Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 1994, volum 18, núm. 3
Segona època

Entitats patrocinadores:

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

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 1994, volum 18, núm. 3

Sumari

Articles originals

Estadística General i Investigació Operativa

Estadística Oficial

Secció docent i problemes

EDITORIAL

La part d'Estadística Oficial presenta dos articles complementaris sobre el tema fonamental i complex de les nomenclatures i classificacions en les estadístiques econòmiques.

Aquesta qüestió és bàsica perquè la coordinació i la comparabilitat estadística són essencials en les finalitats següents, que se situen en tres plans diferents: per aconseguir que els valors numèrics procedents dels diversos elements que formen la població d'un treball estadístic mesurin el mateix i que, per tant, siguin homogenis i agregables; per aconseguir que les diverses estadístiques d'un mateix sistema estadístic com el de Catalunya siguin, dintre del possible, comparables, almenys per als conceptes bàsics, els períodes de temps fixats en les distintes àrees geogràfiques i temàtiques i en l'evolució de les sèries cronològiques; finalment, per aconseguir que les estadístiques de diferents sistemes estadístics, com ara el de les Nacions Unides, el comunitari, l'estatal i el de Catalunya siguin comparables, en els aspectes esmentats.

La comparabilitat de les estadístiques en cadascun dels tres plans esmentats és una qualitat que no s'obté espontàniament, sinó que exigeix unes normes rigoroses, molt més minucioses que allò que s'imaginen els ciutadans i, fins i tot, molts investigadors relacionats amb l'estadística. Aquestes normes, per una banda, tenen un aspecte tècnic molt sofisticat i estricte i, per altra banda, requereixen unes disposicions legals que obliguin tots els informants i totes les administracions a complir-les. Així, els aspectes tècnics i els jurídics es relacionen profundament.

La coordinació estadística és una de les principals missions, funcions i preocupacions de tots els instituts oficials d'estadística. En particular, la Llei d'estadística i la Llei del pla estadístic de Catalunya encomanen a l'Institut d'Estadística de Catalunya la coordinació del sistema estadístic de Catalunya i l'homogeneïtzació dels seus resultats.

En aquest marc, l'article "Nomenclatures i classificacions estadístiques", del senyor Josep Maria Martínez López, de l'Institut d'Estadística de Catalunya, ens ofereix una introducció general als problemes i objectius que tenen les classificacions estadístiques i una presentació de les classificacions econòmiques. El treball es concentra en les classificacions de les activitats econòmiques i en les classificacions d'ocupacions. En ambdós temes exposa les diferents classificacions que han estat vigents i les que són vigents, proposades per les Nacions Unides (Oficina Estadística de les Nacions Unides), la Comunitat Europea (Eurostat),

l'Estat espanyol (Instituto Nacional de Estadística) i la Generalitat (Institut d'Estadística de Catalunya). Ambdues presentacions es clouen, respectivament, amb la classificació catalana d'activitats econòmiques (CCAE-93) i la Classificació catalana d'ocupacions (CCO-94).

L'article és interessant per a tots els investigadors i tots els empresaris que empren estadístiques econòmiques en les seves recerques i en les seves decisions, i que han de conèixer bé què hi ha al darrere de les dades numèriques. L'article és necessari per als equips i institucions que elaboren estadístiques econòmiques. Finalment, l'article és bàsic per a aquells que treballen en la construcció de classificacions i la seva implementació informàtica. Així, és interessant notar que per preveure les classes que seran necessàries, es fa servir el "Cluster Analisys".

L'article "La Terminologia estadística i alguns aspectes terminològics de les classificacions estadístiques" de la senyora Griselda Martí Casas, presenta, primer, els problemes generals de la terminologia científica i de la terminologia estadística en català. El treball es concentra en les classificacions d'activitats econòmiques, les classificacions d'ocupacions i les classificacions de béns i serveis. En cadascuna d'elles explica les dificultats que s'han trobat, el mètode que s'ha seguit i el procés de treball que s'ha dut a terme. Hem de destacar que aquestes nomenclatures contenen nombrosíssims noms d'activitats, oficis i productes de tots els rams i de tots els usos.

