical correlations and functions (Cuadras, 2005a).

3 Correspondence analysis on the upper bound

Correspondence analysis (CA) is a multivariate method to visualize categorical data, typically presented as a two-way contingency table $\mathbf{N}$. The distance used in the graphical display of the rows (and columns) of $\mathbf{N}$ is the so-called chi-square distance between the profiles of rows (and between the profiles of columns). This method is described in Benzécri (1973) and Greenacre (1984), and it can be interpreted as the discrete version of (4) – see also Cuadras *et al.* (2000).

Let $\mathbf{N} = (n_{ij})$ be an $I \times J$ contingency table and $\mathbf{P} = n^{-1}\mathbf{N}$ the correspondence matrix, where $n = \sum_{ij} n_{ij}$. Let $\mathbf{r} = \mathbf{P1}$, $\mathbf{D}_r$ =diag($\mathbf{r}$), $\mathbf{c} = \mathbf{P}^\top\mathbf{1}$, $\mathbf{D}_c$ =diag($\mathbf{c}$), the vectors and diagonal matrices with the marginal frequencies of $\mathbf{P}$.

CA uses the singular value decomposition

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc}^\top)\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_\sigma\mathbf{V}^\top, \tag{5}$$

where $\mathbf{D}_\sigma$ is the diagonal matrix of singular values in descending order, and $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns. To represent the I rows of $\mathbf{N}$ we may take as principal coordinates the rows of $\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\sigma$. Similarly, to represent the J columns of $\mathbf{N}$ we may use the principal coordinates contained in the rows of $\mathbf{B} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\sigma$. CA has the advantage that we can perform a joint representation of rows and columns, called the symmetric representation, as a consequence of the transition relations

$$\mathbf{A} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{B}\mathbf{D}_\sigma^{-1}, \quad \mathbf{B} = \mathbf{D}_c^{-1}\mathbf{P}^\top\mathbf{A}\mathbf{D}_\sigma^{-1}. \tag{6}$$

Let us apply CA on the upper bound. Consider the $I \times I$ triangular matrix

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

and similarly the $J \times J$ matrix $\mathbf{M}$. The cumulative joint distribution is $\mathbf{H} = \mathbf{LPM}^{\mathsf{T}}$ and the cumulative marginals are $\mathbf{R} = \mathbf{Lr}$ and $\mathbf{C} = \mathbf{Mc}$. The $I \times J$ matrix $\mathbf{H}^{+} = (h_{ij}^{+})$ with entries

$$h_{ij}^{+} = \min\{\mathbf{R}(i), \mathbf{C}(j)\}, \quad i = 1, \cdots, I, \ j = 1, \cdots, J,$$

contains the cumulative upper bound for table $\mathbf{N}$. The correspondence matrix for this bound is

$$\mathbf{P}^{+} = \mathbf{L}^{-1}\mathbf{H}^{+}(\mathbf{M}^{\mathsf{T}})^{-1}.$$

For instance, if $I = J = 2$ and $\mathbf{r} = (s, 1-s)^{\mathsf{T}}$, $\mathbf{c} = (t, 1-t)^{\mathsf{T}}$, then $\mathbf{R} = (\mathbf{s}, 1)^{\mathsf{T}}$, $\mathbf{C} = (t, 1)^{\mathsf{T}}$ and

$$\mathbf{H}^{+} = \begin{bmatrix} \min\{s,t\} & s \\ t & 1 \end{bmatrix}, \quad \mathbf{P}^{+} = \begin{bmatrix} \min\{s,t\} & s - \min\{s,t\} \\ t - \min\{s,t\} & 1 - s - t + \min\{s,t\} \end{bmatrix}.$$

For a geometric study of Fréchet-Hoeffding bounds in $I \times J$ probabilistic matrices, see Nguyen and Sampson (1985). For a probabilistic study with discrete marginals (binomial, Poisson), see Nelsen (1987).

Table 1: *Survey combining staff-groups with smoking categories (left) and upper bound correspondence matrix (right).*

	Original table Smoking category				*Upper bound* Smoking category			
Staff	(0)	(1)	(2)	(3)	(0)	(1)	(2)	(3)
SM	4	2	3	2	0.057	0	0	0
JM	4	3	7	4	0.093	0	0	0
SE	25	10	12	4	0.166	0.098	0	0
JE	18	24	33	13	0	0.135	0.321	0
SC	10	6	7	2	0	0	0	0.129

Example 1 Table 1, left, (Greenacre, 1984) reports a cross-tabulation of staff-groups (SM=Senior Managers, JM=Junior Managers, SE=Senior Employers, JE=Junior Employers, SC=Secretaries) by smoking category (none(0). light(1), medium(2), heavy(3)) for 193 members of a company. CA on Table 1, right, which contains the

relative frequency upper bound, provides Figure 1. This table is quasi-diagonal. Note the proximity of the rows to the columns, specially along the first dimension.

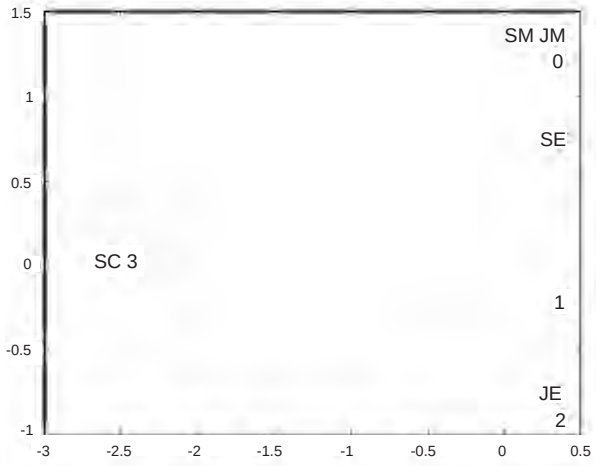

Figure 1: *Symmetric correspondence analysis representation of the upper bound in Table 1, right.*

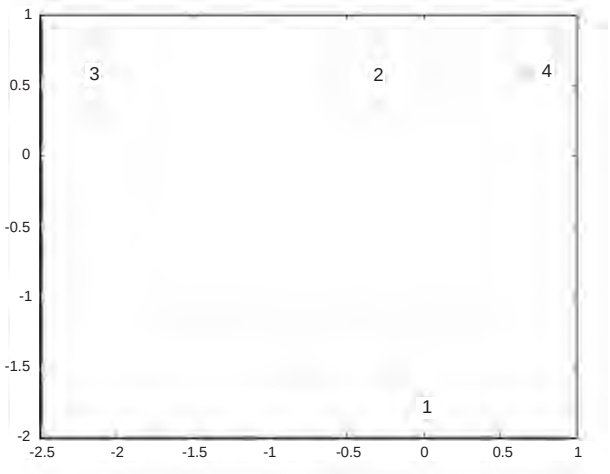

Figure 2: *Symmetric correspondence analysis representation of the upper bound in Table 2, right. Rows and columns are represented on coincident points.*

Example 2 CA is now performed on Table 2, left, an artificial 4×4 table whit the same marginals. Figure 2 exhibits the representation of the relative frequency upper bound. Now this table is diagonal. Note that rows and columns are placed on coincident points.

Table 2: *Artificial contingency table with the same margin frequencies (left) and upper bound correspondence matrix (right).*

	Original table				*Upper bound*			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
(1)	8	6	2	8	0.24	0	0	
(2)	6	0	4	10	0	0.20	0	
(3)	6	4	0	4	0	0	0.14	
(4)	4	10	8	20	0	0	0	0.42

4 Orthogonal expansions

Here we work only with a r. v. X with range $[a, b]$, continuous cdf F, and the above symmetric kernel $K(s,t) = \min\{F(s), F(t)\} - F(s)F(t)$. This kernel is the covariance of the stochastic process $\mathbf{X} = \{X_t,\ t \in [a,b]\}$, where X_t is the indicator of $[X > t]$. If the trace

$$\mathrm{tr}(K) = \int_a^b F(t)(1 - F(t))dt$$

is finite, K can be expanded as

$$K(s,t) = \sum_{k\geq 1} \lambda_k \psi_k(s)\psi_k(t),$$

where $\psi_k, \lambda_k,\ k \geq 1$, are the eigenfunctions and eigenvalues related to the integral operator defined by K. Let us consider the integrals

$$f_k(x) = \int_a^x \psi_k(t)dt.$$

Direct application of (3) shows that $(f_k(X))$ is a sequence of mutually uncorrelated random variables:

$$\mathrm{Cov}(f_i(X), f_j(X)) = \begin{cases} 0 & \text{if } i \neq j, \\ \lambda_i & \text{if } i = j. \end{cases}$$

These variables are principal components of $\mathbf{X}$ and $f_1(X)$ characterizes the distribution of X (Cuadras 2005b).

Examples of principal components $f_n(X)$ and the corresponding variances λ_n are:

1. $(\sqrt{2}/(n\pi))(1 - \cos n\pi X)$, $\lambda_n = 1/(n\pi)^2$, if X is $[0,1]$ uniform.
2. $\left[2J_0(\xi_n \exp(-X/2)) - 2J_0(\xi_n)\right]/\xi_n J_0(\xi_n)$, $\lambda_n = 4/\xi_n^2$, if X is exponential with

unit mean, where ξ_n is the $n-$th positive root of J_1 and J_0, J_1 are the Bessel functions of the first order.

3. $(n(n+1))^{-1/2}[L_n(F(X)) + (-1)^{n+1}\sqrt{2n+1}]$, $\lambda_n = 1/(n(n+1))$, if X is standard logistic, where (L_n) are the Legendre polynomials on $[0,1]$.

4. $c_n[X\sin(\xi_n/X) - \sin(\xi_n)]$, $\lambda_n = 3/\xi_n^2$, if X is Pareto with $F(x) = 1 - x^{-3}$, $x > 1$, where $c_n = 2\xi_n^{-1/2}(2\xi_n - \sin(2\xi_n))^{-1/2}$ and $\xi_n = \tan(\xi_n)$.

Assuming a finite, we can expand $\mathbf{X}$ as $X_t = \psi_1(t)f_1(X) + \psi_2(t)f_2(X) + \ldots$ and from $X_t = X_t^2$, integrating X_t on $[a,b]$ we have $X = a + \int_a^b X_t dt = a + \int_a^b X_t^2 dt$, and the variable X can be expanded in two ways

$$X = a + \sum_{k\geq 1} f_k(b)f_k(X) = a + \sum_{k\geq 1} f_k(X)^2,$$

where the convergence is in the mean-square sense. See Cuadras and Fortiana (1995), Cuadras *et al.* (2006) for other expansions, and Cuadras and Cuadras (2002) for applications in goodness-of-fit assessment. These expansions depend on a countable set of functions, again related to the upper bound.

5 Logit and probit analysis

The upper bound is implicit in some transformations. Suppose that F, the cdf of X, is unknown, whereas Y follows the logistic distribution $G(\alpha + \beta y)$, where

$$G(y) = 1/(1 + \exp(-y)), \quad -\infty < y < +\infty.$$

We may take G as a "model" for F in the sense that H, the cdf of (X,Y), attains the upper bound $H^+(x,y) = \min\{F(x), G(\alpha+\beta y)\}$. In other words, we assume the functional relation $F(X) = G(\alpha + \beta Y)$, with F unknown and G standard logistic (Figure 3). This gives rise to the logistic transformation

$$\ln\left(\frac{F(x)}{1-F(x)}\right) = \alpha + \beta y. \tag{7}$$

If the plot of $\ln[F(x)/(1-F(x))]$ against y is almost linear, then the data fit the upper bound. The probit transformation arises similarly by considering the $N(0,1)$ distribution.

In logit and probit analysis applied in bioassay, the user observes the proportion of $[X > x]$ for x fixed, i.e., observes $F(x)$ rather than X. Then (7) is used, where the parameters α, β should be estimated. Thus the outcomes arise from the above random process $\mathbf{X} = \{X_t,\ t \in [a,b]\}$. It is also worth noting that $F(X)$ is the first principal

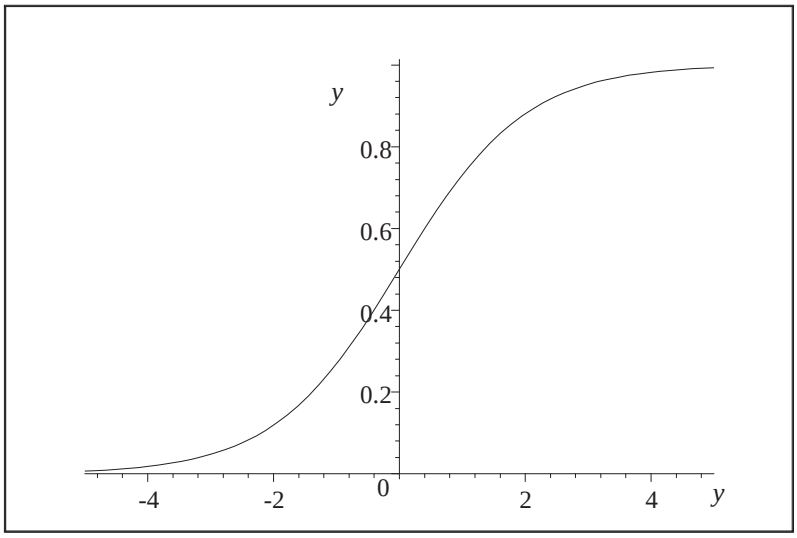

Figure 3: *The logistic curve illustrates the support of the upper bound when a distribution function is considered logistic as a model.*

component of $\mathbf{X}$ only if X is logistic, see Cuadras and Lahlou (2000). Thus logit is better than probit and both transformations can be viewed as a consequence of using the upper bound.

6 Given the regression curve

When Y is increasing in X, an ideal and quite natural relation between X and Y is $F(X) = G(Y)$. Therefore, to predict Y given X when H is unknown, a reasonable way is to use the Fréchet-Hoeffding upper bound. This gives rise to the regression curve

$$y = G^{-1} \circ F(x).$$

Of course, H^+ puts all mass concentrated on this curve.

Let us construct a cdf H_θ with this (or any in general) regression curve. If the ranges of X, Y are the intervals $[a, b], [c, d]$, and $\varphi : [a, b] \to [c, d]$ is an increasing function, the following family

$$H_\theta(x, y) = \theta F(\min\{x, \varphi^{-1}(y)\}) + (1 - \theta)F(x)J_\theta(y), \quad 0 \le \theta < \theta^+, \tag{8}$$

where θ^+ is given below, is a bivariate cdf with marginals F, G, provided that

$$J_\theta(y) = [G(y) - \theta F(\varphi^{-1}(y))]/(1 - \theta)$$

is a cdf. The regression curve is linear in φ and $H_\theta(x,y)$ has a singular part with mass on the curve $y = \varphi(x)$. It can be proved that $P[Y = \varphi(X)] = \theta$ (see Cuadras, 1992, 1996).

With $\varphi = G^{-1} \circ F$ equation (8) reduces to

$$H_\theta(x,y) = \theta F(\min\{x, F^{-1} \circ G(y)\}) + (1-\theta)F(x)G(y), \quad 0 \le \theta \le 1. \tag{9}$$

and the upper bound is attained at $\theta = 1$.

Next, let us find the covariance and prove an inequality. Let $\psi = \varphi^{-1}$, suppose that ψ' exists and X, Y have densities f, g (Lebesgue measure). Differentiation of $J_\theta(y)$ gives $g(y) - \theta f(\psi(y))\psi'(y) > 0$, hence θ is bounded by

$$\theta^+ = \inf_{y\in[c,d]} \{\frac{g(y)}{f(\psi(y))\psi'(y)}\},$$

where we write "ess inf" if necessary. From (3) we have

$$\begin{aligned} \mathrm{Cov}_{H_\theta}(X,Y) &= \int_a^b \int_c^d \theta(F(\min\{x,\psi(y)\}) - F(x)F(\psi(y))d(x)d\varphi(y) \\ &= \theta \mathrm{Cov}(X, \varphi(X)). \end{aligned} \tag{10}$$

Thus the following inequality holds:

$$\inf_{y\in[c,d]} \{\frac{g(y)}{f(\psi(y))\psi'(y)}\}\rho(X,\varphi(X)) \le \rho^+.$$

In particular, if $\varphi(x) = x$ and f, g have the same support $[a,b]$, we obtain

$$\max[\inf_{x\in[a,b]} \{\frac{f(x)}{g(x)}\}, \inf_{x\in[a,b]} \{\frac{g(x)}{f(x)}\}] \le \rho^+.$$

7 Parent distribution of a data set

Let $\chi = \{x_1, x_2, \ldots, x_N\}$ be a sample of X with unknown cdf F, and let F_N be the empirical cdf. We are interested in ascertaining the parent distribution of χ. This problem has been widely studied assuming that F belongs to a finite family of cdf's $\{F_1, \ldots, F_n\}$ (see Marshall *et al.*, 2001).

The maximum Hoeffding correlation is a good similarity measure between two cdf's. Assuming the variables standardized, it can be computed by

$$\rho^+(F_i, F_j) = \int_0^1 F_i^{-1}(u)F_j^{-1}(u)du.$$

Thus, a distance between F_i and F_j, which lies between 0 and $\sqrt{2}$, is given by

$$d_{ij} = \sqrt{2(1 - \rho^+(F_i, F_j))}.$$

We can also compute the correlation and distance between data and any theoretical cdf. Then the $(n + 1) \times (n + 1)$ matrix $D = (d_{ij})$ is a Euclidean distance matrix and we can perform a metric scaling in order to represent the set $\{F_1, \ldots, F_n, F_N\}$ using the two first principal axes (see Mardia *et al.*, 1979). The graphic display may give an indication of the underlying distribution of the sample.

There are other distances between distributions, e.g., the Kolmogorov distance

$$d(F_i, F_j) = \sup_{-\infty<x<\infty} |F_i(x) - F_j(x)|,$$

(Marshall *et al.*, 2001) and the Wasserstein distance (del Barrio *et al.*, 2000)

$$\mathcal{W}_{ij} = \int_0^1 [F_i^{-1}(u) - F_j^{-1}(u)]^2 du,$$

which can be used for the same purpose. However $d(F_i, F_j)$ and $\mathcal{W}_{ij}$ may not give Euclidean distance matrices and are not invariant under affine transformation of the variables. On the other hand, $\mathcal{W}_{ij}$ is directly related to the maximum correlation (see Cuadras and Cuadras, 2002).

Fortiana and Grané (2002, 2003) refined this approach. They used some statistics based on this maximum correlation and obtained asymptotic and exact tests for testing the exponentiality and the uniformity of a sample, which compare with other goodness-of-fit statistics.

Example 3 Suppose that χ is the $N = 50$ sample of $X =$ "sepal length" of *Iris setosa,* the well-known data set used by R. A. Fisher to illustrate discriminant analysis (see Mardia *et al.*, 1979). Suppose the following statistical models:

$$\{U(\text{uniform}), E(\text{exponential}), N(\text{normal}), G(\text{gamma}), LN(\text{log-normal})\}.$$

The matrix of maximum correlations is reported in Table 3. Figure 4 is the metric scaling representation of probability models and data. The closest model is N, so we may decide that this data is drawn from a normal distribution.