L'article té un interès específic per a la fixació i la difusió de la terminologia de l'estadística econòmica. També és una manifestació de la contribució de l'Institut d'Estadística de Catalunya a la normalització del català, en aquest cas a través d'unes classificacions que es fan oficials, en ser publicades en els decrets corresponents.

Finalment, l'article té un interès general per a la normalització del català, tant en el camp científic com en el dels oficis i activitats. Esperem que aquest treball també sigui reproduït en àmbits fora del món estadístic.

Eduard Bonet
Barcelona, abril del 1995

En aquest número, la part d'Estadística General i Investigació Operativa, conté sis articles. Els dos primers versen sobre tests d'ajust de dades a un model, el tercer sobre comparació d'estimadors de quocients de mitjanes, i el quart i cinqué són articles que il·lustren aplicacions de l'anàlisi multivariant. El sisé article estudia procediments a fi d'escollir solucions eficients de programació lineal amb objectius múltiples.

L'estat actual d'articles d'aquesta secció és el següent:

- articles sotmesos (gener 1994 — abril 1995): 30
- articles en revisió: 19
- articles acceptats en espera per a publicar: 8
- articles rebutjats (1994): 7

Carles Cuadras
Barcelona, abril del 1995

Estadística General i Investigació Operativa

QÜESTIIÓ, Vol 18, 3 pp. 317-335, 1994

GOODNESS-OF-FIT TEST FOR THE FAMILY OF LOGISTIC DISTRIBUTIONS

N. AGUIRRE* AND M. NIKULIN†

Chi-squared goodness-of-fit test for the family of logistic distributions is proposed. Different methods of estimation of the unknown parameters θ of the family are compared. The problem of homogeneity is considered.

Key words: Logistic distribution; BAN estimator; method of moments; maximum likelihood method; Chernoff-Lehmann theorem; MVUE; homogeneity problem; chi-squared test.

AMS classification: 62E15, 62E20, 62F03, 62G05, 62G10

1. INTRODUCTION

Let $X_1, \ldots, X_n$ be independent identically distributed random variables and suppose that according to the hypothesis H_0

$$(1)\quad \mathbf{P}\{X_i \leq x\} = F(x;\boldsymbol{\theta}), \qquad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^T \in \Theta \subset \mathbb{R}^s, \qquad x \in \mathbb{R}^1,$$

* N. Aguirre. Mathématiques Appliquées, Université Bordeaux 1, France.

† M. Nikulin. Mathématiques Stochastiques, Université Bordeaux 2, V.A. Steklov Mathematical Institute, St.Petersburg, Russia.

-Article rebut el novembre de 1993.

-Acceptat el juliol de 1994.

where Θ is an open set. We devide the real line into k intervals $I_1, \ldots, I_k$:

$$I_1 \cup \cdots \cup I_k = \mathbb{R}^1, \qquad I_i \cap I_j = \emptyset, \quad i \neq j.$$

We shall suppose that

$$p_i(\boldsymbol{\theta}) = \mathbf{P}\{X_1 \in I_i \mid H_0\} > 0, \quad i = 1, \ldots, k. \tag{2}$$

Let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)^T$ be the vector of frequencies arising as a result of grouping the random variables $X_1, \ldots, X_n$ into the classes $I_1, \ldots, I_k$. We denote

$$X_n^2(\boldsymbol{\theta}) = \mathbf{X}_n^T(\boldsymbol{\theta})\mathbf{X}_n(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{(\nu_i - np_i(\boldsymbol{\theta}))^2}{np_i(\boldsymbol{\theta})}, \tag{3}$$

where

$$\mathbf{X}_n = \left(\frac{\nu_1 - np_1(\boldsymbol{\theta})}{\sqrt{np_1(\boldsymbol{\theta})}}, \ldots, \frac{\nu_k - np_k(\boldsymbol{\theta})}{\sqrt{np_k(\boldsymbol{\theta})}} \right)^T. \tag{4}$$