The uniform distribution U, the second closest distribution to the data, may be another candidate. To decide between normal and uniform we may proceed as follows.

First, we assume normality and perform the integral transformation $y = \Phi(x)$ on the standardized sample, where Φ is the $N(0, 1)$ cdf, and correlate the transformed sample χ^*, say, with the principal components $(f_n(X))$, see Section 4. Let $r_k = \text{Cor}(\chi^*, f_k(U))$

Table 3: *Maximum Hoeffding correlations among several distributions and data. U (uniform), E (exponential), N (normal), G (gamma), LN (log-normal).*

	U	*E*	*N*	*GA*	*LN*	*Data*
U	1					
E	0.8660	1				
N	0.9772	0.9032	1			
G	0.9472	0.9772	0.9730	1		
LN	0.6877	0.8928	0.7628	0.8716	1	
Data	0.9738	0.8925	0.9871	0.9660	0.7452	1

be the coefficient of correlation between χ^*, with empirical cdf F_N^*, and $f_k(U)$, where the correlation is taken with respect to the upper bound $H_N^+(x,u) = \min[F_N^*(x), u]$. The theoretical correlations are:

$$\rho_k = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ 4\sqrt{6}/(k\pi^2) & \text{if } k \text{ is even.} \end{cases}$$

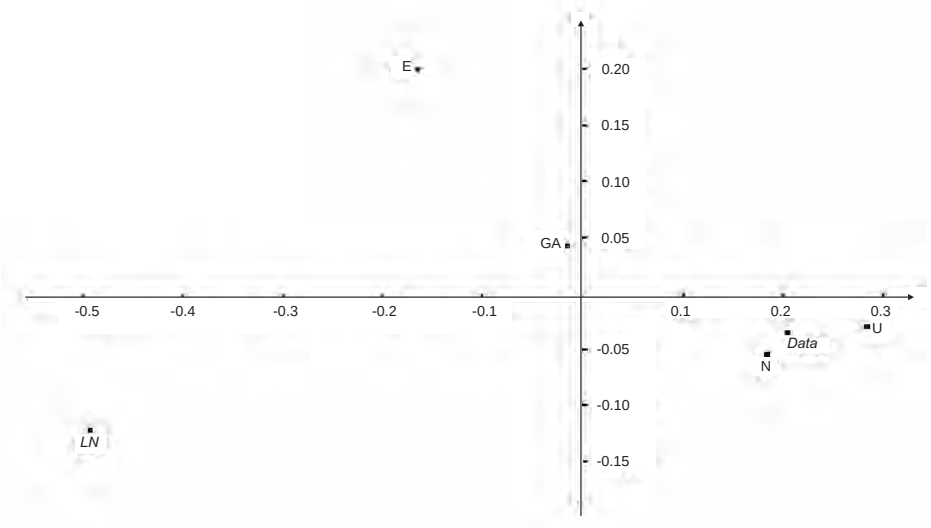

Figure 4: *Metric scaling representation of a sample (Data) and the distributions uniform (U), exponential (E), gamma (GA), normal (N) and log-normal (LN), using the maximum correlation between two distributions (Cuadras and Fortiana, 1994).*

It can be proved that $\rho^+(\chi^*, U) = \sum_{k\geq 1} \rho_k r_k$. Thus

$$\rho^+(\chi^*, U) = (4\sqrt{6}/\pi^2) \sum_{k=0}^{\infty} (r_{2k+1})/(2k+1)^2.$$

This agreement coefficient $\rho^+(\chi^*, U)$ between the transformed sample and the uniform distribution has an expansion similar to the expansions of Cramér-von Mises and Anderson-Darling statistics used in goodness-of-fit.

Second, we perform analogous computations for the original sample χ assuming that it is drawn from an uniform distribution, the alternative model. This gives Table 4, where $\rho^+(\chi, U) = 0.9738$, $\rho^+(\chi^*, U) = 0.9915$, $\rho^+(\chi^*, f_1(U)) = 0.9792$, etc. These results give support to the normality of the sample. See Cuadras and Fortiana (1994), Cuadras and Lahlou (2000) and Cuadras and Cuadras (2002) for further aspects of this graphic test.

Table 4: *Maximum correlations between normal and uniform and correlations among principal directions and data.*

	Theoretical	Normal	Uniform
ρ^+	1	0.9915	0.9738
ρ_1	0.9927	0.9792	0.9435
ρ_2	0	0.0019	−0.0102
ρ_3	0.1103	0.1713	0.2847
ρ_4	0	−0.0277	−0.0324

8 Kendall's tau and Spearman's rho

Any proposal of coefficient of stochastic dependence between X and Y should evaluate the difference between the joint cdf H and the independence FG. Thus

$$A(X, Y) = c \int_{\mathbf{R}^2} (H(x, y) - F(x)G(y))d\mu,$$

where c is a normalizing constant and μ is a suitable measure. The maximum value for $A(X, Y)$, where H has margins F, G, is attained at the upper bound:

$$\max A(X, Y) = c \int_{\mathbf{R}^2} (H^+(x, y) - F(x)G(y))d\mu.$$

However, as proposed by Hoeffding (1940), it is quite convenient to have a coefficient "scale invariant", that is, it should remain unchanged by monotonic transformations of X and Y. The integral transformation $u = F(x)$ and $v = G(y)$ is a monotonic transformation that provides the copula C_H, i.e., a bivariate cdf with uniform marginals on $\mathbf{I} = [0, 1]$, such that

$$H(x, y) = C_H(F(x), G(y)).$$

This copula exists (Sklar's theorem) and is unique if F, G are continuous. Thus we can construct bivariate distributions $H = C(F, G)$ with given univariate marginals F, G by

using copulas C. ("Copula" as a function which links marginals was coined by Sklar (1959). The same concept was called "uniform representation" by G. Kimeldorf and A. Sampson in 1975, and "dependence function" by P. Deheuvels and J. Galambos in 1978).

Kendall's τ and Spearman's ρ_s are coefficients of dependence computed from the copula C_H by using $d\mu = dC_H$ and $d\mu = dudv$, respectively. They are defined by

$$\begin{aligned}\tau &= 4\int_{\mathbf{I}^2}(C_H(u,v) - uv)\, dC_H(u,v) \\ &= 4\int_{\mathbf{I}^2} C_H(u,v)\, dC_H(u,v) - 1,\end{aligned}$$

and

$$\begin{aligned}\rho_s &= 12\int_{\mathbf{I}^2}(C_H(u,v) - uv)\, dudv \\ &= 12\int_{\mathbf{I}^2} C_H(u,v) dudv - 3.\end{aligned}$$

Then $\tau = \rho_s = 1$ when $H = H^+$, that is, when $F(X) = G(Y)$ (a.s.).

Spearman's ρ_s is Pearson's correlation between $F(X)$ and $G(Y)$ and can be 0 even if there is stochastic dependence. For example, $\rho_s = 0$ for the copula $C = uv + \theta(2u^3 - 3u^2 + u)(2v^3 - 3v^2 + v)$, $|\theta| \leq 1$. On the other hand, $F(X)$ is the first principal dimension for the logistic distribution (Section 4). Then we may extend ρ_s by obtaining Pearson's correlation $\mathrm{Cor}(f_1(X), g_1(Y))$ between the principal dimensions $f_1(X)$ and $g_1(Y)$. This may improve the measure of stochastic dependence between X and Y, with applications to testing independence (Cuadras, 2002b, 2002c).

Table 5: *Some parametric families, their properties and the upper bound.*

Family	Spearman	Kendall	Constant	Archimedian	Upper bound
FGM	yes	yes	yes	no	no
Normal	yes	yes	yes	no	yes
Plackett	yes	no	yes	no	yes
Cuadras-Augé	yes	yes	yes	no	yes
Regression (Fréchet)	yes	yes	yes	no	yes
Clayton-Oakes	yes	yes	yes	yes	yes
AMH	yes	yes	yes	yes	no
Frank	yes	yes	no	yes	yes
Raftery	yes	yes	no	no	yes
Gumbel-Barnett	no	no	no	yes	no
Gumbel-Hougaard	no	yes	yes	yes	yes
Joe	no	no	no	yes	yes

9 Some bivariate families

In this section we present some parametric families of bivariate distributions, in terms of F, G rather than copulas, as some aspects such as constant quantity and regression are not well manifested with uniform marginals. To get the corresponding copula simply replace F, G by u, v. For instance, for the FGM family the copula is $C_\alpha = uv[1+\alpha(1-u)(1-v)]$. References for these families can be found in Cuadras (1992, 1996, 2002, 2005), Druet-Mari and Kotz (2001), Hutchinson and Lai (1991), Joe (1997), Kotz *et al.* (2000), Mardia (1970) and Nelsen (1999). Table 5 summarizes some aspects, e.g., whether or not ρ_s and τ can be given in closed form and the family contains the upper bound.

9.1 FGM

The Farlie-Gumbel-Morgenstern family provides a simple and widely used example of distribution H with marginals F, G. This family does not reach H^+ and can be seen as the first term in the diagonal expansion (4).

1. Cdf : $H_\alpha = FG[1+\alpha(1-F)(1-G)], \quad -1 \le \alpha \le +1.$
2. Constant quantity : $\alpha = (H-FG)/[FG(1-F)(1-G)].$
3. Spearman: $\rho_s = \alpha/3.$
4. Kendall: $\tau = 2\alpha/9.$
5. Maximal correlation : $\rho_1 = |\alpha|/3.$
6. Fréchet-Hoeffding bounds : $H^- < H_{-1} < H_0 = FG < H_{+1} < H^+.$

9.2 Normal

Let N_ρ and n_ρ be the cdf and pdf, respectively, of the standard bivariate normal with correlation coefficient ρ. The distribution H_ρ is obtained by the "translation method", as described by Mardia (1970).

1. Cdf : $H_\rho = N_\rho(\Phi^{-1}F, \Phi^{-1}G), \quad -1 \le \rho \le +1.$
2. Constant quantity : $\frac{\rho}{1-\rho^2} = \dfrac{\partial^2 \log n_\rho}{\partial x \partial y}$ (normal marginals).
3. Spearman: $\rho_s = \frac{6}{\pi}\arcsin(\rho/2).$
4. Kendall: $\tau = \frac{2}{\pi}\arcsin(\rho).$
5. Maximal correlation : $\rho_1 = \rho.$
6. Fréchet-Hoeffding bounds : $H_{-1} = H^- < H_0 = FG < H_1 = H^+.$

9.3 Plackett

The Plackett family arises in the problem of correlating two dichotomized variables X, Y, when the ranges are divided into four regions and the correlation is computed as a function of the association parameter ψ. Then H is defined such that

$$\psi = \frac{H(1-F-G+H)}{(F-H)(G-H)},$$

is constant. $\psi \geq 0$ is the cross product ratio in 2×2 contingency tables.

1. Cdf: $H_\psi = \left[S - \left\{S^2 - 4\psi(\psi-1)FG\right\}^{1/2}\right] / \{2(\psi-1)\}, \quad \psi \geq 0,$ where $S = 1 + (F+G)(\psi-1)$.
2. Constant quantity: $\psi = H(1-F-G+H)/[(F-H)(G-H)]$.
3. Spearman: $\rho_s = \dfrac{\psi+1}{\psi-1} - \dfrac{2\psi}{(\psi-1)^2}\ln\psi$.
4. Kendall: τ not in closed form.
5. Fréchet-Hoeffding bounds: $H_0 = H^- < H_1 = FG < H_\infty = H^+$.

9.4 Cuadras-Augé

The Cuadras-Augé family is obtained by considering a weighted geometric mean of the independence distribution and the upper Fréchet-Hoeffding bound. The corresponding copula is $C_\theta = (\min\{u,v\})^\theta (uv)^{1-\theta}$. Although obtained independently by Cuadras and Augé (1981), C_θ is the survival copula of the Marshall and Olkin (1967) bivariate distribution when the variables are exchangeable. (The survival copula for H is $C_{\overline{H}}$ such that $\overline{H} = C_{\overline{H}}(\overline{F}, \overline{G})$, where $\overline{F} = 1-F$, $\overline{G} = 1-G$, $\overline{H} = 1-F-G+H$). The canonical correlations for this family constitutes a continuous set. Kimeldorf and Sampson (1975) proposed a copula C_λ also related to Marshall-Olkin. C_λ is given in Block and Sampson (1988) in the form $C_\lambda = u+v-1+(1-u)^\lambda(1-v)^\lambda \min\{(1-u)^\lambda, (1-v)^\lambda\}$, $0 \leq \lambda \leq 1$. See Muliere and Scarsini (1987) for an unified treatment of C_θ, C_λ and other related copulas. A generalization of C_θ, proposed by Nelsen (1991), is $C_{\alpha,\beta} = \min\{u^\alpha, v^\beta\}u^{1-\alpha}v^{1-\beta}$ for $\alpha, \beta \in [0,1]$.

1. Cdf : $H_\theta = (\min\{F,G\})^\theta (FG)^{1-\theta}, \quad 0 \leq \theta \leq 1.$
2. Constant quantity: $\theta = \ln(H/FG)/\ln(H^+/FG)$.
3. Spearman: $\rho_s = 3\theta/(4-\theta)$.
4. Kendall: $\tau = \theta/(2-\theta)$.

5. Maximal correlation: $\rho_1 = \theta$.
6. Fréchet-Hoeffding bounds: $H_0 = FG < H_1 = H^+$.

9.5 Regression (Fréchet)

A family with a given correlation coefficient $0 \leq r \leq 1$ can be constructed taking (X, X) with probability r and (X, Y), where X, Y are independents, with probability $(1 - r)$. The cdf is then

$$H^*(x, y) = rF(\min\{x, y\}) + (1 - r)F(x)G(y).$$

A generalization is the regression family H_r defined below (Cuadras, 1992). Family (9) is a particular case. This family extends the weighted mean of the upper bound and independence, $H_\theta = \theta H^+ + (1-\theta)FG$, proposed by Fréchet (1951) and studied by Nelsen (1987) and Tiit (1986). Note that $H_\theta \neq H_r$ have the same copula. See also Section 6.

1. Cdf : $H_r(x, y) = rF(\min\{x, y\}) + (1 - r)F(x)J(y), \quad 0 \leq r < 1,$
   where $J(y) = [G(y) - rF(y)] / (1 - r)$ is a univariate cdf.
2. Spearman: $\rho_s = r$.
3. Kendall: $\tau = r(r + 2)/3$.
4. Constant quantity : $r = [H(x, y) - F(x)G(y)]/[F(\min\{x, y\}) - F(x)F(y)]$.
5. Fréchet-Hoeffding bounds : $H_0 = FG < H_1 \leq H^+$, with $H_1 = H^+$ if $F = G$.

9.6 Clayton-Oakes

The Clayton-Oakes distribution is a bivariate model in survival analysis, which satisfies for all failure times s and t, the equation

$$h(s, t)\bar{H}(s, t) = \frac{1}{c}\int_s^\infty h(u, v)du \int_t^\infty h(u, v)dv$$

where $\bar{H} = 1 - F - G + H$ and h is the density.

1. Cdf: $H_c = \max\{(F^{-c} + G^{-c} - 1)^{-1/c}, 0\}, \quad -1 \leq c < \infty.$
2. Spearman: $\rho_s = 12 \int_0^1 \int_0^1 (u^{-c} + v^{-c})^{-1/c} dudv - 3$ (see Hutchinson and Lai, 1991, p. 240).

3. Kendall: $\tau = c/(c+2)$.
4. Constant quantity : $c^{-1} = h\bar{H}/(\frac{\partial}{\partial x}\bar{H}\frac{\partial}{\partial y}\bar{H})$.
5. Fréchet-Hoeffding bounds : $H_{-1} = H^- < H_0 = FG < H_\infty = H^+$.

This family is also known as: 1) Kimeldorf and Sampson, 2) Cook and Johnson, 3) Pareto.

9.7 Frank

Frank's copula $C = -\theta^{-1}\ln(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1})$ arises in the context of associative functions. It is characterized by the property that $C(u,v)$ and $\hat{C}(u,v) = u+v-C(u,v)$ are associative, that is, $C(C(u,v),w) = C(u,C(v,w))$ and similarly for $\hat{C}$. Statistical aspects of Frank's family were given by Nelsen (1986) and Genest (1987).

1. Cdf: $H_\theta = -\theta^{-1}\ln(1 + \dfrac{(e^{-\theta F}-1)(e^{-\theta G}-1)}{e^{-\theta}-1}), \quad -\infty \le \theta \le \infty$.
2. Spearman: $\rho_s = 1 - (12/\theta)[D_1(\theta) - D_2(\theta)]$.
3. Kendall: $\tau = 1 - (4/\theta)[1 - D_1(\theta)]$, where $D_k(x) = \frac{k}{x^k}\int_0^x \frac{t^k}{e^t-1}dt$.
4. Fréchet-Hoeffding bounds: $H_{-\infty} = H^- < H_0 = FG < H_\infty = H^+$.

9.8 AMH

The Ali-Mikhail-Haq distribution is obtained by considering the odds in favour of failure against survival. Thus $(1-F)/F = K$ must be non-increasing and $F = 1/(1+K)$. The bivariate extension is $H = 1/(1+L)$, where L is the corresponding bivariate odds function. Some conditions to H gives the model below.

1. Cdf: $H_\alpha = FG/[1 - \alpha(1-F)(1-G)], \quad -1 \le \alpha \le +1$.
2. Constant quantity: $\alpha = (H - FG)/[H(1-F)(1-G)]$.
3. Spearman: $\rho_s = -\frac{12(1+\alpha)}{\alpha^2}\text{diln}(1-\alpha) - \frac{24(1-\alpha)}{\alpha^2}\ln(1-\alpha) - \frac{3(\alpha+12)}{\alpha}$, where $\text{diln}(1-\alpha) = \int_0^\alpha x^{-1}\ln(1-x)dx$ is the dilogarithmic function.
4. Kendall: $\tau = \frac{3\alpha-2}{3\alpha} - \frac{2(1-\alpha)^2}{3\alpha^2}\ln(1-\alpha)$.
5. Fréchet-Hoeffding bounds: $H^- < H_{-1} < H_0 = FG < H_1 < H^+$.

9.9 Raftery

The Raftery distribution is generated by considering

$$X = (1-\theta)Z_1 + JZ_3, \quad Y = (1-\theta)Z_2 + JZ_3,$$

where Z_1, Z_2 and Z_3 are independent and identically distributed exponential with $\lambda > 0$, and J is Bernoulli of parameter θ, independent of Z's. This distribution, via Sklar's theorem, generates a family.

1. Cdf: $H_\theta = H^+ + \dfrac{1-\theta}{1+\theta}(FG)^{1/(1-\theta)}\{1 - [\max\{F,G\}]^{-(1+\theta)/(1-\theta)}\}, \quad 0 \le \theta \le 1.$
2. Spearman: $\rho_s = \dfrac{\theta(4-3\theta)}{(2-\theta)^2}.$
3. Kendall: $\tau = \dfrac{2\theta}{3-\theta}.$
4. Fréchet-Hoeffding bounds: $H_0 = FG < H_1 = H^+.$

9.10 Archimedian copulas

For the independence copula $C = uv$ we have $-\ln C = -\ln u - \ln v$. A fruitful generalization of this additivity, related to $\varphi(t) = -\ln t$, is the key idea for defining the so-called Archimedian copulas (Genest and MacKay, 1987). The cdf is described by a function $\varphi : \mathbf{I} \to [0, \infty)$ such that

$$\varphi(1) = 0, \quad \varphi'(t) < 0, \quad \varphi''(t) > 0,$$

for all $0 < t < 1$, conditions which guarantee that φ has inverse. These copulas are defined as

$$\begin{array}{ll} C(u,v) & = \varphi^{-1}[\varphi(u) + \varphi(v)] \quad \text{if } \varphi(u) + \varphi(v) \le \varphi(0), \\ & = 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{array}$$

For example, the AMH copula $C = uv/[1 - \alpha(1-u)(1-v)]$ satisfies

$$1 + (1-\alpha)\frac{1-C}{C} = \left[1 + (1-\alpha)\frac{1-u}{u}\right]\left[1 + (1-\alpha)\frac{1-v)}{v}\right]$$

i.e., the above relation for $\varphi(t) = \ln[1 + (1-\alpha)(1-t)/t] = \ln[\{1 - \alpha(1-t)\}/t]$.

Archimedian copulas play an important role because they have interesting properties. For instance:

1. Probability density: $c(u,v) = -\varphi''(C(u,v))\,\varphi'(u)\varphi'(v)/\left[\varphi'(C(u,v))\right]^3$.
2. Kendall's tau: $\tau = 4\int_0^1 [\varphi(t)/\varphi'(t)]dt + 1$.
3. C has a singular component if and only if $\varphi(0)/\varphi'(0) \neq 0$. In this case

$$P[\varphi(U) + \varphi(V) = \varphi(0)] = -\frac{\varphi(0)}{\varphi'(0)}.$$

Table 6: *Basic copulas and some Archimedian families. Kendall's tau can not be given in closed form for Gumbel-Barnett and Joe.*

Copula	cdf	$\varphi(t)$	
Lower bound	$C^- = \max\{u+v-1, 0\}$	$(1-t)$	
Independence	$C^0 = uv$	$-\ln t$	
Upper bound	$C^+ = \min\{u, v\}$	Not archimedian	
Family	**cdf**	$\varphi(t)$	**Bounds**
Gumbel-Barnett	$FG\exp(-\theta \ln F \ln G)$ $0 < \theta \leq 1$	$\ln(1-\theta \ln t)$	$C_0 = C^0$ $C_1 < C^+$
Gumbel-Hougaard	$\exp(-[(-\ln F)^\theta + (-\ln G)^\theta]^{1/\theta})$ $1 \leq \theta < \infty, \quad \tau = 1 - 1/\theta$	$(-\ln t)^\theta$	$C_1 = C^0$ $C_\infty = C^+$
Joe	$1 - [\overline{F}^\theta + \overline{G}^\theta - \overline{F}^\theta\overline{G}^\theta]^{1/\theta}$ $1 \leq \theta < \infty, \overline{F} = 1 - F$	$-\ln[1-(1-t)^\theta]$	$C_1 = C^0$ $C_\infty = C^+$

Table 6 summarizes the Archimedian property for the three basic copulas and three Archimedian families. The Gumbel-Hougaard family can be obtained by compounding. The copula C_H is the only extreme-value distribution (i.e., C_H^n is also a cdf) which is Archimedian. The constant quantity is $\ln H(x,x)/\ln F(x)$ if $F = G$.

9.11 Shuffles of Min

Can complete dependence be very close to independence? Apparently not, as the opposite of stochastic independence between X and Y is the relation $Y = \varphi(X)$ (a.s.), where φ is a one-to-one function. When φ is monotonic non-decreasing, the distribution of (X, Y) is the upper bound $H^+ = \min\{F, G\}$, so $\rho_s = \tau = 1$.

Let us consider the related copula $C^+ = \min\{u, v\}$. A family of copulas, called shuffles of Min, has interesting properties and can be constructed from C^+. The support of this copula can be described informally by placing the mass of C^+ on $\mathbf{I}^2$, which is cut vertically into a finite number of strips. The strips are then shuffled with some of them

flipped around the vertical axes of symmetry and then reassembled to form the square again. A formal definition is given in Mikusinski *et al.* (1992).

If the copula of (X, Y) is a shuffle of Min, then it can be arbitrarily close to the independence copula $C^0 = uv$. It can be proved that, for any $\varepsilon > 0$, there exists a shuffle of Min C_ε such that $\sup_{u,v\in\mathbf{I}} |C_\varepsilon(u, v) - uv| < \varepsilon$. Statistically speaking, we may have a bivariate sample, where (x, y) are completely dependent, but being impossible to distinguish from independence. This family is even dense, that is, we may approximate any copula by a shuffle of Min.

10 Additional aspects

Here we consider more statistical and probabilistic concepts where the bivariate upper bound is also present.

10.1 Multivariate generation

Any bivariate cdf H can be generated by a copula C, i.e., $H = C(F, G)$, where F, G are univariate cdf's.

One is tempted to use multivariate marginals F, G of dimensions p and q and a bivariate copula C to construct $H = C(F, G)$. But, is H a cdf? As proved by Genest *et al.* (1995), the answer is no, except for the independence copula $C^0 = uv$. In particular, the upper bound is not useful for this purpose. For instance, if F_1, F_2, G are univariate cdf's for (X_1, X_2, Y) with $F = F_1F_2$, and we consider $H = \min\{F, G\}$, then $\min\{F_i, G\}$ is the distribution of $(X_i, Y), i = 1, 2$. Therefore, $F_1(X_1) = G(Y)$ and $F_2(X_2) = G(Y)$, which contradicts the independence of X_1 and X_2.

10.2 Distances between distributions

If X and Y have univariate cdf's F and G and joint cdf H, we can define a distance between X and Y (and between F and G) by using

$$d_\alpha(X, Y) = E_H|X - Y|^\alpha,$$

assuming that $E(X^\alpha)$ and $E(Y^\alpha)$ exist. For $\alpha > 1$ it can be proved (see Dall'Aglio, 1972) that the minimum of $d_\alpha(X, Y)$ when H has marginals F, G is obtained when $H^+ = \min\{F, G\}$. The case $\alpha = 2$ corresponds to the maximum correlation ρ^+ and was proved by Hoeffding (1940).

Several authors (J. Bass, S. Cambanis, R. L. Dobrushin, G. H. Hardy, C. L. Mallows, A. H. Tchen, S. S. Vallender, W. Whitt and others) have considered the extreme bounds

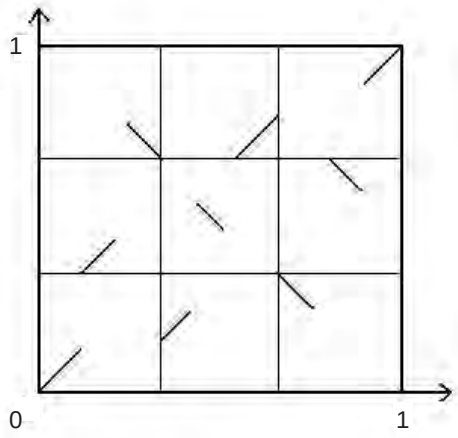

Figure 5: *A simple example of the support of a Shuffle of Min.*

for $E_H XY$, $E_H|X-Y|^\alpha$, $E_H f(|X-Y|)$ with $f'' > 0$, $\int \varphi(x,y)dH(x,y)$ where φ is superadditive (i.e., $\varphi(x,y) + \varphi(x',y') > \varphi(x',y) + \varphi(x,y')$ for all $x' > x$ and $y' > y$) and $E_H k(X,Y)$ when $k(x,y)$ is a quasi-monotone function (property quite similar to superadditivity). Thus, the supremum of $E_H XY$, $\int \varphi dH$ and $E_H k(X,Y)$ are achieved by the upper bound H^+. For instance, the supremum of $E_H k(X,Y)$ is

$$E_{H^+} k(X,Y) = \int_0^1 k(F^{-1}(u), G^{-1}(u))du.$$

See Tchen (1980).

10.3 Convergence in probability

The upper bound can also be applied to study the convergence in distribution and probability. Suppose that (X_n) is a sequence of r.v.'s with cdf's (F_n), which converges in probability to X with cdf F. Then it can be proved that this occurs if and only if

$$H_n(x,y) \to \min\{F(x), F(y)\},$$

where H_n is the joint cdf of (X_n, X). See Dall'Aglio (1972).

10.4 Lorenz curve and Gini coefficient

The Lorenz curve is a graphical representation of the distribution of a positive r. v. X. It is used to study the distribution of income. If X with cdf F ranges in (a,b), this curve is

defined by

$$L(y) = \frac{\int_a^y x dF(x)}{\int_a^b x dF(x)},$$

and can be given in terms of F

$$L(F) = \frac{\int_0^u F^{-1}(v)dv}{\int_0^1 F^{-1}(v)dv}.$$

The Lorenz curve is a convex curve in $\mathbf{I}^2$ under the diagonal from $(0,0)$ to $(1,1)$. Deviation from this diagonal indicates social inequality, see Figure 6. A global measure of inequality is the Gini coefficient $\mathcal{G}$, defined as twice the area between the curve and the diagonal:

$$\mathcal{G} = 1 - 2\int_0^1 L(F)dF.$$

For example, if X is Pareto with cdf $F(x) = 1 - (x/a)^c$ if $x > a$ then

$$L(F) = 1 - (1-F)^{1-1/c} \quad \text{and} \quad \mathcal{G} = 1/(2c-1).$$

The optimum social equality corresponds to $c = \infty$ and the maximum inequality to $c = 1$. However, for $c = 1$ the mean does not exist.

Assuming that X has finite variance $\sigma^2(X)$, Gini's coefficient can also be expressed as

$$\begin{aligned}
\mathcal{G} &= \int_a^b \int_a^b |x-y| dF(x) dF(y) \\
&= 2\int_a^b F(t)(1-F(t))dt \\
&= 4\mathrm{Cov}(X, F(X)) \\
&= \frac{2}{\sqrt{3}}\sigma(X)\mathrm{Cor}(X, F(X)).
\end{aligned}$$

But $\mathrm{Cor}(X, F(X))$ is the maximum Hoeffding correlation between X and U, where U is uniformly distributed. Thus, if $\sigma^2(X)$ exists, the maximum social inequality is $\mathcal{G} = 1/3$ and is attained when X is uniform, that is, when poor, middle and rich classes have the same proportions. Note that Pareto with $c = 2$ also gives $\mathcal{G} = 1/3$, but in this case $\sigma^2(X)$ does not exist.

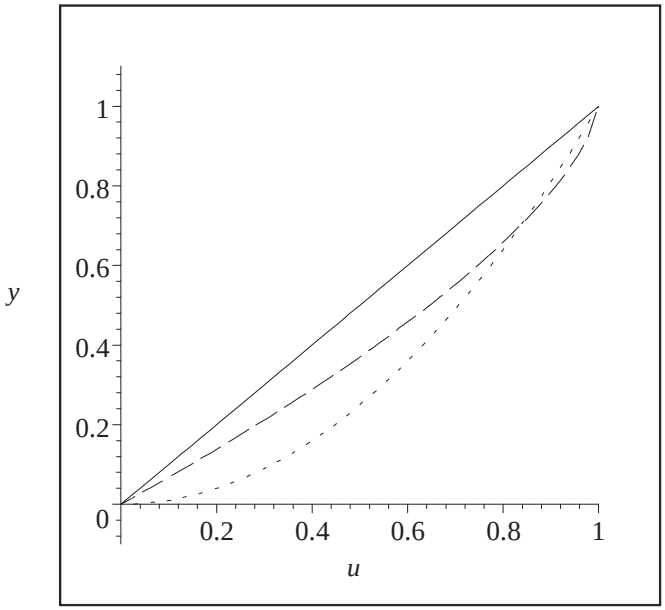

Figure 6: *Lorenz curve. Diagonal (solid line), Pareto with $c = 3$ (dash line) and uniform (dots line). The dots curve indicates maximum inequality (assuming finite variance) and this curve is related to the upper bound.*

10.5 Triangular norms and quasi-copulas

The theory of triangular norms (T-norms) is used in the study of associative functions, probabilistic metric spaces and fuzzy sets. See Schweizer and Sklar (1983), Alsina *et al.* (2006).

A T-norm T is a mapping from $\mathbf{I}^2$ into $\mathbf{I}$ such that $T(u,1) = u$, $T(u,v) = T(v,u)$, $T(u_1,v_1) \leq T(u_2,v_2)$ whenever $u_1 \leq u_2$, $v_1 \leq v_2$, and $T(T(u,v),w) = T(u,T(v,w))$.

Examples of T-norms are $C^- = \max\{u+v-1,0\}$, $C^+ = \min\{u,v\}$, $C^0 = uv$ and Z defined by $Z(a,1) = Z(1,a) = a$ and $Z(a,b) = 0$ otherwise. Note that Z is not a copula. It is readily proved that

$$Z < C^- < C^0 < C^+.$$

Thus the bivariate upper bound is the supremum of the partial ordered set of the T-norms.

A quasi-copula $Q(u,v)$ is a function $Q : \mathbf{I}^2 \to \mathbf{I}$ satisfying $Q(0,v) = Q(u,0) = 0$, $Q(u,1) = Q(1,u) = u$, Q is non-decreasing in each of its arguments, and Q satisfies the Lipschiptz's condition

$$|Q(u',v') - Q(u,v)| \leq |u'-v'| + |u-v|.$$

The quasi-copulas were introduced by Alsina *et al.* (1993) to study operations on univariate distributions not derivable from corresponding operations on random variables on the same probability space. For example, if X and Y are independent with cdf's F and G, the convolution $F * G(x) = \int F(x-y)dG(y)$ provides the cdf of $X+Y$. However, the geometric mean $\sqrt{FG}$ can not be the cdf of the random variable $K(X,Y)$, for a Borel-measurable function K (see Nelsen, 1999). Any copula is a quasi-copula and again the bivariate upper bound is the supremum of the partial ordered set of the quasi-copulas.

11 Bayes tests in contingency tables

We show here that the upper bound is also related to the test of comparing two proportions from a Bayesian perspective.

The problem of choosing intrinsic priors to perform an objective analysis is discussed in Casella and Moreno (2004). We choose a prior distribution which puts positive mass on the null hypothesis. This distribution depends on a positive parameter measuring dependence, which can be estimated via Pearson's contingency coefficient. Thus this test can be approached by the chi-square test and improved by obtaining the Bayes factor. Once again, the upper bound appears in this context.

Suppose that k_1, k_2 are binomial independent $B(n_1, p_1), B(n_2, p_2)$, respectively. We consider the test of hypothesis

$$H_0 : p_1 = p_2 \quad vs. \quad H_1 : p_1 \neq p_2. \tag{11}$$

Writing $k_i' = n_i - k_i, i = 1, 2$, the classic (asymptotic) approach is based on the chi-square statistic

$$\chi_1^2 = n\phi^2, \tag{12}$$

where $n = n_1 + n_2$ and

$$\phi^2 = (k_1 k_2' - k_1' k_2)^2 / [(k_1 + k_2)(k_1' + k_2')(k_1 + k_1')(k_2 + k_2')]$$

is the squared phi-coefficient.

Let us suppose that (p_1, p_2) is an observation of a random vector (P_1, P_2), with support $\mathbf{I}^2$, following one of the following copulas:

$$\begin{aligned} C_1(p_1, p_2) &= \theta_1 \min\{p_1, p_2\} + (1-\theta_1)(p_1 p_2), \quad 0 \leq \theta_1 \leq 1, \\ C_2(p_1, p_2) &= \min\{p_1, p_2\}^{\theta_2} (p_1 p_2)^{1-\theta_2}, \quad 0 \leq \theta_2 \leq 1, \quad (p_1, p_2) \in \mathbf{I}^2. \end{aligned}$$

C_1 is the copula related to the regression family, see (8), and was implicit in Fréchet

(1951). Copula C_2, proposed by Cuadras and Augé (1981), is the survival copula of the Marshall-Olkin distribution. These copulas satisfy (see also Sections 9.4 and 9.5):

1. There is independence for $\theta_i = 0$ and functional dependence for $\theta_i = 1$, that is,

$$P_1 = P_2 \text{ (a.s.) if } \theta_i = 1.$$

2. The pdf's with respect to the measure $\nu = \mu^2 + \mu_1$ are

$$\begin{aligned} c_1(p_1, p_2) &= (1 - \theta_1) + \theta_1 I_{\{p_1 = p_2\}}, \\ c_2(p_1, p_2) &= (1 - \theta_2) \max\{p_1, p_2\}^{-\theta_2} + \theta_2 p_1^{(1-\theta_2)} I_{\{p_1 = p_2\}}, \end{aligned}$$

   where $I_{\{p_1=p_2\}}$ is the indicator function, μ^2 and μ_1 are the Lebesgue measures on $\mathbf{I}^2$ and the line $p_1 = p_2$, respectively. Thus:

$$\int_0^{p_1} \int_0^{p_2} c_i(u_1, u_2) d\nu = C_i(p_1, p_2), \quad i = 1, 2.$$

3. These distributions have a singular part:

$$P_{C_1}[P_1 = P_2] = \theta_1, \quad P_{C_2}[P_1 = P_2] = \frac{\theta_2}{2 - \theta_2}.$$

4. The parameter θ measures stochastic dependence. Actually θ is the correlation coefficient for C_1 (see (10) for $\varphi(x) = x$) and the maximum correlation for C_2 (Cuadras, 2002a):

$$\theta_1 = \mathrm{Cor}_{C_1}(P_1, P_2), \qquad \theta_2 = \max_{\psi_1, \psi_2} \mathrm{Cor}_{C_2}(\psi_1(P_1), \psi_2(P_2)).$$

With these prior distributions (11) can be expressed as

$$H_0 : \theta = 1 \quad vs. \quad H_1 : 0 \le \theta < 1, \tag{13}$$

where θ is either θ_1 or θ_2. Note that to accept H_0 is equivalent to say that the copula reaches the upper Fréchet-Hoeffding bound.

As it has been presented in Section 9, there are other parametric copulas reaching the upper bound $\min\{p_1, p_2\}$, i.e., the hypothesis H_0. However in most of these copulas the probability of having $p_1 = p_2$ is zero and the estimation of the dependence parameter is not available from the contingency table.