Theorem. (K. PEARSON, 1900)

If $\boldsymbol{\theta}$ is known or given by the hypothesis H_0, then

$$\lim_{n \to \infty} \mathbf{P}\{X_n^2(\boldsymbol{\theta}) \geq x \mid H_0\} = \mathbf{P}\{\chi_{k-1}^2 \geq x\} \tag{5}$$

It is known that if the value of the parameter $\boldsymbol{\theta}$ is unknown and is estimated relative to the observed values of $X_1, \ldots, X_n$, the limiting distribution of the Pearson's statistics $X_n^2(\boldsymbol{\theta}_n^*)$ is given by the asymptotic properties of the estimator $\boldsymbol{\theta}_n^*$ which is substituted into (3) in place of $\boldsymbol{\theta}$.

Here we shall give some results concerning this problem.

2. FISHER'S THEOREM

Following Cramer (1946) we suppose that

1) $p_i(\boldsymbol{\theta}) > c > 0$, $i = 1, \ldots, k$, $(k \geq s + 2)$;

2) $\frac{\partial^2 p_i(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l}$ are continuous functions;

3) the information matrix of Fisher

$$\mathbf{J} = \mathbf{J}(\boldsymbol{\theta}) = \left\| \sum_{l=1}^{k} \frac{1}{p_l(\boldsymbol{\theta})} \frac{\partial p_l(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial p_l(\boldsymbol{\theta})}{\partial \theta_j} \right\|_{s \times s} = \mathbf{B}^T(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta}) \tag{6}$$

exists, and rank$\mathbf{J} = s$, where

$$\mathbf{B}(\boldsymbol{\theta}) = \left\| \frac{1}{\sqrt{p_l(\boldsymbol{\theta})}} \frac{\partial p_l(\boldsymbol{\theta})}{\partial \theta_j} \right\|_{k \times s}. \tag{7}$$

In this case $n\mathbf{J}$ is the information matrix of Fisher of the statistic $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)^T$.

Let $\tilde{\boldsymbol{\theta}}_n$ is the minimum chi-squared estimator for $\boldsymbol{\theta}$,

$$X_n^2(\tilde{\boldsymbol{\theta}}_n) = \min_{\boldsymbol{\theta} \in \Theta} X_n^2(\boldsymbol{\theta}), \tag{8}$$

or an estimator asymptotically equivalent to it. As it was shown by Cramer(1946), a root $\tilde{\boldsymbol{\theta}}_n$ of the system

$$\sum_{i=1}^{k} \frac{\nu_i}{np_i(\boldsymbol{\theta})} \frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_j} = 0, \qquad j = 1, \ldots, s. \tag{9}$$

is such an estimator and, under H_0 as $n \to \infty$, the vector $\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ satisfies the asymptotic relation

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \mathbf{J}^{-1}(\boldsymbol{\theta})\mathbf{B}^T(\boldsymbol{\theta})\mathbf{X}_n(\boldsymbol{\theta}) + o(\mathbf{1}_s), \tag{10}$$

where $o(\mathbf{1}_s)$ is a random vector converging to $\mathbf{0}_s$ in $\mathbf{P}_{\boldsymbol{\theta}}$ - probability, and hence $\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ is asymptotically normally distributed with parameters $\mathbf{0}_s$ and $\mathbf{J}^{-1}(\boldsymbol{\theta})$.