Inference on the parameter θ_2 is discussed in Ruiz-Rivas and Cuadras (1988) and Ocaña and Ruiz-Rivas (1990). However, if only the sufficient statistic (k_1, k_2) is

available, in view of the above property, we may take the sample correlation between indicators of the events in a 2×2 contingency table. The square of this correlation is just the squared phi-coefficient ϕ^2. Accordingly, we may estimate θ by ϕ and the classic approach for testing (13) is again by means of (12).

Testing (11) or (13) may be improved using the Bayesian perspective. The likelihood function is

$$L(k_1, k_2; p_1, p_2) = p_1^{k_1}(1-p_1)^{k_1'} p_2^{k_2}(1-p_2)^{k_2'}.$$

Under $H_0 : p_1 = p_2 = p$ this function reduces to

$$L(k_1, k_2; p_1, p_2) = p^{k_1+k_2}(1-p)^{k_1'+k_2'}.$$

The Bayes factor in testing (11), expressed as $(p_1, p_2) \in \omega$ *vs.* $(p_1, p_2) \in \Omega - \omega$, where θ is interpreted as another unknown parameter, is

$$B_i = \frac{\int_\omega L(k_1, k_2; p_1, p_2) dC_i(p_1, p_2)}{\int_{\Omega-\omega} L(k_1, k_2; p_1, p_2) dC_i(p_1, p_2)}, \quad i = 1, 2.$$

Thus we obtain for copulas C_1 and C_2

$$B_1 = \frac{\int_0^1 p^{k_1+k_2}(1-p)^{k_1'+k_2'} dp}{\int_0^1 \int_0^1 (1-\theta_1) p_1^{k_1}(1-p_1)^{k_1'} p_2^{k_2}(1-p_2)^{k_2'} dp_1 dp_2 d\theta_1},$$

$$B_2 = \frac{\int_0^1 p^{k_1+k_2}(1-p)^{k_1'+k_2'} dp}{\int_0^1 \int_0^1 \int_0^1 (1-\theta_2) p_1^{k_1}(1-p_1)^{k_1'} p_2^{k_2}(1-p_2)^{k_2'} \max\{p_1, p_2\}^{-\theta_2} dp_1 dp_2 d\theta_2}.$$

High and low values of B_1 and B_2 give evidence for H_0 and H_1, respectively.

Finally, we can approach the more general hypothesis

$$H_0 : p_2 = \phi(p_1) \quad vs. \quad H_1 : p_2 \neq \phi(p_1),$$

where ϕ is a monotonic function, by using the family (8) as prior distribution, which also puts positive mass to the null hypothesis.

Example 4 Table 7 is a 2×2 contingency table summarizing the results of comparing surgery with radiation therapy in treating cancer, and was used by Casella and Moreno (2004). For this table we obtain

$$\widehat{\theta} = 0.1208,\ \chi_1^2 = 0.599,\ B_1 = 5.982,\ B_2 = 5.754.$$

The Bayes factors B_1, B_2, as well as χ^2_1, give support to the null hypothesis (the proportions are equal).

Table 7: *Contingency table combining treatment and cancer.*

	Cancer controlled	Cancer not controlled	
Surgery	21	2	23
Radiation therapy	15	3	18
	36	5	41

References

Alsina, C., Frank, M. J. and Schweizer, B. (2006). *Associative Functions: Triangular Norms and Copulas*. World Scientific, Singapore.

Alsina, C., Nelsen, R. B., and Schweizer, B. (1993). On the characterization of a class of binary operations on distribution functions. *Statistics and Probability Letters*, 17, 75-89.

Benzécri, J. P. (1973). *L'Analyse des Données. I. La Taxinomie. II. L'Analyse des Correspondances*. Dunod, Paris.

Block, H. W. and Sampson, A. R. (1988). Conditionally ordered distributions. *Journal of Multivariate Analysis*, 27, 91-104.

Casella, G. and Moreno, E. (2004). Objective Bayesian analysis of contingency tables. *Technical Report*, 2002-023. Department of Statistics, University of Florida.

Cuadras, C. M. (1992). Probability distributions with given multivariate marginals and given dependence structure. *Journal of Multivariate Analysis*, 42, 51-66.

Cuadras, C. M. (1996). A distribution with given marginals and given regression curve. In *Distributions with Fixed Marginals and Related Topics* (Eds. L. Rüschendorf, B. Schweizer and D. Taylor), pp. 76-84, IMS Lecture Notes-Monograph Series, Vol. 28, Hayward.

Cuadras, C. M. (2002*a*). On the covariance between functions. *Journal of Multivariate Analysis*, 81, 19-27.

Cuadras, C. M. (2002*b*). Correspondence analysis and diagonal expansions in terms of distribution functions. *Journal of Statistical Planning and Inference*, 103, 137-150.

Cuadras, C. M. (2002*c*). Diagonal distributions via orthogonal expansions and tests of independence. In *Distributions with Given Marginals and Statistical Modelling*, (Eds. C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena), pp. 35-42, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Cuadras, C. M. (2005*a*). Continuous canonical correlation analysis. *Research Letters in Information and Mathematical Sciences*, 8, 97-103.

Cuadras, C. M. (2005*b*). First principal component characterization of a continuous random variable. In *Advances on Models, Characterizations and Applications* (Eds. N. Balakrishnan, I. Bairamov and O. Gebizlioglu), pp. 189-199, Chapman & Hall/CRC-Press, New York.

Cuadras, C. M. and Augé, J. (1981). A continuous general multivariate distribution and its properties. *Communications in Statistics-Theory and Methods*, A10, 339-353.

Cuadras, C. M. and Cuadras, D. (2002). Orthogonal expansions and distinction between logistic and normal. In *Goodness-of-fit Tests and Model Validity*, (Eds. C. Huber-Carol, N. Balakrishnan, M. S. Nikulin and M. Mesbah), pp. 327-339, Birkhaüser, Boston.

Cuadras, C. M. and Fortiana, J. (1994). Ascertaining the underlying distribution of a data set. In *Selected Topics on Stochastic Modelling*, (Eds. R. Gutierrez and M. J. Valderrama), pp. 223-230, World Scientific, Singapore.

Cuadras, C. M. and Fortiana, J. (1995). A continuous metric scaling solution for a random variable. *Journal of Multivariate Analysis*, 52, 1-14.

Cuadras, C. M., Fortiana, J. and Greenacre, M. J. (2000). Continuous extensions of matrix formulations in correspondence analysis, with applications to the FGM family of distributions. In *Innovations in Multivariate Statistical Analysis* (Eds. R. D. H. Heijmans, D. S. G. Pollock and A. Satorra.), pp. 101-116, Kluwer Ac. Publ., Dordrecht.

Cuadras, C. M., Fortiana, J. and Rodriguez-Lallena, J. A. (Eds.) (2002). *Distributions with Given Marginals and Statistical Modelling*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Cuadras, C. M. and Lahlou, Y. (2000). Some orthogonal expansions for the logistic distribution. *Communications in Statistics-Theory and Methods*, 29, 2643-2663.

Cuadras, C. M., Cuadras, D. and Lahlou, Y. (2006). Principal directions for the general Pareto distribution. *Journal of Statistical Planning and Inference*, 136, 2572-2583.

Dall'Aglio, G. (1972). Fréchet classes and compatibility of distribution functions. *Symposia Mathematica*, 9, 131-150.

del Barrio, E., Cuesta-Albertos, J. A. and Matrán, M. (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *TEST*, 9, 1-96.

Druet-Mari, D. and Kotz, S. (2001). *Correlation and Dependence*. Imperial College Press, London.

Fortiana, J. and Grané, A. (2002). A scale-free goodness-of-fit statistic for the exponential distribution based on maximum correlations. *Journal of Statistical Planning and Inference*, 108, 85-97.

Fortiana, J. and Grané, A. (2003). Goodness-of-fit tests based on maximum correlations and their orthogonal decompositions. *Journal of the Royal Statisitcal Society, Series B*, 65, 115-126.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges son données. *Annales de l'Université de Lyon, Série* 3, 14, 53-77.

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, 74, 540-555.

Genest, C. and MacKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, 40, 280-283.

Genest, C., Quesada-Molina, J. J. and Rodriguez-Lallena, J. A. (1995). De l'impossibilté de construire des lois a marges données a partir de copules. *Comptes Rendus Academie Sciences Paris*, 320, 723-726.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press, London.

Hoeffding, W. (1940). Maszstabinvariante Korrelationstheorie. *Schriften des Mathematischen Institiuts und des Instituts für Angewandte Mathematik der Universitat Berlin*, 5, 179-233.

Hutchinson, T. P. and Lai, C. D. (1991). *The Engineering Statistician's Guide to Continuous Bivariate Distributions.* Rumsby Scientific Pub., Adelaide.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman and Hall, London.

Kimeldorf, G. and Sampson, A. (1975). Uniform representation of bivariate distributions. *Communications in Statistics*, 4, 617-627.

Kotz, S., Balakrishnan, N. and Johnson, N. L. (2000). *Continuous Multivariate Distributions*. Wiley, New York.

Mardia, K. V. (1970). *Families of Bivariate Distributions*. Griffin, London.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, London.

Marshall, A. W., and Olkin, I. (1967). A multivariate exponential distribution. *Journal of the American Statistical Association*, 62, 30-44.

Marshall, A. W., Meza, J. C. and Olkin, I. (2001). Can data recognize the parent distribution? *Journal of Computational and Graphical Statistics*, 10, 555-580.

Mikusisnki, P., Sherwood, H., Taylor, M. D. (1992). Shuffles of Min. *Stochastica*, 13, 61-74.

Muliere, P. and Scarsini, M. (1987). Characterization of Marshall-Olkin type class of distributions. *Annals Institute Statistical Mathematics*, 39, 429-441.

Nelsen, R. B. (1986). Properties of a one-parameter family of bivariate distributions with given marginals. *Communications in Statistics-Theory and Methods*, 15, 3277-3285.

Nelsen, R. B. (1987). Discrete bivariate distributions with given marginals and correlation. *Communications in Statistics-Theory and Methods*, 16, 199-208.

Nelsen, R. B. (1991). Copulas and Association. In *Advances in Probability Distributions with Given Marginals* (Eds. G. Dall'Aglio, S. Kotz and G. Salinetti), pp. 51-74, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer, New York.

Nguyen, T. T. and Sampson, A. R. (1985). The geometry of certain fixed marginal probability distributions. *Linear Algebra and Its Applications*, 70, 73-87.

Ocaña, J. and Ruiz-Rivas, C. (1990). Computer generation and estimation in a one-parameter system of bivariate distributions with specified marginals. *Communications in Statistics-Simulation and Computation*, 19, 37-55

Ruiz-Rivas, C. and Cuadras, C. M. (1988). Inference properties of a one-parameter curved family of distributions with given marginals. *Journal of Multivariate Analysis*, 27, 447-456.

Schweizer, B. and Sklar, A. (1983). *Probabilistic Metric Spaces*. North-Holland, New York.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229-231.

Tchen, A. H. (1980). Inequalities for distributions with given marginals. *The Annals of Probability*, 8, 814-827.

Tiit, E. (1986). Random vectors with given arbitrary marginal and given correlation matrix. *Acta et Commentationes Universitatis Tartuensis*, 733, 14-39.

SORT 30 (1) January-June 2006, 85-90

A matrix function useful in the estimation of linear continuous-time models

Heinz Neudecker

Cesaro

Abstract

In a recent publication Chen & Zadrozny (2001) derive some equations for efficiently computing e^A and $\nabla\, e^A$, its derivative. They employ an expression due to Bellman (1960), Snider (1964) and Wilcox (1967) for the differential de^A and a method due to Van Loan (1978) to find the derivative $\nabla\, e^A$. The present note gives *a*) a short derivation of $\nabla\, e^A$ by way of the Bellman-Snider-Wilcox result, *b*) a shorter derivation without using it. In both approaches there is no need for Van Loan's method.

MSC: 15A69

Keywords: matrix derivatives, vectorization, matrix exponential

1 Introduction

In a recent publication Chen & Zadrozny (2001) consider the matrix exponential

$$e^A = I_n + A + \frac{1}{2!}A^2 + \cdots + \frac{1}{k!}A^k + \cdots .$$

Their aim is to find $\nabla\, e^A$ which is *implicitly* defined by

$$d \operatorname{vec} e^A = \left(\nabla\, e^A\right) d \operatorname{vec} A$$

Address for correspondence: Oosterstraat, 13. 1741 GH Schagen. The Netherlands.
E-mail: H.Neudecker@uva.nl
Received: February 2005
Accepted: June 2006

or *explicitly* as

$$\nabla e^A = \frac{\partial \operatorname{vec} e^A}{\partial (\operatorname{vec} A)'}.$$

They introduce a matrix function

$$K(t) = \int_{\tau=0}^{t} e^{(t-\tau)A'} \otimes e^{\tau A}\, d\tau$$

and recall that

$$d\, e^A = \int_{\tau=0}^{1} e^{\tau A}\ (d\, A)\ e^{(1-\tau)A}\, d\tau.$$

See Bellman (1960, p. 171, (7)), Snider (1960) and Wilcox (1967). It then follows by vectorization that

$$d \operatorname{vec} e^A = K(1)\, d \operatorname{vec} A.$$

The authors compute $K(1)$ by employing a method due to Van Loan (1978). It turns out that $K(1)$ is the submatrix in the northeast corner of e^C, where

$$C = \begin{pmatrix} A' \otimes I_n & I_{n^2} \\ 0 & I_n \otimes A \end{pmatrix}.$$

The submatrix is denoted by $G_1(1)$. In this note we shall use two methods to find ∇e^A: *a*) one by way of the Bellman-Snider-Wilcox result, *b*) another more direct one without using that earlier result.

2 The derivation through *K*(1)

Starting from the expression

$$K(1) = \int_{\tau=0}^{1} \left(e^{(1-\tau)A'} \otimes I_n\right)\left(I_n \otimes e^{\tau A}\right)\, d\tau,$$

which is also to be found in Chen & Zadrozny, we further develop

$$\begin{aligned} K(1) &= \int_{\tau=0}^{1} e^{(1-\tau)A' \otimes I_n}\, e^{\tau\, I_n \otimes A}\, d\tau = \\ &= \int_{\tau=0}^{1} e^{(1-\tau)A' \otimes I_n + \tau\, I_n \otimes A}\, d\tau = \end{aligned}$$

$$= \int_{\tau=0}^{1} e^{A'\otimes I_n + \tau(I_n\otimes A - A'\otimes I_n)}\, d\tau =$$

$$= e^{A'\otimes I_n} \int_{\tau=0}^{1} e^{\tau(I_n\otimes A - A'\otimes I_n)}\, d\tau =$$

$$= e^{A'\otimes I_n}\left[I_{n^2} + \frac{1}{2!}(I_n\otimes A - A'\otimes I_n) + \cdots + \frac{1}{(k+1)!}(I_n\otimes A - A'\otimes I_n)^k + \cdots\right] =$$

$$= I_{n^2} + \frac{1}{2!}(I_n\otimes A + A'\otimes I_n) + \cdots + \frac{1}{(k+1)!}(I_n\otimes A + A'\otimes I_n)^{(k)} + \cdots$$

where for *commuting* A and B:

$$(A+B)^{(i)} = \sum_{j=1}^{i-1} A^j B^{i-j} + A^i + B^i, \quad (A+B)^{(1)} = A + B.$$

Clearly $I_n \otimes A$ and $A' \otimes I_n$ commute. We used Properties 3, 4 and 5 of the Appendix. It is clear that $K(1) = G_1(1)$, given the following computation. Consider

$$C = \begin{pmatrix} P & I \\ O & Q \end{pmatrix} \quad \text{and} \quad e^C = \begin{pmatrix} R & S \\ T & U \end{pmatrix}.$$

Then $T = O$, $R = I + \sum_{i=1}^{\infty} \frac{1}{i!} P^i = e^P$, $U = I + \sum_{i=1}^{\infty} \frac{1}{i!} Q^i = e^Q$ and

$$S = I_{n^2} + \sum_{i=1}^{\infty} \frac{1}{(i+1)!}(P+Q)^{(i)}.$$

Hence $G_1(1) = I_{n^2} + \sum_{i=1}^{\infty} \frac{1}{(i+1)!}(A'\otimes I + I\otimes A)^{(i)} = K(1)$. We can also define

$$C = \begin{pmatrix} I_n\otimes A & I_{n^2} \\ 0 & A'\otimes I_n \end{pmatrix}$$

to get the same $G_1(1)$.

3 A direct derivation of ∇e^A

Still simpler is to proceed as follows. Differentiation of e^A yields

$$d\,e^A = dA + \frac{1}{2!}\{(dA)A + AdA\} + \frac{1}{3!}\left\{(dA)A^2 + A(dA)A + A^2dA\right\} + \cdots$$

and from this by vectorization

$$\begin{aligned} d\,\mathrm{vec}\,e^A &= d\,\mathrm{vec}\,A + \frac{1}{2!}(I_n \otimes A + A' \otimes I_n)\,d\,\mathrm{vec}\,A + \frac{1}{3!}\left\{I_n \otimes A^2 + A' \otimes A + (A')^2 \otimes I\right\} \\ &\quad \times d\,\mathrm{vec}\,A + \cdots = \\ &= \left[I_{n^2} + \frac{1}{2!}(I_n \otimes A + A' \otimes I_n) + \frac{1}{3!}(I_n \otimes A + A' \otimes I_n)^{(2)} + \cdots\right] d\,\mathrm{vec}\,A = \\ &= \left[I_{n^2} + \sum_{i=1}^{\infty} \frac{1}{(i+1)!}(I_n \otimes A + A' \otimes I_n)^{(i)}\right] d\,\mathrm{vec}\,A. \end{aligned}$$

Hence

$$\nabla e^A = \frac{\partial\,\mathrm{vec}\,e^A}{\partial\,(\mathrm{vec}\,A)'} = I_{n^2} + \sum_{i=1}^{\infty} \frac{1}{(i+1)!}(I_n \otimes A + A' \otimes I)^{(i)}\,.$$

4 Appendix

Some algebraic properties:

1. $\mathrm{vec}\,ABC = (C' \otimes A)\,\mathrm{vec}\,B$.
2. $(A \otimes B)(C \otimes D) = AC \otimes BD$.
3. $e^{I \otimes A} = I \otimes e^A, \quad e^{A \otimes I} = e^A \otimes I$.
4. $e^A \cdot e^B = e^{A+B}$ for commuting A and B.
5. $e^B\left[I + \frac{1}{2!}(A-B) + \frac{1}{3!}(A-B)^2 + \cdots\right] = I + \frac{1}{2!}(A+B) + \frac{1}{3!}(A+B)^{(2)} + \cdots$

   where $(A+B)^{(i)} = \sum_{j=1}^{i-1} A^j B^{i-j} + A^i + B^i$ for commuting A and B.

References

Bellman, R. (1960). *Introduction to Matrix Analysis*. McGraw-Hill, New York.

Chen, B. and Zadrozny, P. A. (2001). Analytic derivatives of the matrix exponential for estimation of linear continuous-time models. *Journal of Economic Dynamics & Control*, 25, 1867-1879.

Neudecker, H. (1971). On a theorem of Snider and Wilcox. *METU Journal of Pure and Applied Sciences*, 4, 217-220.

Snider, R. F. (1964). Perturbation variation methods for a quantum Boltzmann equation. *Journal of Mathematical Physics*, 5, 1580-1587.

Van Loan, C. F. (1978). Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control*, 23, 395-404.

Wilcox, R. M. (1967). Exponential operators and parameter differentiation in quantum physics. *Journal of Mathematical Physics*, 8, 962-982.

SORT 30 (1) January-June 2006, 91-100

About one problem of Bernoulli and Euler from the theory of statistical estimation

Mikhaïl Nikulin

Université Bordeaux

Abstract

We consider some results by D. Bernoulli and L. Euler on the method of maximum likelihood in parametric estimation. The statistical analysis is made by considering a parametric family with a shift parameter.

MSC: 62H03, 62F10

Keywords: Asymptotic normality, sample mean, Bernoulli's estimator, Euler's estimator, maximum likelihood, unbiased estimator.

1 Introduction

Kendall (1961) published a paper on Daniel Bernoulli and the maximum likelihood. This paper quotes two papers: Bernoulli (1961) and Euler (1961). The paper of D. Bernoulli and the commentary by Euler appeared in Latin (1777). An interesting discussion about this problem can be found in Stigler (1997).

We shall consider here one contribution of D. Bernoulli and L. Euler in the estimation of parameters, in particular on the method of maximum likelihood. For more aspects about the principle of maximum likelihood (ML) in estimation, see, for example, Huber and Nikulin (1997).

Address for correspondence: Mikhaïl Nikulin. Université Bordeaux
Received: April 2006
Accepted: May 2006

Bernoulli and Euler considered the problem of statistical estimation of the parameter θ of the probability density

$$p(x;\theta) = \begin{cases} \dfrac{2}{\pi}\sqrt{1-(x-\theta)^2} & \text{if} \quad |x-\theta| \le 1, \quad |\theta| < \infty. \\ 0 & \text{otherwise.} \end{cases}$$

Bernoulli proposed to estimate θ by the ML method. Euler agreed with Bernoulli, but he provided a different estimator. Who was right? This question was posed by L.N. Bolshev in 1969. We shall consider here both approaches under a more general case.

Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$ be a sample, where $X_1, X_2, \cdots, X_n$ are independent identically distributed random variables with density

$$p_k(x;\theta) = \begin{cases} \dfrac{\Gamma(2k)}{\Gamma(k)\Gamma(k)}\left(\dfrac{1}{2}\right)^{2k-1}\left[1-(x-\theta)^2\right]^{k-1} & \text{if} \quad |x-\theta| \le 1, \\ 0 & \text{if} \quad |x-\theta| > 1. \end{cases}$$

We have to estimate the shift parameter θ, where k is given, with $k \in [1, 2]$.

The family $\{p_k(x;\theta)\}$ is quite rich. In particular, if $k = 1$ it contains the uniform distribution with support

$$\theta - 1 \le x \le \theta + 1, \quad |\theta| < \infty.$$

If $k = 1.5$ (the case of Bernoulli) the graph of $p_{1.5}(x;\theta) = p(x;\theta)$ is a half ellipse with parameters $a = 1$ and $b = 2/\pi$, and two tangents to the extremes of the curve, $x = \theta - 1$ and $x = \theta + 1$, orthogonal to the axis OX.

L. N. Bolshev proposed to find the ML estimate of θ, as Bernoulli did for the case $k = 1.5$.

Let us denote

$$L(\theta) = \prod_{i=1}^{n} p_k(X_i;\theta)$$

the likelihood function, obtained with the data $\mathbf{X}$, and let $\hat{\theta}_n$ be the value of θ that maximises $L(\theta)$:

$$L(\hat{\theta}_n) = \max_{|\theta|<\infty} L(\theta),$$

with the constraint $\max_i |X_i - \theta| \le 1$, i.e., $0 \le X_{(n)} - X_{(1)} \le 2$, where $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ is the ordinal statistic.

As is well known, it is more convenient to consider $\ln L(\theta)$. In our case we have

$$\ln L(\theta) = (k-1)\sum_{i=1}^{n} \ln\left[1-(X_i-\theta)^2\right] + n\ln\frac{\Gamma(2k)}{\Gamma(k)\Gamma(k)2^{2k-1}}. \tag{1.1}$$

To find $\hat{\theta}$ we must solve the ML equation:

$$\frac{\partial}{\partial\theta}\ln L(\theta) = 0. \tag{1.2}$$

From (1.1) and (1.2) we obtain:

$$\sum_{i=1}^{n}\frac{X_i-\theta}{1-(X_i-\theta)^2} = 0. \tag{1.3}$$

Following Euler, this equation can be expressed as:

$$\sum_{i=1}^{n}\frac{1}{1+X_i-\theta} = \sum_{i=1}^{n}\frac{1}{1-X_i+\theta}.$$

It is worth noting that (1.3) does not depend on k. A solution $\hat{\theta}_n$ is the ML estimator, which was proposed by Bernoulli. One can verify that $\hat{\theta}_n$ satisfies the equation

$$\hat{\theta}_n = \frac{\sum_{i=1}^{n}\left\{\frac{1}{1-(X_i-\hat{\theta}_n)^2}X_i\right\}}{\sum_{i=1}^{n}\frac{1}{1-(X_i-\hat{\theta}_n)^2}}. \tag{1.4}$$

More exactly, we can say that (1.4) is (1.3) "solved" with respect to θ. We can find $\hat{\theta}_n$ by using iterative procedures.

The same problem was considered by Euler, who knew Bernoulli's result. Euler proposed the estimator

$$\theta_n^* = \frac{\sum_{i=1}^{n}\left\{\left[1-(X_i-\theta_n^*)^2\right]X_i\right\}}{\sum_{i=1}^{n}\left[1-(X_i-\theta_n^*)^2\right]}. \tag{1.5}$$

Note the difference between θ_n^* and $\hat{\theta}_n$, as in (1.4) and (1.5) the observations X_i have different weights.

Clearly, to estimate θ we can also take the arithmetic mean

$$\bar{X}_n = \frac{1}{n}(X_1+\cdots+X_n).$$

Our aim is to compare these three estimators $\hat{\theta}_n$, θ_n^* and $\overline{X}_n$. We can suppose $\theta = 0$, i.e. we can consider that $\hat{\theta}_n, \theta_n^*$ and $\overline{X}_n$ are estimators of zero, as they are invariant under translation when the loss is quadratic.

2 Arithmetic mean

Since $\mathrm{E}(X_i) = 0$, we obtain that $\overline{X}_n$ is an unbiased estimator of 0, and $\mathrm{Var}(X_i) = \mathrm{E}(X_i)^2$. To compute $\mathrm{Var}(X_i)$ let us find the moments. It is evident that

$$\mathrm{E}(X_i)^{2m+1} = 0, \quad m = 0, 1, 2, \cdots \tag{2.1}$$

On the other hand we have ($\theta = 0$):

$$\begin{aligned} \mathrm{E}(X_i)^{2m} &= \frac{\Gamma(2k)}{\Gamma(k)\Gamma(k)} \left(\frac{1}{2}\right)^{2k-1} \hat{E} \int_{-1}^{1} x^{2m}(1-x^2)^{k-1} dx \\ &= \frac{\Gamma(2k)}{\Gamma(k)\Gamma(k)} \left(\frac{1}{2}\right)^{2k-1} B\left(m + \frac{1}{2}, k\right) \\ &= \frac{\Gamma(2k)\Gamma\left(m + \frac{1}{2}\right)}{\Gamma(k)\Gamma\left(m + k + \frac{1}{2}\right)} \left(\frac{1}{2}\right)^{2k-1}. \end{aligned} \tag{2.2}$$

Table 1

$k \backslash^m$	1	2	3
1	$\frac{1}{3}$	$\frac{1}{5}$	$\frac{1}{7}$
$\frac{3}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{5}{64}$
2	$\frac{1}{5}$	$\frac{3}{35}$	$\frac{1}{21}$

From (2.2) we obtain the Table 1 for some values of $\mathrm{E}(X_i)^{2m}$.
From (2.1), (2.2) and Table 1 it follows that

$$\mathrm{Var}(X_i) = \mathrm{E}(X_i)^2 = \frac{\sqrt{\pi}}{2^{2k}} \frac{\Gamma(2k)}{\Gamma(k)\Gamma\left(k + \frac{3}{2}\right)}. \tag{2.3}$$

(In the particular case $k = 1.5$, i.e., in the cases of Bernoulli and Euler we have $\mathrm{Var}(X_i) = 1/4$). From (2.3) we obtain

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n}\frac{\sqrt{\pi}}{2^{2k}}\frac{\Gamma(2k)}{\Gamma(k)\Gamma\left(k+\frac{3}{2}\right)}, \tag{2.4}$$

and hence we have

$$\mathrm{Var}(\bar{X}_n) = \begin{cases} \dfrac{1}{3n} & \text{if } k = 1, \\ \dfrac{1}{4n} & \text{if } k = \frac{3}{2}, \\ \dfrac{1}{5n} & \text{if } k = 2. \end{cases} \tag{2.5}$$

Furtheremore, from the central limit theorem for any $k \in [1, 2]$ and any $x \in \boldsymbol{R}$

$$\lim_{n\to\infty} \mathrm{P}\left(\left\{\bar{X}_n \le x\sqrt{\frac{1}{n}\frac{\sqrt{\pi}}{2^{2k}}\frac{\Gamma(2k)}{\Gamma(k)\Gamma\left(k+\frac{2}{3}\right)}}\right\}\right) = \Phi(x), \tag{2.6}$$

where $\Phi(x)$ is the cdf of the $N(0, 1)$ distribution.

3 Euler's estimator θ_n^*

Recall that $\theta = 0$. From (1.5) θ_n^* is the solution of the equation

$$\theta_n^* \sum_{i=1}^{n}\left[\left(1 - X_i^2\right) + 2X_i\theta_n^* - (\theta_n^*)^2\right] = \sum_{i=1}^{n}\left[\left(1 - X_i^2\right) + 2X_i\theta_n^* - (\theta_n^*)^2\right]X_i. \tag{3.1}$$

Let us denote $\bar{X}_n^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$. It is evident that $\{\bar{X}_n\}$ converges in probability to 0, i.e., $\mathrm{P}\left(\lim_{n\to\infty}\bar{X}_n\right) = 0$, and similarly $\mathrm{P}\left(\lim_{n\to\infty}\bar{X}_n^3\right) = 0$. Furthermore, we have

$$\mathrm{P}\left(\lim_{n\to\infty}\bar{X}_n^2\right) = \mathrm{E}(X_i)^2 = \begin{cases} \dfrac{1}{3} & \text{if } k = 1, \\ \dfrac{1}{4} & \text{if } k = \frac{3}{2}, \\ \dfrac{1}{5} & \text{if } k = 2. \end{cases} \tag{3.2}$$

Next, let us consider equation (3.1) in terms of $\bar{X}_n$

$$\theta_n^*(1-3\bar{X}_n^2)=\bar{X}_n-\bar{X}_n^3-3\bar{X}_n(\theta_n^*)^2+(\theta_n^*)^3, \tag{3.3}$$

which is an equation of third degree and hence there exists at least one real root. From (3.2) we have

$$\mathrm{P}\left(\lim_{n\to\infty}(1-3\bar{X}_n^2)\right)=1-3\mathrm{E}(X_i)^2\geq 0,$$

so, by taking limits in both sides of (3.3), we get the equation

$$(1-3\mathrm{E}(X_i)^2)\theta^*=(\theta^*)^3, \tag{3.4}$$

whose three roots

$$\theta_1^*=0, \qquad \theta_2^*=\sqrt{1-3\mathrm{E}(X_i)^2}, \qquad \theta_3^*=-\sqrt{1-3\mathrm{E}(X_i)^2}, \tag{3.5}$$

are not random, where $\mathrm{P}\left(\lim_{n\to\infty}\theta_n^*\right)=\theta^*$. Clearly, the roots θ_i^* of (3.5) are very close to the roots θ_{ni}^* of (3.4).

Now, if we consider once again equation (3.3) we can write

$$\sqrt{n}\theta_{n1}^*=\frac{\sqrt{n}\left[\left(\bar{X}_n-\bar{X}_n^3\right)-3\bar{X}_n\theta_{n1}^{*2}\right]}{1-3\bar{X}_n^2-\theta_{n1}^{*2}}. \tag{3.6}$$

It is evident that the numerator of (3.6) is asypmtotically normal distributed with parameters

$$\mu_n=0, \quad \sigma_n^2=\mathrm{E}(X_i-X_i^3)^2.$$

We also have

$$\mathrm{P}\left(\lim_{n\to\infty}(1-3\bar{X}_n^2-\theta_{n1}^{*2})\right)=1-3\mathrm{E}(X_i)^2=\begin{cases}0 & \text{if } k=1,\\ \dfrac{1}{4} & \text{if } k=\frac{3}{2},\\ \dfrac{2}{5} & \text{if } k=2.\end{cases} \tag{3.7}$$

What do this result mean? If $k=1$, then

$$\theta_{n1}^*=\sqrt[3]{-\bar{X}_n+\bar{X}_n^3+3\bar{X}_n\theta_{n1}^{*2}+\theta_{n1}^{*2}(1-3\bar{X}_n^2)},$$

so

$$n^{1/6}\theta^*_{n1} = \sqrt[3]{-\sqrt{n}(\bar{X}_n - \bar{X}^3_n) + 3\hat{E}\theta^{*2}_{n1}\sqrt{n}\bar{X}_n + \sqrt{n}(1 - 3\bar{X}^2_n)\theta^*_{n1}}.$$

On the other hand, since

$$\mathrm{P}\left(\lim_{n\to\infty}\theta^*_{n1}\right) = 0 \quad \text{and} \quad \mathrm{P}\left(\lim_{n\to\infty}\theta^{*2}_{n1}\right) = 0,$$

we obtain

$$\mathrm{P}\left(\lim_{n\to\infty}\theta^{*2}_{n1}\sqrt{n}\bar{X}_n\right) = 0, \quad \mathrm{P}\left(\lim_{n\to\infty}\sqrt{n}(1 - 3\bar{X}^2_n)\theta^*_{n1}\right) = 0$$

and hence we have

$$\begin{aligned}\lim_{n\to\infty}\mathrm{P}\left(\left\{n^{1/6}\theta^*_{n1} < x\right\}\right) &= \lim_{n\to\infty}\mathrm{P}\left(\left\{\sqrt{n}\theta^{*3}_{n1} < x^3\right\}\right)\\ &= \lim_{n\to\infty}\mathrm{P}\left(\left\{-\sqrt{n}(\bar{X}_n - \bar{X}^3_n) < x^3\right\}\right)\\ &= \Phi\left(x^3/\sqrt{\mathrm{E}(X_i - X_i^3)^2}\right).\end{aligned} \tag{3.8}$$

Since

$$\mathrm{E}(X_i - X_i^3)^2 = \mathrm{E}(X_i)^2 - 2\mathrm{E}(X_i)^4 + \mathrm{E}(X_i)^6,$$

from (2.2) we find that

$$\mathrm{E}(X_i - X_i^3)^2 = \begin{cases} \dfrac{8}{105} & \text{if } k = 1,\\ \dfrac{5}{64} & \text{if } k = 1.5,\\ \dfrac{8}{105} & \text{if } k = 2.\end{cases} \tag{3.9}$$

Hence from (3.8) and (3.9) it follows that if $k = 1$ then the sequence $\{n^{1/6}\theta^*_{n1}\}$ converges in distribution as $n \to \infty$ to a random variable $Z^{1/3}$, where Z is normal $N(0, 8/105)$. Thus

$$\lim_{n\to\infty}\mathrm{P}\left(\{n^{1/6}\theta^*_{n1} \le x)\}\right) = \Phi\left(x^3/\sqrt{\mathrm{E}(X_i - X_i^3)^2}\right).$$

With the help of (3.8) and (3.9) it follows that

$$\mathrm{E}(n^{1/6}\theta^*_{n1})^2 = n^{1/3}\mathrm{E}(\theta^{*2}_{n1}) \cong 3\sqrt{\frac{105}{8}}\int_{-\infty}^{\infty}\varphi\left(\sqrt{\frac{105}{8}}x^3\right)x^4dx < \infty,$$

where $\varphi(x) = \Phi'(x)$, i.e., when $k = 1$

$$\mathrm{E}(\theta^{*2}_{n1}) = O\left(\frac{1}{n^{1/3}}\right). \tag{3.10}$$

On the other side if $1 < k \le 2$, then $\mathrm{E}(1 - 3X_i^2) > 0$ and hence the sequence $\{\sqrt{n}\theta^*_{n1}\}$ converges in distribution as $n \to \infty$ to a random variable Z normal $N(0, \sigma^2)$, where $\sigma^2 = [\mathrm{E}(X_i - X_i^3)^2]/[\mathrm{E}(1 - 3X_i^2)]^2$. In particular, from (3.7) and (3.9) (see also Table 1) we obtain

$$\frac{\mathrm{E}(X_i - X_i^3)^2}{\left[\mathrm{E}(1 - 3X_i^2)\right]^2} = \begin{cases} \dfrac{5}{64}\Big/\left(\dfrac{1}{4}\right)^2 = \dfrac{5}{4} & \text{if } k = 1.5, \\ \dfrac{8}{105}\Big/\left(\dfrac{2}{5}\right)^2 = \dfrac{10}{21} & \text{if } k = 2. \end{cases} \tag{3.11}$$

A simple comparison of (3.10), (3.11) and (2.5) shows that $\bar{X}_n$ is better than Euler's estimator. Finally, note the different rates of convergence of the estimators $\hat{\theta}_n$ and θ^*_{n1}.