Theorem. (FISHER(1928), CRAMER(1946))

If the regularity conditions of Cramer hold then

$$\lim_{n \to \infty} \mathbf{P}\{X_n^2(\tilde{\boldsymbol{\theta}}_n) \geq x \mid H_0\} = \mathbf{P}\{\chi_{k-s-1}^2 \geq x\}. \tag{11}$$

3. CHERNOFF-LEHMANN'S THEOREM

We suppose that the regularity conditions of Chernoff-Lehmann's (1954) hold:

1) $F(x;\boldsymbol{\theta})$ has probability density $f(x;\boldsymbol{\theta})$, where all

$$\frac{\partial^2 f(x;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_l} \tag{12}$$

are continuous functions on $\mathbf{R}^1 \times \Theta$;

2) the information matrix of Fisher

$$\mathbf{I}(\boldsymbol{\theta}) = \|I_{ij}\|_{s\times s} = \mathbf{E}_{\boldsymbol{\theta}}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{\Lambda}^T(\boldsymbol{\theta}), \tag{13}$$

corresponding to one observation X_1 exists and is positive definite for any $\boldsymbol{\theta} \in \Theta$, where

$$\boldsymbol{\Lambda}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} \ln f(X_1;\boldsymbol{\theta}); \tag{14}$$

($n\mathbf{I}(\boldsymbol{\theta})$ is the amount of information of Fisher about $\boldsymbol{\theta}$ in sample $\mathbb{X} = (X_1,\ldots,X_n)^T$).

3) differentiation with respect to parameters under the integral sign of

$$\int f(x;\boldsymbol{\theta})dx = 1 \tag{15}$$

is permissible, i.e.

$$\frac{\partial}{\partial\theta_i}\int f(x;\boldsymbol{\theta})dx = \int \frac{\partial}{\partial\theta_i} f(x;\boldsymbol{\theta})dx = 0, \qquad i = 1,\ldots,s; \tag{16}$$

4) the matrix $\mathbf{W} = [w_{ij}]$ with elements

$$w_{ij} = \int_{I_j} \frac{\partial}{\partial\theta_i} f(x;\boldsymbol{\theta})dx \tag{17}$$

has rank s;

5) the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ exists,

$$L(\hat{\boldsymbol{\theta}}_n) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}), \tag{18}$$

where

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i;\boldsymbol{\theta}). \tag{19}$$

As it is known (see for example Rao (1965)), $\hat{\boldsymbol{\theta}}_n$ is a solution of the likelihood equation

$$\boldsymbol{\Lambda}(\boldsymbol{\theta}) = \mathbf{0}_s, \tag{20}$$

and, under H_0 as $n \to \infty$, the vector $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ satisfies the asymptotic relation

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\mathbf{I}^{-1}(\boldsymbol{\theta})\boldsymbol{\Lambda}(\boldsymbol{\theta}) + o(\mathbf{1}_s), \tag{21}$$

from where it follows that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ is asymptotically normally distributed with parameters $\mathbf{0}_s$ and $\mathbf{I}^{-1}(\boldsymbol{\theta})$ and hence $\hat{\boldsymbol{\theta}}_n$ is asympotically efficient estimator. We say that $\hat{\boldsymbol{\theta}}_n$ is a BAN estimator.

Theorem. (CHERNOFF-LEHMANN, 1954)

If the regularity conditions 1)-5) hold then

$$\lim_{n\to\infty} \mathbf{P}\{X_n^2(\hat{\boldsymbol{\theta}}_n) \geq x \mid H_0\} = \mathbf{P}\{\chi^2_{k-s-1} + \sum_{i=1}^{s} \lambda_i \xi_i^2 \geq x\}, \tag{22}$$

where $\chi^2_{r-s-1}, \xi_1, \ldots, \xi_s$ are independent, $\xi_i \sim N(0,1)$ and $\lambda_i = \lambda_i(\boldsymbol{\theta})$, $0 < \lambda_i(\boldsymbol{\theta}) < 1, \quad i = 1, 2, \ldots, s$, are the roots of the equation

$$| (1-\lambda)\mathbf{I}(\boldsymbol{\theta}) - \mathbf{J}(\boldsymbol{\theta}) | = 0. \tag{23}$$

Remark 1. We note here that in continuous case $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)^T$ is not sufficient statistic, and hence the matrix $\mathbf{I}(\boldsymbol{\theta}) - \mathbf{J}(\boldsymbol{\theta})$ is positive definite.