4 Bernoulli's (ML) estimator

Let us consider the statistic

$$T = \frac{X_{(1)} + X_{(n)}}{2},$$

as was done, for example, by Voinov and Nikulin (1993, pp. 47-51). Clearly $\mathrm{E}(T = 0)$ and from (2.2) it follows that

$$\mathrm{Var}(T = 2)\left[\frac{\Gamma(k)\Gamma(k+1)}{\Gamma(2k)}\right]^{2/k}\left[\Gamma\left(\frac{2}{k}+1\right) - \Gamma^2\left(\frac{1}{k}+1\right)\right]\frac{1}{n^{2/k}}(1 + o(1)). \tag{4.1}$$

From (4.1) with large values of n we have

$$\mathrm{Var}(T) \sim \begin{cases} \dfrac{2}{n^2} < \frac{1}{n} & \text{if } k = 1, \\ \dfrac{1}{18}\left(\dfrac{3\pi}{2}\right)^{4/3}\left[\Gamma\left(\dfrac{1}{3}\right) - \Gamma^2\left(\dfrac{2}{3}\right)\right]\dfrac{1}{n^{4/3}} < \dfrac{1}{4n} & \text{if } k = 3/2, \\ \dfrac{2}{3}\left(1 - \dfrac{\pi}{4}\right)\dfrac{1}{n} < \dfrac{1}{5n} & \text{if } k = 2. \end{cases} \tag{4.2}$$

Since $\mathrm{Var}(X_{(n)}) = \mathrm{Var}(X_{(1)}) = O(n^{-2/k})$, we also have

$$\mathrm{Var}(\hat{\theta}_n) = O(n^{-2/k}). \tag{4.3}$$

Now it is clear (compare (4.2), (4.3), and (2.5)), that Bernoulli's estimator $\hat{\theta}_n$ is better than $\overline{X}_n$ and θ^*_{n1}. It is also clear that, in practice, for $1 \le k \le 2$ it is reasonable to use the above statistic T to estimate θ.

Acknowledgements

The author would like to thank Professors T. Smith (Queen's University, Kingston, Canada), C. M. Cuadras (University of Barcelona, Spain), and V. Solev (Steklov Mathematical Institute, St Petersburg, Russia) for helpful discussions and remarks.

Bibliography

Bernoulli, D. (1777). Memoirs of the Academy of St. Petersburg, *Acta Acad. Petrop.*, 3-33.

Bernoulli, D. (1961). The most probable choice between several discrepant observations and the formation therefrom of the most likely induction, *Biometrika*, 1961, 48, 3-13.

Euler, L. (1961). Observation on the foregoing dissertation of Bernoulli, *Biometrika*, 1961, 48, 13-18.

Huber, C., Nikulin, M. (1997). Remarques sur le maximum de vraisemblance. *Qüestiió*, 21, 37-58.

Kendall, M. G. (1961). Studies in the history of probability and statistics XI. Daniel Bernoulli and maximum likelihood, *Biometrika*, 48, 1-18.

Stigler, S. M. (1997). Daniel Bernoulli, Leonhard Euler, and maximum likelihood (including a new translation of a paper by D. Bernoulli). In: *Festschrift for Lucien Le Cam*, (Eds. David Pollard, Eric Torgensen, Grace L. Yang), Springer. New York, pp. 345-357.

Voinov, V., Nikulin, M. (1993). *Unbiased Estimators and their Applications*. Vol. 1: *Univariate Case*, Kluwer Academic Publishers. Dordrecht.

SORT 30 (1) January-June 2006, 101-122

Improving small area estimation by combining surveys: new perspectives in regional statistics*

Alex Costa[1], Albert Satorra[2] and Eva Ventura[2]

[1] *Statistical Institute of Catalonia (IDESCAT)* [2] *Universitat Pompeu Fabra*

Abstract

A national survey designed for estimating a specific population quantity is sometimes used for estimation of this quantity also for a small area, such as a province. Budget constraints do not allow a greater sample size for the small area, and so other means of improving estimation have to be devised. We investigate such methods and assess them by a Monte Carlo study. We explore how a complementary survey can be exploited in small area estimation. We use the context of the Spanish Labour Force Survey (EPA) and the Barometer in Spain for our study.

MSC: 62J07, C2J10, 62H12

Keywords: composite estimator, complementary survey, mean squared error, official statistics, regional statistics, small areas

1 Introduction

The 1978 Spanish constitution established that the Spanish Government has the exclusive competence over the statistics that are of interest for the whole country (Article 149.1.31a). At the same time, the statutes of the autonomous communities (regions) state that the regional administrations have the exclusive competence over the statistics that

* The authors are grateful to Xavier López and Maribel Garcia, statisticians at IDESCAT, for their help at several stages of this research. The comments by Nicholas T. Longford on a previous version of this paper are very much appreciated.

Address for correspondence: Albert Satorra and Eva Ventura. Department of Economics and Business. Universitat Pompeu Fabra.

Received: February 2006

Accepted: June 2006

are of interest to the region (e.g., Article 33, 1979, Statute of Catalonia).[1] These laws lead to an interesting overlap of competences, since many surveys and administrative registers are of interest to both the country and the regions.

The country's official statistics bureaux have a much longer-standing tradition and greater resources, and are usually in charge of producing survey-based statistics with a country-wide scope. However, the statistics produced at the country level are sometimes not satisfactory for the region. This may arise for three reasons: 1) an issue that is relevant for the region but not for the country is not reported by the survey; 2) data collected at the country level is not reliable at the regional level; 3) statistics collected at the country level may not provide reliable information for small areas of a region.

For an example of the first case, consider the tourism surveys conducted in Catalonia, for which information on cross-border day trips to France or Andorra is highly relevant. The general questionnaire for the Spanish surveys does not include information about these trips since for the country as a whole these trips are of little importance. An example of the second case is that despite being over-represented in the National Labour Force Survey (EPA), less populous regions, such as Murcia and Navarra, still have too small subsample sizes to make any reliable inferences about them. The third problem is that territorially disaggregated information at the county or municipality levels, or for small islands (Isla de Hierro in the Canary Islands, or Formentera in the Balearic Islands, for example), is very important for regions, but at the overall country level they have a much lower priority.

A regional statistics office could conduct a similar survey, duplicated and improved for its purpose, but that would amount to wasting resources and would increase the burden on respondents. Some subjects (companies, households, and the like) would receive virtually identical survey questionnaires, so they are likely to develop an impression that the national and regional statistical offices do not coordinate their activities. In addition to inducing a negative attitude towards official statistics, duplication of the respondents' costs might be unacceptable.

As an alternative, regional statistical offices may ask the National Statistics Institute (INE) to modify its survey design to meet the regions' needs: to expand the questionnaire to cover issues of regional interest, or to increase the sample size to achieve sufficient precision for the inferences of interest. These changes would not cause problems in sporadically conducted surveys. An example is the Survey of Time Usage conducted in Spain on a single occasion in 2004. INE agreed with some regions to increase the subsample size in their areas. This option may not be available in some ongoing (annual, or quarterly) INE surveys that do not meet the regions' needs due to problems related to reliability or territorial disaggregation. Reasons of technical, legal, or professional nature make modifications of the design of the ongoing surveys problematic. The national offices could not cope with the myriad of requests of various kinds from the

1. Or the Article 135, 2006, of the recently approved Statute of Catalonia.

regional offices. In this paper, we investigate an analytical solution to these problems that is based on supplementing the country survey with auxiliary information available for some of the small areas of interest.

The use of auxiliary information is not a new idea in small area estimation. When the direct estimator for a particular small area is not satisfactory, one may resort to an indirect estimator. The direct estimator uses only information or data from the area and the variable of interest. Direct estimators are usually unbiased, though they may have large variances. An indirect estimator uses information from the small area of interest as well as from other areas and other variables, or even from other data sources. Indirect estimators are based on implicit or explicit models that incorporate information from other sources. For example, information obtained in a survey can be combined with the one collected in a census or an administrative register. Indirect estimators are usually biased, although their variances are smaller than those of the direct (unbiased) estimators, and the trade-off of bias and variance is usually in their favour.

The novelty of our approach is that we use the information of an auxiliary survey instead of census or administrative records. We combine the information of a country-wide survey, called the reference survey (RS), with the information from a complementary survey conducted by the regional statistics office and tailored to the specific needs of the small area.

A complementary survey (CS) is conducted at the regional level and records variables that correlate with the variables in the RS. CS covers one or several regions of the country, or part of a region. We regard CS as a "light survey" since the data will be faster and cheaper to collect than for the RS. For example, in the case of unemployment, a subject in the CS identifies him or herself as unemployed by the response to a single question. In contrast, the RS follow certain guidelines set forth by the International Labour Organization to classify the subject as unemployed (actively searching for work, available to begin working immediately, and so forth), employed and economically inactive. CS can also simplify the process of contacting the subjects (persons, companies, households, etc.) by using telephone contact systems (Computer Assisted Telephone Interviewing, CATI) or other automated survey methods. So, CS provides results similar to those of RS at a much lower cost; however, as CS records the values of a slightly different variable than RS, its results are biased. This is the price for the less elaborate questionnaire, with looser wording.

We differentiate three types of CS:

1) A general complementary survey (GCS) covers all the regions of the country at one or several points in time. With data from many areas we can remove the bias of GCS estimators relative to RS. One example of GCS is the Economically Active Population Survey (EPA) conducted by INE as RS, and the Barometer of Spain conducted by the Centre for Sociological Research (CIS) as GCS. In the Barometer, respondents are asked if they are unemployed. Information from the EPA and CIS is available for all the Spanish regions for several years.

2) With a regional complementary survey (RCS) we can assess the bias at the regional level, but not at the small area level because there are no data to compare RS and RCS at the small area level. An example is the Survey on Information and Communication Technologies in Catalonia, conducted by the Statistical Institute of Catalonia (IDESCAT) (RCS) by means of CATI, complementing an equivalent RS conducted by INE. RS has a clustered sampling design in which the 41 counties of Catalonia are not well represented. In contrast, the design of the RCS ensures an even coverage of the counties. In this example, the bias of RCS cannot be disaggregated to counties.

3) A local complementary survey (LCS) is a survey conducted in a specific small area. The bias is unknown since RS does not produce valid results for this small area. One example is a survey similar to the EPA in a single small municipality in Catalonia which has very sparse or no representation in EPA.

The bias of the survey relative to RS can be explicitly modeled in GCS but not in RCS or LCS. We investigate how information from CS can be integrated with RS for making inferences about small areas. We consider the specific context in which EPA is RS and the Barometer is CS, in this case, a GCS. The Barometer contains a few questions regarding the subject's employment status which are at face value highly correlated with the corresponding variable in EPA.

The accuracy in small area estimation can be increased by: *a)* increasing the sample size in the area of interest; *b)* borrowing strength from neighbouring areas (using indirect or composite estimators); *c)* borrowing strength from CS, especially when the variables in RS and CS are highly correlated. We explore all these alternatives, with emphasis on combining the options b) and c). The performance of the estimators and the contribution of the complementary information will be assessed by simulation.

Parallel work on the use of CS has been conducted by Costa *et al.* (2006), who study the Survey on the Uses of Information and Communication Technologies in Catalan households; INE conducts a country-wide survey while IDESCAT is in charge of the RCS.

The present paper is organized as follows. Section 2 reviews established small area estimators, with emphasis on estimating labour statistics. Section 3 describes the specific context of estimating rates of unemployment in Spain. Section 4 assesses the performance of the alternative small area estimators by simulations. Section 5 summarizes the main findings of the paper.

2 Estimators for small areas

In this section we consider a two-stage clustered sampling design, motivated by the sampling design of EPA that is considered later in the paper. We consider a binary

variable Y that takes the values $Y_{ij} = 1$ if the characteristic under study is present for subject ij, and $Y_{ij} = 0$ otherwise. Here i ($i = 1, 2, \ldots, n$) and j ($j = 1, 2, \ldots, m_i$) denote *primary sample unit* (PSU) and *secondary sampling unit* (SSU), respectively. We use the convention that capitals (X, Y) denote population values and lowercases (x, y) sample values. Their indexing is implied; that is, in X_{ij} we use population indexing and in x_{ij} we use sample indexing. For every sample, we have a variable W of sampling weights, with w_{ij} representing the sampling weight of subject ij.

The population is divided in K small areas, indexed by $k = 1, 2, \ldots, K$. We use the notation $Y_{k.ij}$ for the values of variable Y on units of area k. For sampling data, the symbol + in the subscript denotes the weighted summation over the sample; for example, $y_{k,i+} = \sum_{j=1}^{m_i} w_{k,ij} y_{k,ij}$. For population data, the symbol + indicates summation without weighting.

Our target is the population ratio $\theta_k = Y_{k,+}/X_{k,+}$ of two totals, for each area k. We consider also the overall population ratio $\theta = Y_+/X_+$. It is assumed that the denominator is positive. Several estimators are considered.

2.1 Direct estimator

A direct estimator of θ_k uses only data from area k. It is defined as

$$\hat{\theta}_k = \frac{\hat{Y}_k}{\hat{X}_k} ,$$

where $\hat{Y}_k = y_{k+}$ and $\hat{X}_k = x_{k+}$. Here the summation extends only over the (say, n_k) PSUs that intersect with area k (we assume $n_k > 2$). Straightforward application of the delta-method yields the following estimator of variance $\mathrm{V}(\hat{\theta}_k)$

$$\hat{V}(\hat{\theta}_k) = \frac{1}{\hat{X}_k^2} \left\{ \hat{V}(\hat{X}_k) - 2\hat{\theta}_k \, \widehat{\mathrm{cov}}(\hat{Y}_k, \hat{X}_k) + \hat{\theta}_k^2 \, \hat{V}(\hat{X}_k) \right\}, \tag{1}$$

where

$$\widehat{\mathrm{cov}}(\hat{Y}_k, \hat{X}_k) = \frac{n_k}{n_k - 1} \sum_{i=1}^{n_k} (z_{k,i}{}^{(y)} - \bar{z}^{(y)})(z_{k,i}{}^{(x)} - \bar{z}^{(x)}), \tag{2}$$

$$z_{k,i}{}^{(y)} = y_{k,i+} \quad \text{and} \quad \bar{z}^{(y)} = n_k^{-1} \sum_{i=1}^{n_h} z_{k,i}{}^{(y)},$$

and similarly for x. We compute $\hat{V}(\hat{X}_k)$ and $\hat{V}(\hat{Y}_k)$ as $\widehat{\mathrm{cov}}(\hat{X}_k, \hat{X}_k)$ and $\widehat{\mathrm{cov}}(\hat{Y}_k, \hat{Y}_k)$, respectively.

In the general case of L strata, the sample values are $y_{h,ij}$, were h denotes strata. The direct estimator of the overall population ratio $\theta = Y_+/X_+$ is $\hat{\theta} = \hat{Y}/\hat{X}$, where $\hat{Y} = y_{+,++}$

and $\hat{X} = x_{+,++}$ (summation over strata, PSUs within the strata, and units within the PSU). The estimator of var $(\hat{\theta})$ is like (1) and (2) with subscript k suppressed and a summation over L strata added to the right hand side of (2).

Information on population totals of some auxiliary variables would allow us to calculate post-stratified or ratio-estimators. This will not be pursued in this study. For more information on those estimators, the reader can consult Rao (2003), Ghosh and Rao (1994) or Mancho (2002). López (2000) considers some of these estimators in the context of small area estimation for EPA in Canary Islands. We consider $\hat{\theta}_k$ and $\hat{\theta}$ as the only direct estimators in this study.

2.2 Small area estimators without auxiliary information

An indirect estimator of θ_k uses data from outside area k. As an alternative to $\hat{\theta}_k$ we may adopt the overall-country direct estimator $\hat{\theta} = \hat{Y}/\hat{X}$ for every area k. This is an indirect estimator. Being based on much more data than $\hat{\theta}_k$, $\hat{\theta}$ has a much smaller variance than $\hat{\theta}_k$, but is biased for θ_k, unless the θ_k's are all equal. In this case, $\hat{\theta}$ is much more efficient than $\hat{\theta}_k$. But if the θ_k's vary substantially across areas, the bias of $\hat{\theta}$ will be large, and so will be its mean squared error (MSE). An attractive alternative estimator to both $\hat{\theta}_k$ and $\hat{\theta}$ is the composite estimator $\hat{\theta}_k^{(c)}$ defined as the convex combination

$$\hat{\theta}_k^{(c)} = \phi_k \hat{\theta} + (1 - \phi_k)\hat{\theta}_k \tag{3}$$

with $0 \leq \phi_k \leq 1$. The coefficient ϕ_k is chosen so as to minimize the MSE and is equal to

$$\phi_k = \frac{\text{var}\,(\hat{\theta}) - \text{cov}\,(\hat{\theta}_k, \hat{\theta})}{(\theta_k - \theta)^2 + \text{var}\,(\hat{\theta}_k) + \text{var}\,(\hat{\theta}) - 2\,\text{cov}\,(\hat{\theta}_k, \hat{\theta})}\,. \tag{4}$$

The denominator of (4) is positive; in fact, it is equal to $E\{(\hat{\theta}_k - \hat{\theta})^2\}$. Clearly, ϕ_k depends on some unknown parameters, and itself has to be estimated. Since

$$\hat{\theta} = \frac{\sum_{k=1}^{K} \hat{Y}_k}{\hat{X}} = \frac{\sum_{k=1}^{K} \hat{X}_k \hat{\theta}_k}{\hat{X}} = \sum_{k=1}^{K} q_k \hat{\theta}_k,$$

where $q_k = \hat{X}_k/\hat{X}$, $\text{cov}\,(\hat{\theta}_k, \hat{\theta}) = \sigma_k^2 q_k$, so the optimal weight is

$$\phi_k = \frac{\sigma_k^2(1 - q_k)}{(\theta_k - \theta)^2 + \sigma_k^2(1 - 2q_k) + \sigma^2}\,, \tag{5}$$

where σ_k^2 and σ^2 are the respective sampling variances of $\hat{\theta}_k$ and $\hat{\theta}$ (see Longford, 1999).

When q_k is very small and the survey is large (e.g., the number K of small areas is

large and the sample sizes of most of them are small), we can ignore both q_k and the variance σ^2; then, ϕ_k is approximated by

$$\phi_k = \frac{\sigma_k^2}{(\theta_k - \theta)^2 + \sigma_k^2} .$$

We could estimate σ_k^2 from the sample data from area k and use $(\hat{\theta}_k - \hat{\theta})^2$ as an estimator of the denominator in (5). Our experience, shows that this results in a very unstable estimator of ϕ_k (see Costa, Satorra and Ventura 2003, 2004). One way to overcome this difficulty is by averaging the estimators $\hat{\sigma}_k$'s among several areas (or several variables). For example, Purcell and Kish (1979) use a weight common to all areas that minimizes the between-area average of the mean squared errors. By assuming that the within-area variances of Y are equal, the pooled estimator of their common variance is

$$\hat{\sigma}_w^2 = \frac{1}{n - K} \sum_{k=1}^{K} (n - 1)\hat{\sigma}_k^2 . \tag{6}$$

By using the following estimator of the square of the bias

$$b^2 = \frac{1}{K} \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta})^2, \tag{7}$$

and ignoring both q_k and the var $(\hat{\theta})$, we estimate the weight as

$$\hat{\phi}_k = \frac{\hat{\sigma}_w^2}{\hat{\sigma}_w^2 + b^2} . \tag{8}$$

Expressions (3) and (8) define the *classic composite* estimator (Costa, Satorra and Ventura, 2003).

2.3 Auxiliary information

In the literature on small area estimation, we find many indirect estimators that incorporate auxiliary information. Usually this information consists of data from a census or an administrative register. Rao (2003) describes the regression synthetic estimator, which combines direct estimators with those obtained from census or records, encompassing a large variety of estimators, such as ratio or count estimators.

We use auxiliary information that arises from a CS carried out in each of several small areas. In particular, we assume that for each of a set of small areas we have the direct estimator $\hat{\theta}_k$ as well as an estimator $\hat{\delta}_k$ derived from CS. For these estimators,

consider the simple regression equation

$$\hat{\theta}_k = \alpha + \beta\hat{\delta}_k + \epsilon_k, \tag{9}$$

$k = 1, 2, \ldots K$, and the *fitted estimator* of θ_k by the OLS regression,

$$\hat{\theta}_k^F = \hat{\alpha} + \hat{\beta}\hat{\delta}_k \, .$$

When historical data are available for the RS and CS across several areas, the fitted estimator $\hat{\theta}_k^F$ could be based on more advanced regression than just simple OLS. If we have RS and CS at several time points ($t = 1, \ldots, T$) and several areas ($k = 1, 2, \ldots, K$), we could estimate θ_k by an analysis of covariance model. The regression could also involve other covariates. As more variables are incorporated into the regression that links $\hat{\theta}_k$ with $\hat{\delta}_k$, the synthetic estimator $\hat{\theta}_k^F$ will be more efficient for θ_k, although using too many covariates may inflate the sampling variance. For simplicity, we only consider the OLS regression (9). In the Monte Carlo set-up of Section 4, however, we also involve an estimator that is based on a covariate (fixed-effects, FE) regression model that serves as a benchmark for maximum information attainable from CS.

Even though the variance of $\hat{\theta}_k^F$ may be substantially smaller than var $(\hat{\theta}_k)$, $\hat{\theta}_k^F$ may be biased. We improve both $\hat{\theta}_k$ and $\hat{\theta}_k^F$ estimators by considering the composite estimator

$$\hat{\theta}_k^{(c)}(CS) = \phi_k\hat{\theta}^F + (1 - \phi_k)\hat{\theta}_k \tag{10}$$

where

$$\phi_k = \frac{\text{var}\,(\hat{\theta}_k) - \text{cov}\,(\hat{\theta}_k^F, \hat{\theta})}{\Delta_k^2 + \text{var}\,(\hat{\theta}_k) - \text{var}\,(\hat{\theta}_k^F)} \tag{11}$$

Here $\Delta_k = \theta_k - E(\hat{\theta}_k^F)$ has to be estimated. If α and β were known, var $(\hat{\theta}_k^F) = \beta^2$ (var $\hat{\delta}_k$). When the regression parameters are estimated, then

$$\text{var}\,\hat{\theta}_k^F = E\left(\text{var}\,\hat{\theta}_k^F \mid \hat{\Theta}\right) + \text{var}\,\left(E(\hat{\theta}_k^F \mid \hat{\Theta})\right),$$

where $\hat{\Theta} = (\hat{\alpha}, \hat{\beta})$ stands for the vector of estimated regression coefficients. Since the expected value of $(\hat{\theta}_k - \hat{\theta}_k^F)^2$ coincides with the denominator in (11), the weight of (11) is estimated as

$$\hat{\phi}_k = \frac{\hat{\sigma}_k^2 - \tilde{\sigma}_k^2}{(\hat{\theta}_k - \hat{\theta}_k^F)^2},$$

where $\hat{\sigma}_k^2$ and $\tilde{\sigma}_k^2$ are the respective estimators of the variances of $\hat{\theta}_k$ and $\hat{\theta}_k^F$. An alternative estimator of ϕ_k is more stable,

$$\hat{\phi}_k = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + b^2}, \tag{12}$$

where $\hat{\sigma}^2$ is given by (6) and

$$b^2 = \frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta}_k^F)^2. \tag{13}$$

The estimator $\hat{\theta}_k^c(CS)$ defined by equations (10), (12) and (13) is called the composite complementary survey (CCS) estimator based on OLS regression (or CCS-OLS).

In Section 4, we assess by Monte Carlo the efficiency of the CCS, direct, indirect and classic composite small area estimators. The efficiency of the CCS estimator will be compared against a direct estimator based solely on RS but with sample size increased by r%, with $r = 10, 20, 50, 100$. For the Monte Carlo study we construct an artificial population that resembles Spain in some aspects related to the labour force. The EPA and Barometer Surveys are used for this construction.

3 EPA and Barometer surveys

This section describes general aspects of estimation of unemployment rates in Spain, both at country and area levels. The main source of information about unemployment in Spain is the EPA conducted by the INE, our RS. The Barometer is our CS.

EPA is a quarterly survey that uses a lengthly questionnaire, panel design and face-to-face interview (paper and pencil, PAPI); in contrast, the Barometer is a monthly survey that uses CATI as the interviewing system and an entirely new sample each month. EPA is designed for reliable estimation of several labour market statistics, including the unemployment rate at country level. The Barometer uses the self-perceived labour status of the interviewed individuals as a proxy for unemployment status.

While all the provinces of Spain are represented in the EPA, this is not always the case for the Barometer. Even if a province is represented in the survey, its sample size may be very small. To have a more informative study, we grouped the 50 Spanish provinces (plus the two autonomous cities in the north of Africa) into 25 areas, according to their geographical proximity and similarity of their labour markets. With this grouping, each area is represented by at least 140 observations in CS. The second and the sixth column of Table 1 shows the sample sizes of EPA and CIS surveys across the 25 areas for the fourth quarter of 2003. To give the same time frame to both surveys, the quarterly estimate by Barometer is taken to be the average across three months of the monthly unemployment rates of the Barometer.

While EPA follows the standard International Labour Organization methodology, the Barometer simply asks the subjects how they perceive their employment status. The unemployment rates in the two surveys differ for two reasons: different definitions of

the target variables and different samples (sampling error).

Table 1 shows the direct estimates of unemployment rates for the 25 areas for men, women and for all adults. This distinction is important because in Spain the labour market participation of women is lower than of men, so the properties of the estimators for these two categories may differ. The table shows that the estimates of the unemployment rates differ considerably between the two surveys; see, for example, Asturias and Navarra-Rioja. The sample correlation of the sets of estimates are 0.54, 0.66 and 0.26, for the total (T), men (M) and women (W), respectively. Such large correlations supports the use of the Barometer as complementary information for the EPA.

Table 1: *Sample sizes and unemployment rates (in %) for the EPA and CIS surveys (fourth quarter of 2003; Ss = Sample size, T = Total, M = Men, W = Women).*

	EPA survey				CIS survey			
	Ss	T	M	W	Ss	T	M	W
Almería-Granada (AGR)	4 864	15.65	10.22	23.38	324	13.96	12.92	16.30
Málaga(MAL)	3 441	17.33	13.68	23.08	235	18.59	11.90	29.13
Cádiz-Huelva (CHU)	5 287	23.51	18.37	31.97	246	34.24	32.35	38.98
Córdoba-Jaén (CJA)	6 640	18.56	12.83	28.27	237	20.19	16.53	24.37
Sevilla (SEV)	6 411	17.19	12.80	24.01	248	16.74	11.47	25.29
Aragón (ARA)	6 589	6.20	3.71	9.94	232	11.58	7.60	18.26
Asturias (AST)	4 522	10.03	7.00	14.35	218	23.20	12.31	36.59
Baleares (BAL)	3 539	9.38	8.50	10.59	141	13.38	6.41	24.65
Canarias (CAN)	7 748	12.10	9.37	16.10	297	18.37	9.20	33.08
Cantabria (CNT)	3 578	10.32	8.06	13.77	102	22.52	11.50	38.62
Albacete-C. Real (ACR)	4 971	9.28	4.80	16.59	150	28.14	18.89	45.28
Cuenca-Guad.-Tol. (CGT)	6 253	9.89	5.58	17.44	173	12.86	7.44	22.12
Castilla-León (CLE)	15 143	10.91	6.09	18.37	491	11.85	9.50	16.80
Barcelona (BCN)	7 448	9.54	7.37	12.47	919	13.69	11.44	16.44
Gerona-Lérida-Tarr. (GLT)	7 721	7.00	5.09	9.62	259	10.90	5.55	18.04
Alicante-Castellón (ACS)	6 405	10.38	8.16	13.67	345	20.04	18.07	23.28
Valencia (VAL)	5 858	10.08	7.20	14.20	393	20.84	12.85	31.55
Extremadura (EXT)	6 167	17.11	12.51	24.75	201	22.06	13.41	40.97
La Coruña (LCO)	3 472	15.44	10.14	22.17	241	21.11	15.35	29.19
Lugo-Orense-Pont. (LOP)	6 921	11.99	7.73	17.55	293	18.26	15.98	21.83
Madrid (MAD)	7 765	7.00	5.47	9.08	966	15.62	10.51	21.20
Murcia (MUR)	4 043	10.49	7.09	15.87	198	14.03	12.16	18.44
Navarra-Rioja (NRI)	5 362	5.99	4.36	8.45	151	16.81	6.29	30.56
Álava-Guipúzcoa (AGU)	4 489	7.06	5.48	9.28	194	15.21	6.82	26.89
Vizcaya (VIZ)	3 037	11.43	10.06	13.28	212	19.30	17.21	21.47

Using historical data of the EPA and the Barometer (CIS) surveys of unemployment rates, we fit – for men, women, and all adults, in turn – the following regression with area-fixed effects:

$$\text{EPA}\%_{kt} = \alpha + u_k + \beta\,\text{CIS}\%_{kt} + \epsilon_{kt}, \tag{14}$$

where $t = 1, 2, \ldots, T$ and $k = 1, 2, \ldots K$ denote the quarter and the small area respectively. Here u_k is a fixed effect and α and β are the intercept and slope, respectively. The model is estimated using data from the first quarter of 2001 to the fourth quarter of 2003. The monthly data of the Barometer has been averaged over quarters. Table 3 shows parameter estimates with standard errors and t-values, as well as the corresponding R^2 fit measures. The area-effects and the estimated u_k's based on this model are reported in Table 4. These regression estimates and area-effects will be used in the Monte Carlo study to obtain a benchmark estimator (CCS-FE) for unemployment rates that combines RS and CS surveys.

4 Monte Carlo study

In this section, we describe the Monte Carlo study that evaluates the performance of the small area estimators described in Section 2. Based on an EPA sample, we consider an artificial population that resembles that of Spain. We undertake estimation of unemployment rates of total adult population, or just male and female. We expect that a small area estimator that uses auxiliary information will outperform the estimators based only on RS. The gains, however, may diminish when the subsample size of RS in the area of interest is large.

4.1 Design of the simulations

We adopt the sample of the EPA survey in Oct.-Dec. 2003 (see Table 1) as our population. We estimate the unemployment rates of men, women and all adults in this population. The 25 target areas are listed in Table 1. In the simulations, CIS has approximately 2550 monthly observations (7650 observations per quarter). In each area we have a small RS sample and typically a large CS sample.

For simplicity, and to focus on the comparison of the estimators, we apply to the adopted EPA population stratified (by area) simple random sampling (with replacement) proportional to the area size. The sample sizes range from 2500 to 25000, but are fixed within replications. The CS sample is drawn from the realized CIS sample (sample size 7650) treated as the population. The sample size is fixed at 7650. Unemployment rates for men and women are estimated from the total sample by considering just the men and

women respectively. Knowing the population values, the MSE's can be estimated with high precision governed by the number of replications.

We use the following RS sample sizes: $5000, 10000, 12500, 25000$, so that the average within-area subsample sizes are $200, 400, 500$ and 1000. Further, the sample for the direct estimator is boosted by $10, 25, 50$ and 100%, but the subsample added is not used in evaluating the other estimators.

For small RS sample sizes, we expect the CCS estimator to be more efficient than the direct estimator, even with a substantially boosted sample size. The specific value of r for which the efficiency of both estimators is similar is likely to vary with the sample size of the RS sample.

The Monte Carlo simulations comprise replications of the following steps.

A) we draw a sample of size m from the population ($N = 147000$). Similarly, we draw a CIS sample. We evaluate the direct, indirect, classic composite, and the composite estimator based on the CS-based composite estimators.

B) We boost the sample size by $r = 10, 25, 50, 100\%$, by drawing an additional subsample from EPA and we evaluate the direct estimator.

Steps A) and B) are replicated 1000 times.

Table 2 summarizes the settings of sample sizes across areas. Our targets are the unemployment rates in the 25 areas, for men, women and all the working force. For each pair of samples, EPA and CIS, we compute the direct, indirect and classic composite, and and two CS-based composite estimators.

Table 2: *Monte Carlo study: average sample sizes across small areas, and sample size increase (in %) of the direct estimator.*

Average Sample size	Sample size increase (%)			
	10	25	50	100
100	110	125	150	200
200	220	250	300	400
400	440	500	600	800
500	550	625	750	1000
1000	1100	1250	1500	2000

For each estimator $\tilde{\theta}$ and area k we evaluate the relative root mean square error as

$$\text{RRMSE}(\tilde{\theta}, k) = \frac{\sqrt{\sum_{j=1}^{1000} \left(\tilde{\theta}_k{}^{(j)} - \theta_k\right)^2 / 1000}}{\theta_k},$$

where $\tilde{\theta}_k^{(j)}$ is the jth replicate of a specific small area estimator of θ_k. Smallest RRMSE is preferred. The average, median and maximum value of RRMSE across areas is recorded for each estimator.

4.2 Incorporating auxiliary information

Denote the RS and CS direct estimators of the target θ_k by $\hat{\theta}_k$ and $\hat{\delta}_k$, respectively. Because the concepts measured in RS and CS differ slightly, $\hat{\delta}_k$ is likely to be biased. In the fitted estimator $\hat{\theta}_k^F$ we use both $\hat{\theta}_k$ and $\hat{\delta}_k$.

In each replication, we fit the OLS regression (9) with the unemployment rates of EPA and CS as the respective $\hat{\theta}_k$ and $\hat{\delta}_k$ estimators, and compute the OLS fitted $\hat{\theta}_k^F$(OLS). Estimators $\hat{\theta}_k^F$(OLS) and $\hat{\theta}_k$ are then combined to obtain the CS-based CCS-OLS estimator. In parallel, we compute the fitted estimator $\hat{\theta}_k^F$(FE) that is based on the regression coefficients and area effects of the FE regression reported in Section 3. Since this estimator uses more information than the others we can expect it to be more efficient. The FE model is taken as a benchmark model for the information of CS on RS, since among all the regression alternatives that we could think of, many will use more information than the simple OLS model, but less than the FE model.

Two CCS estimators are considered, the CCS-OLS which uses the $\hat{\theta}_k^F$(OLS) obtained from the OLS regression, and the CCS-FE which uses the $\hat{\theta}_k^F$(FE) of the FE regression. Only CCS-OLS is feasible in applications, since CCS-FE uses information that will not generally be available. While the CCS-OLS does fit the OLS regression in each replication, the CCS-FE is based on just one regression fit common to the whole Monte Carlo study. Table 3 reports the regression coefficients used in obtaining the CCS-FE. The area fix-effects that arise from the FE regression are reported in Table 4. The alternative of a random-effect regression model was considered. No substantial change of the performance of the CCS estimator was observed.

Table 3: *Parameter estimates, se and t-values of the fixed effect regression, for total* (T), *men* (M) *and women* (W) *unemployment rates. Various R^2s are reported: overall (R^2), within (R_w^2) and between (R_b^2).*

	T			M			W		
Model par.	$\hat{\beta}$	se ($\hat{\beta}$)	t-test	$\hat{\beta}$	se ($\hat{\beta}$)	t-test	$\hat{\beta}$	se ($\hat{\beta}$).	t-test
β	0.06	0.02	3.22	0.02	0.02	1.34	0.05	0.02	2.82
α	10.42	0.34	31.08	7.81	0.20	39.81	15.70	0.43	36.25
R_w^2	0.04			0.01			0.03		
R_b^2	0.69			0.70			0.66		
R^2	0.41			0.32			0.35		

Table 4: *Estimated fixed area effects. (*T = *Total*, M = *Men*, W = *Women).*

	Unemployment rate		
Small area	T	M	W
Almería-Granada	3.68	2.65	5.47
Málaga	3.69	4.22	3.56
Cádiz-Huelva	11.08	8.47	16.40
Córdoba-Jaén	7.77	5.65	12.77
Sevilla	7.98	6.80	10.28
Aragón	–5.63	–4.43	–7.83
Asturias	–2.14	–1.44	–3.07
Baleares	–3.68	–2.10	–6.65
Canarias	–0.75	0.24	–1.99
Cantabria	–1.49	–1.29	–2.07
Albacete-Ciudad Real	–2.68	–2.53	–2.16
Cuenca-Guadalajara-Toledo	–1.23	–2.48	1.50
Castilla-León	–0.79	–1.76	0.65
Barcelona	–1.44	–0.47	–3.53
Gerona-Lérida-Tarragona	–4.13	–3.21	–6.36
Alicante-Castellón	–1.95	–1.26	–3.40
Valencia	–0.72	–0.17	–1.81
Extremadura	5.45	4.09	8.28
La Coruña	0.99	0.78	0.74
Lugo-Orense-Pontevedra	0.03	–0.55	0.04
Madrid	–4.05	–2.93	–6.42
Murcia	–0.35	–0.46	–0.25
Navarra-Rioja	–5.74	–4.39	–8.24
Álava-Guipúzcoa	–4.20	–3.35	–5.93
Vizcaya	0.28	–0.09	0.01

5 Results

Tables 5-7 show results of the simulations for the whole labour force, men and women respectively. The tables have identical layouts giving summaries of RRMSEs for the different estimators (columns) and sample sizes (blocks of rows). For each estimator and sample size, the average, mean and maximum RRMSE across the 25 areas is given. The column at the extreme right contains the RRMSE summaries for the benchmark CCS-FE.

Table 5: *Estimation of total unemployment rates. For different sample sizes, the table shows the RRMSE average (across areas) of the various estimators evaluated in the Monte Carlo study. The estimators are: RS-based estimators (direct with sample size boosted by r%; indirect, $\hat{\theta}$; composite, $\hat{\theta}_k^{(c)}$), and the CCS estimators based on fixed effects (FE) and simple (OLS) regression. All the values of RRMSE have been multiplied by* 1000.

	r%							CS-based	
Summary	$0^{(0)}$	10	25	$50^{(b)}$	100	$\hat{\theta}^{(I)}$	$\hat{\theta}_k^{(c)(1)}$	$OLS^{(2)}$	$FE^{(+)}$
	Average sample size **100**								
Average	419	403	376	345	295	326	323	307	251
Median	415	401	373	347	294	237	312	307	254
Max	578	571	549	501	411	1332	562	550	353
	Average sample size **200**								
Average	298	285	269	245	210	316	251	240	188
Median	298	285	272	239	208	226	236	238	190
Max	421	410	377	350	292	1322	428	454	255
	Average sample size **400**								
Average	210	203	191	172	151	311	191	184	142
Median	211	191	186	164	156	219	179	182	143
Max	299	286	286	242	218	1313	303	335	193
	Average sample size **500**								
Average	190	181	171	158	137	310	174	170	132
Median	188	178	171	157	137	221	167	169	134
Max	264	243	239	229	212	1318	268	305	179
	Average sample size **1000**								
Average	138	132	124	115	100	308	131	129	105
Median	137	132	120	113	097	220	129	127	108
Max	211	209	209	205	193	1319	188	210	140

[(i)] These superscripts show the symbols used in Figures 1 and 2 to represent the RRMSEs.