Remark 2. Let us consider the density family

$$f(x;\boldsymbol{\theta}) = h(x) exp\{\sum_{m=1}^{s} \theta_m x^m + v(\boldsymbol{\theta})\}, \qquad x \in \mathcal{X} \subseteq \mathbb{R}^1, \tag{24}$$

$\mathcal{X}$ is open in $\mathbf{R}^1$, $\mathcal{X} = \{x : f(x;\boldsymbol{\theta}) > 0\}, \qquad \boldsymbol{\theta} \in \Theta.$

The family (24) is very rich: it contains Poisson, normal distributions etc. It's evident that

$$\mathbf{U}_n = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \ldots, \sum_{i=1}^{n} X_i^s\right)^T \tag{25}$$

is complet minimal sufficient statistics for the family (24).

We suppose that

1) the support $\mathcal{X}$ does not depend on $\boldsymbol{\theta}$;

2) the matrix of Hessen

$$\mathbf{H}_v(\boldsymbol{\theta}) = -\left\| \frac{\partial^2}{\partial\theta_i \partial\theta_j} v(\boldsymbol{\theta}) \right\|_{s\times s} \tag{26}$$

of the function $v(\boldsymbol{\theta})$ is positive definite;

3) the moment $a_s(\boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{\theta}} X_1^s$ exists.

In this case, using the results of Zacks (1971), it is not difficult to show (see, for example, Dzhaparidze and Nikulin (1991)) that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{U}_n)$ and the method of moments estimator $\bar{\boldsymbol{\theta}}_n = \bar{\boldsymbol{\theta}}_n(\mathbf{U}_n)$ of $\boldsymbol{\theta}$ coincide, i.e. $\bar{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n$. Let

$$\mathbf{a}(\boldsymbol{\theta}) = (a_1(\boldsymbol{\theta}), \ldots, a_s(\boldsymbol{\theta}))^T \qquad \text{and} \qquad \mathbf{T}_n = \frac{1}{n}\mathbf{U}_n.$$

One can verify that

$$\mathbf{a}(\boldsymbol{\theta}) = -\frac{\partial}{\partial\boldsymbol{\theta}} v(\boldsymbol{\theta}),$$

and hence the likelihood equation is $\mathbf{T}_n = \mathbf{a}(\boldsymbol{\theta})$, i.e. $\hat{\boldsymbol{\theta}}_n$ is root of this equation. On the other hand we have $\mathbf{E}_{\boldsymbol{\theta}}\mathbf{T}_n \equiv \mathbf{a}(\boldsymbol{\theta})$, and hence from the properties of the statistics $\mathbf{U}_n$ it follows that $\mathbf{T}_n$ is the **MVUE** of $\mathbf{a}(\boldsymbol{\theta})$, and $\bar{\boldsymbol{\theta}}_n$ is the root of the same equation $\mathbf{T}_n = \mathbf{a}(\boldsymbol{\theta})$, which we used to find $\hat{\boldsymbol{\theta}}_n$. Hence $\hat{\boldsymbol{\theta}}_n = \bar{\boldsymbol{\theta}}_n$, i.e. under the conditions 1)-3) the method of moments gives for the family (24) an asymptotically efficient (BAN) estimator. We remark that in general *an estimator based on the method of moments is not asymptotically efficient, and hence does not verify the Chernoff-Lehmann theorem.* In "Handbook of the logistic distribution" in chap. 13, it is reported that this theorem is applied by Massaro and d'Agostino using $\bar{\boldsymbol{\theta}}_n = (\bar{\mathbf{X}}_n, s_n^2)^T$, (the moments method estimator of $\boldsymbol{\theta} = (\mathbf{E}X_1, \mathbf{Var}X_1)^T$) for the family of the logistic distributions. But $\bar{\boldsymbol{\theta}}_n$ is not efficient and ever not asymptotically efficient for the logistic family, and hence is not BAN, since this family does not belong to the exponential family (24) and $(\bar{\mathbf{X}}_n, s_n^2)^T$ is not sufficient statistic in this situation. Hence, the tables of critical points, proposed by Massaro et d'Agostino in section 13.9 are not valid.