Of course, the RRMSEs are reduced for the direct estimator when the sample size is boosted. Boosting with $r = 0\%$ is the direct estimator $\hat{\theta}_k$. The efficiency of the composite estimator $\hat{\theta}_k^{(c)}$ is comparable with boosting of the sample by about 50% for small sample size (100), but much less (about 10%) for large sample (1000). That is, the composite estimator $\hat{\theta}_k^{(c)}$ is much more effective for small sample sizes than for large ones; its effectiveness, over the direct estimator, decreases with sample size.

The composite estimator CCS-OLS that makes use of the CS is only slightly more efficient than $\hat{\theta}_k^{(c)}$ for all sample sizes, but the maximum over the areas is slightly higher in most cases. The benchmark CCS-FE estimator (last column of the table), is more efficient than all the other estimators. In fact, its performance is comparable to boosting of the sample by 100%.

Table 6: *Estimation of unemployment rates for men. For different sample sizes, the table shows the RRMSE average (across areas) of the various estimators evaluated in the Monte Carlo study. The estimators are: RS-based estimators (direct with sample size boosted by r%; indirect, $\hat{\theta}$; composite, $\hat{\theta}_k^{(c)}$), and the CCS estimators based on fixed effects (FE) and simple (OLS) regression. All the values of RRMSE have been multiplied by* 1000.

	r%							CS-based	
Summary	$0^{(0)}$	10	25	$50^{(b)}$	100	$\hat{\theta}^{(I)}$	$\hat{\theta}_k^{(c)(1)}$	$OLS^{(2)}$	$FE^{(+)}$
	Average sample size **100**								
Average	669	638	601	546	469	404	478	487	388
Median	645	603	571	525	469	313	447	462	374
Max	915	892	852	776	685	1526	795	803	529
	Average sample size **200**								
Average	475	450	423	385	330	385	371	374	281
Median	470	429	409	372	328	300	347	339	272
Max	663	631	590	558	473	1503	643	658	384
	Average sample size **400**								
Average	335	316	300	270	238	373	285	287	205
Median	325	299	292	263	231	300	266	264	198
Max	492	454	438	390	344	1484	462	495	294
	Average sample size **500**								
Average	298	283	268	247	213	372	259	263	185
Median	294	267	266	233	206	297	250	244	186
Max	411	392	380	347	299	1493	428	454	245
	Average sample size **1000**								
Average	212	204	193	178	153	368	196	199	141
Median	206	195	185	170	150	296	194	190	144
Max	292	286	275	243	214	1488	311	331	184

[(i)] These superscripts show the symbols used in Figures 1 and 2 to represent the RRMSEs.

Similar conclusions are arrived at by inspecting the results for men and women. The RRMSEs for women tend to be larger than for men for a fixed sample size, because their rates or unemployment are higher than for men and RRMSEs are approximately proportional to $\sqrt{p/(1-p)}$, where p is the unemployment rate.

The tables contain a lot of detail that is difficult to digest and do not indicate the performance of the estimators for the individual areas. Figures 1 and 2 display RRMSEs of four small area estimators: direct, marked as 0; indirect, I; composite, 1; and CCS-OLS, 2. It shows also the benchmark estimator, marked as +; and the direct estimator with sample size boosted by 50%, marked as b. We regard estimators 0,1,2 and I as feasible because they use information that would normally be available. Estimators + and b are a benchmark and comparator, respectively. They use information that would not be available in practice. At the outset, I is discarded as competitor of 0, 1 and 2.

Table 7: *Estimation of unemployment rates for women. For different sample sizes, the table shows the average (across areas) of RRMSE of the various estimators evaluated in the Monte Carlo study. The estimators are: RS-based estimators (direct with sample size boosted by r%; indirect, $\hat{\theta}$; composite, $\hat{\theta}_k^{(c)}$), and the CCS estimators based on fixed effects (FE) and simple (OLS) regression. All the values of RRMSE have been multiplied by* 1000.

	r%							CS-based	
Summary	$0^{(0)}$	10	25	$50^{(b)}$	100	$\hat{\theta}^{(I)}$	$\hat{\theta}_k^{(c)(1)}$	$OLS^{(2)}$	$FE^{(+)}$
	Average sample size **100**								
Average	539	514	476	437	378	315	386	365	327
Median	537	512	475	445	376	215	383	345	335
Max	757	736	716	648	540	1164	618	621	443
	Average sample size **200**								
Average	376	361	338	309	267	304	292	273	235
Median	383	368	327	305	266	200	279	265	240
Max	556	521	493	444	377	1155	485	517	331
	Average sample size **400**								
Average	269	259	241	218	190	298	228	210	180
Median	262	261	237	211	192	192	217	210	182
Max	419	374	374	314	269	1147	393	410	267
	Average sample size **500**								
Average	238	230	215	197	172	296	206	194	163
Median	234	228	211	194	170	191	198	195	161
Max	352	324	314	283	250	1151	333	381	232
	Average sample size **1000**								
Average	172	166	155	143	126	293	157	151	129
Median	168	164	152	136	123	190	153	155	130
Max	252	241	238	229	218	1153	239	284	192

[(i)] These superscripts show the symbols used in Figures 1 and 2 to represent the RRMSEs.

The areas are ordered according to their unemployment rates. The diagrams show, for instance, that estimator I has serious weaknesses although it is the most efficient for a few areas in the middle of the range. In general, the composite estimators 1 and 2 are the most efficient among the feasible estimators.

In Figure 1 we have a graphical representation of the RRMSE for the alternative estimators in the case of large sample size (1000). Small areas are on the vertical axis and different symbols represent the different estimators. RRMSE is on the horizontal axis, so that efficiency corresponds to being on the left. For this large sample case, the indirect estimator (I) is generally very inefficient: each RRMSE summary has been truncated at 0.25, except for areas whose unemployment rate is close to the national rate. The composite estimators, the RS-based (1) and CS-based (2) are the most efficient estimators, after the benchmark estimator CCS-FE (+). The composite estimators 1 and

2 (RS and CS based, respectively) are generally as efficient as b, direct estimator with sample size boosted by 50%.

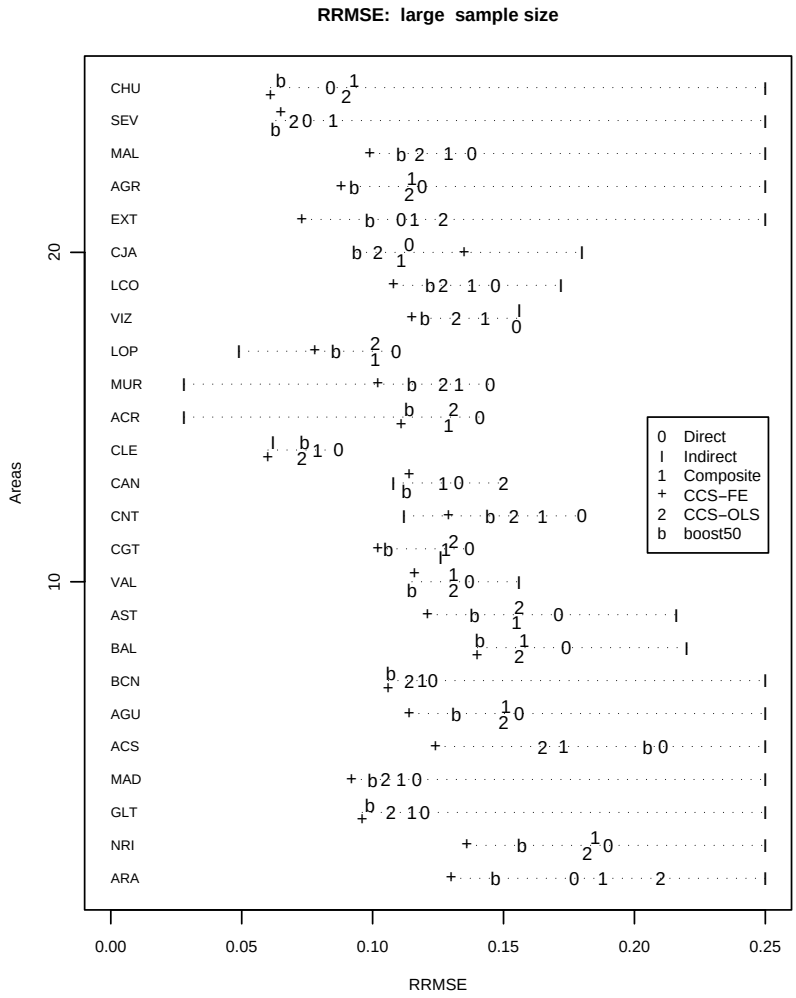

Figure 1: *RRMSEs for the areas and for estimators in the case of large sample (average small area sample size is* 1000*). The areas have been ordered in increasing order of magnitude of their rate of unemployment. Values of RRMSE have been truncated at* 0.25*.*

Figure 2 shows the same information, using the same layout and symbols, for small sample size (100). Again, the indirect estimator (I) is the best for the areas which unemployment value is at the middle range (around the national rate), although they are very inefficient for the areas with extreme rates of unemployment. In those areas, the composite RS-based and the feasible new CS-based estimators (1 and 2, respectively)

are more efficient than the direct estimator b for most of the areas. For most of the areas, 2 is the most efficient, after the benchmark estimator +.

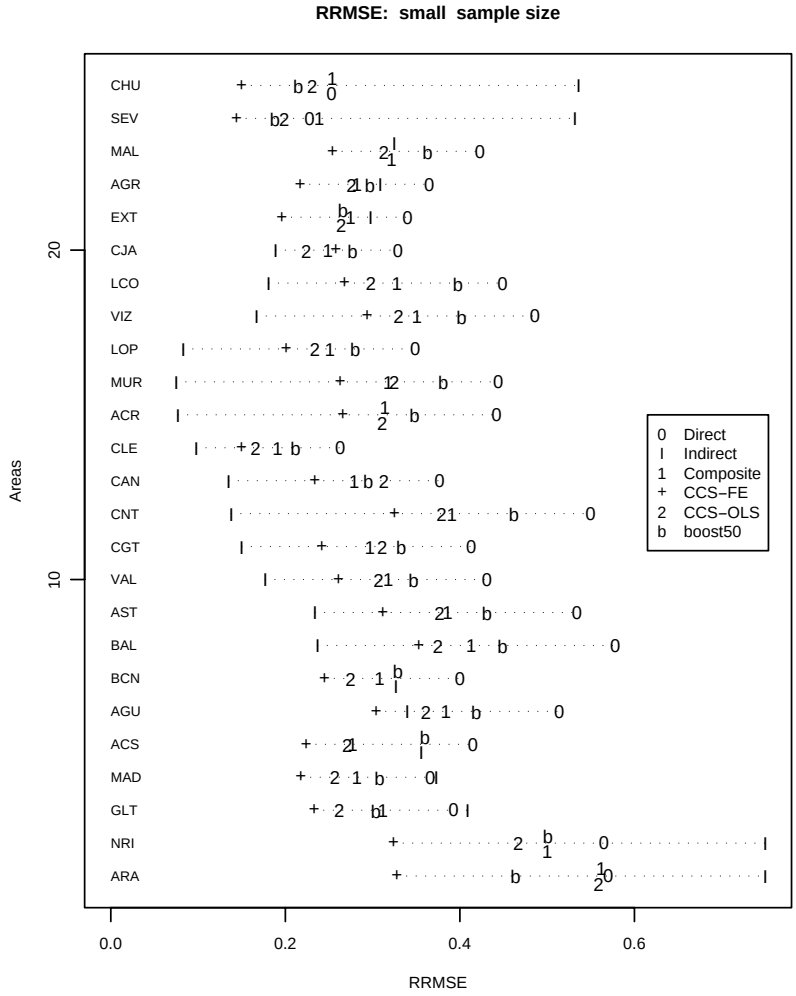

Figure 2: *RRMSEs for the areas and for estimators in the case of small sample (average small area sample size is* 100*). The areas have been ordered in increasing order of magnitude of their rate of unemployment. Values of RRMSE have been truncated at* 0.75*.*

For an area and setting of the simulations (sample size), we define the pattern of RRMSEs by their order for the estimates 0, 1 and 2. For example, for the setting in Figure 1 (large sample size), the pattern for Aragón is 012, which means that the RRMSE for 0 is the smallest and the RRMSE for 2 is the largest of the three; see bottom of the diagram. We say that estimator 2 is the winner for an area if the pattern

of RRMSEs is 201 or 210. In Figure 2, we see that 2 is the winner in 22 areas and 1 is the winner in the remaining three areas. This insight can not be gained from Tables 5-7. With large sample size, estimator 2 wins only 17 areas, so it is still preferable, but less decisively so. We also see that 2 is more efficient than b in 22 areas for small sample, but in only two areas for large sample.

Tables 5-7 and Figures 1 and 2 corroborate the prior expectation that composite estimators outperform direct estimators in almost all settings and for almost all areas, and that the indirect estimator is efficient only in areas with small sample size.

We summarize our findings from the simulations as follows:

1) CCS-OLS (with sample data at one time point) is less efficient than the benchmark estimator CCS-FE.
2) Only for very large samples (1000), CCS-OLS has no gains over the direct estimator. For smaller samples, CCS-OLS is comparable with the benchmark estimator with sample size boosted by up to 50%.
3) A substantial part of the gains attained by the benchmark estimator is attained also by the CCS-OLS estimator.
4) The CCS-OLS estimator is slightly more efficient than the estimators that use information solely from RS.
5) The behaviour of the small area estimators does not change much whether we consider total, male or female unemployment rates.
6) In the context of the estimation of Spanish unemployment rates and for moderate area sample sizes (say, 200 subjects in the area), the simplest CCS-OLS estimator is comparable with an increase of sample size by up to 50%.

As a concluding remark, our results show that regional statistics are not in conflict with the statistics produced at the country-wide level. Rather, a regional survey can be combined with a country survey to improve the precision of estimators for small areas, avoiding the costly solution of increasing the region's subsample size.

6 References

Costa, A, Garcia, M., Lopez, X., and Pardal, M. (2006). Estimació de les taxes de desocupació comarcal a Catalunya. Aplicació d'estimadors de petita àrea amb combinació d'enquestes, Working Document, IDESCAT, Barcelona.

Costa, A., Satorra, A. and Ventura, E. (2003). An empirical evaluation of small area estimators, *SORT (Statistics and Operations Research Transactions)*, 27 (1), 113-135.

Costa, A., Satorra, A. and Ventura, E. (2004). Using composite estimators to improve both domain and total area estimation, *SORT (Statistics and Operations Research Transactions)*, 28 (1), 69-86.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9 (1), 55-93.

Longford, N. T. (2004). Missing data and small-area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society A*, 167, 341-373.

López, R. (2000). Estimaciones para áreas pequeñas. *Estadística Española*, 42, 146, 291-338.

Mancho, J. (2002). Técnicas de estimación en áreas pequeñas. *Cuaderno Técnico del Eustat*.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley Series in Survey Methodology.

StataCorp. (2003) *Stata Statistical Software: Release 8.0*. College Station, TX: Stata Corporation.

Book reviews

MISSING DATA AND SMALL-AREA ESTIMATION

Longford, N.T.

Modern Analytical Equipment for the Survey Statistician.
Springer, 2005
(Statistics for Social Science and Public Policy);
ISBN: 1-852-33760-5

The analysis of large scale surveys usually present the problem of dealing with incomplete data jointly with making inference in a detailed geographical division of the country. Specific statistical methods are necessary to address each one of these topics, which are extensively covered by the present book.

Three differentiated parts divide the eleven chapters of this book. The first and the second parts are focused on the missing data and the small-area estimation problems, respectively. The third part, a single chapter, addresses the problem of model uncertainty presenting a solution inspired on the basis of small-area estimation.

Chapters 1 through 5 constitute the first part of the book devoted to the missing data problem. The book starts with an introductory chapter of survey sampling terminology. The concepts of type of estimator and efficiency are defined in this first chapter. The problem of incompleteness of data is described in Chapter 2. The definitions of complete, incomplete dataset and data analysis are depicted with a detailed example, which leads to introduce the nature of the nonresponse process in surveys and their mechanisms. To deal with those, the single imputation methods, their related models and the EM algorithm are presented in Chapter 3, whereas the method of multiple imputation is detailed in Chapter 4. This chapter makes emphasis on whether to apply multiple or single imputation in an incomplete dataset. Alternative applications of multiple imputation such as measurement error or data editing are posed in the last sections of this chapter. Finally, Chapter 5 presents four case studies of missing data.

Chapters 6 through 10 address the small area estimation problem developing the key idea of similarity among small-areas. Chapter 6 introduces the concept of similarity, discusses the selection procedure between two estimators, suggests composite estimators and finishes with the concept of spatial similarity. Chapter 7 presents models for small area estimation, describes their computational procedures and discusses the

model selection issues. Chapter 8 develops methods that allow to estimate the quantity of interest when auxiliary information allows for greater or equal precision than without it. Chapter 9 highlights on the non-asymptotic nature of the small-area estimators and notes their small-sample variances and precisions. The small-area estimation section of the book finishes in Chapter 10 with four case studies.

The third section is developed in Chapter 11 and argues against the current practice of adopting a model due to failing to find evidence against it. This chapter presents limitations of model selection and introduces the concept of synthetic estimation. In brief, the selection of model and estimators is replaced by linear combinations or predictors based on alternative models. Applications of synthetic estimation are discussed for the analysis of variance and linear regression. This chapter concludes with other applications of synthetic estimation such as meta-analysis.

Every chapter includes suggested readings and exercises, and the author points out that code for the data analyses is available through request. Although computing preferences of the author are S-plus and R, section 5.5 describes other software available for missing data analysis.

To sum up, I think this is an excellent book and it thoroughly covers methods to deal with incomplete data problems and small-area estimation. It is a useful and suitable book for survey statisticians, as well as for researchers and graduate students interested on sampling designs.

Ramon Clèries Soler[1,2]

[1] Servei d'Epidemiologia i Registre del Càncer de l'Institut Català d'Oncologia. Av. Gran Via Km 2,7. HOSPITALET DE LLOBREGAT 08907.

[2] Dept. Economia de l'Empresa. EU Empresarials. Universitat Autònoma de Barcelona Edifici S. C/ dels Emprius, 2. SABADELL 08202.

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 2005 (volume 29)

Annual subscription rates:
– Spain: €22 (VAT included)
– Other countries: €25 (VAT included)

Price for individual issues (current and back issues):
– Spain: €9/issue (VAT included)
– Other countries: €11/issue (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