4. ROY'S EXTENSION OF THE CHERNOFF-LEHMANN THEOREM

We consider here the result of Dahiya and Gurland (1970,1972), concerning the chi-squared test of Pearson with random intervals, which is an extension of

the unpublished result of Roy (1956) and Chernoff-Lehmann's theorem. It can be found more information about Chernoff-Lehmann's theorem in the paper of LeCam, Mahan and Singh (1983).

Let $\boldsymbol{\theta}_n^*$ is an $\sqrt{n}$ - consistent estimator for $\boldsymbol{\theta}$ such that

$$\sqrt{n}(\boldsymbol{\theta}_n^* - \boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{v}(X_j) + \mathrm{o}(\mathbf{1}_s), \tag{27}$$

where a function $\mathbf{v} = (v_1, \ldots, v_s)^T$ is such that

$$\mathbf{E}_{\boldsymbol{\theta}}\mathbf{v}(X_1) = \mathbf{0}_s \qquad \text{and} \qquad \mathbf{Var}_{\boldsymbol{\theta}}\mathbf{v}(X_1) = \mathbf{V} \qquad \text{is finite.} \tag{28}$$

As it follows from (10) and (21) $\tilde{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_n$ satisfy (27). For each $\boldsymbol{\theta} \in \Theta$ let us define a partition of the real line into k classes-intervals $I_1, I_2, \ldots, I_k$ is defined with boundary points $x_0 = -\infty$, $x_1 = \gamma_1(\boldsymbol{\theta}_n^*), \ldots, x_{k-1} = \gamma_{k-1}(\boldsymbol{\theta}_n^*)$, $x_k = +\infty$ depending on $\boldsymbol{\theta}_n^*$ such that

$$I_i = I_i(\boldsymbol{\theta}_n^*) = \{x : \gamma_{i-1}(\boldsymbol{\theta}_n^*) \le x < \gamma_i(\boldsymbol{\theta}_n^*)\}, \qquad i = 1, \ldots, k, \tag{29}$$

where $\gamma_i(\boldsymbol{\theta})$ are continuous functions having partial derivatives.

Let $\boldsymbol{\nu}^* = (\nu_1^*, \ldots, \nu_k^*)^T$ be a vector of frequencies obtained as a result of grouping the observations $X_1, \ldots, X_n$ by the intervals $I_1(\boldsymbol{\theta}_n^*), I_2(\boldsymbol{\theta}_n^*), \ldots, I_k(\boldsymbol{\theta}_n^*)$ with random boundaries, and let

$$\begin{aligned} p_i(\boldsymbol{\theta}_n^*) &= p_i(\boldsymbol{\theta}_n^*, \boldsymbol{\theta}_n^*) = \mathbf{P}_{\boldsymbol{\theta}_n^*}\{X_1 \in I_i(\boldsymbol{\theta}_n^*) \mid H_0\} = \\ &= = F(\gamma_i(\boldsymbol{\theta}_n^*); \boldsymbol{\theta}_n^*) - F(\gamma_{i-1}(\boldsymbol{\theta}_n^*); \boldsymbol{\theta}_n^*). \end{aligned} \tag{30}$$

To test H_0 Roy (1956) proposed to consider the statistic

$$X_R^2(\boldsymbol{\theta}_n^*) = X_n^2(\boldsymbol{\theta}_n^*, \boldsymbol{\theta}_n^*) = \sum_{i=1}^{k} \frac{(\nu_i^* - np_i(\boldsymbol{\theta}_n^*))^2}{np_i(\boldsymbol{\theta}_n^*)}. \tag{31}$$

Theorem. (Roy(1956), Dahiya & Gurland (1972))

If $\boldsymbol{\theta}_n^*$ satisfies (27) and the Chernoff-Lehmann conditions hold then

$$\lim_{n\to\infty} \mathbf{P}\{X_R^2(\boldsymbol{\theta}_n^*) \ge x \mid H_0\} = \mathbf{P}\{\sum_{i=1}^{k} \lambda_i \xi_i^2 \ge x\}, \tag{32}$$

where $\xi_1, \xi_2, \ldots, \xi_k$ are mutually independent standard normal random variables, $\xi_i \sim N(0,1)$, $\lambda_1 = \lambda_1(\boldsymbol{\theta}), \ldots, \lambda_k = \lambda_k(\boldsymbol{\theta})$ are the characteristic roots of

the matrix $\mathbf{D}^{-1}\boldsymbol{\Sigma}$, where $\mathbf{D} = \mathbf{D}(\boldsymbol{\theta})$ is the diagonal matrix with the elements $p_1(\boldsymbol{\theta}), \ldots, p_k(\boldsymbol{\theta})$ on the main diagonal,

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{D} - \mathbf{p}\mathbf{p}^T - \mathbf{U}^T\mathbf{M} - \mathbf{M}^T\mathbf{U} + \mathbf{U}^T\mathbf{V}\mathbf{U}, \tag{33}$$

$$\mathbf{U}^T = [U_{ij}]_{k\times s}, \qquad U_{ij} = U_{ij}(\boldsymbol{\theta}) = \int_{g_{i-1}(\boldsymbol{\theta})}^{g_i} (\boldsymbol{\theta}) \frac{\partial f(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dx,$$

$$\mathbf{M}^T = [M_{ij}]_{k\times s}, \qquad M_{ij} = M_{ij}(\boldsymbol{\theta}) = \int_{g_{i-1}(\boldsymbol{\theta})}^{g_i} (\boldsymbol{\theta}) v_j(\boldsymbol{\theta}) f(x;\boldsymbol{\theta}) dx.$$

For example, if $\boldsymbol{\theta}_n^* = \hat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimator, which satisfies (21), then, as it was shown by Roy (1956), in (32) $k - s - 1$ of λ_i are equal to 1, one of the λ_i is equal to 0, and the remaining s lie between 0 and 1. It is obvious that this is an extension of the Chernoff-Lehmann theorem to the case of random cell boundaries.

Remark 3. As it was shown by Dahiya and Gurland (1972) if the density function $f(x;\boldsymbol{\theta})$ of X_1 belongs to a location and scale family

$$f(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{\theta_2}} f\left(\frac{x-\theta_1}{\sqrt{\theta_2}}\right), \qquad \boldsymbol{\theta} = (\theta_1, \theta_2)^T, \qquad |\,\theta_1\,| < \infty, \theta_2 > 0, \tag{34}$$

then it is possible to choose the grouping intervals in order to have the asymptotic distribution of X_R^2 independent of $\boldsymbol{\theta}$. For example, let suppose that we test hypothesis H_0 according to which X_i follows the normal distribution $N(\theta_1, \theta_2)$,

$$\mathbf{E}\{X_i \mid H_0\} = \theta_1, \qquad \mathbf{Var}\{X_i \mid H_0\} = \theta_2, \tag{35}$$

and let $\bar{\boldsymbol{\theta}}_n = (\bar{X}_n, s_n^2)^T$, where

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \qquad s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2. \tag{36}$$

$\bar{\boldsymbol{\theta}}_n$ is the method of moments estimator for $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$. Since $f(x;\boldsymbol{\theta})$ belongs to the exponential family (24) of order 2, $s = 2$, the method of moments gives as a result the maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_n = \bar{\boldsymbol{\theta}}_n$.

If in (29) we choose

$$\gamma_i(\boldsymbol{\theta}) = \theta_1 + c_i\sqrt{\theta_2}, \tag{37}$$

and hence $\gamma_i(\hat{\boldsymbol{\theta}}_n) = \bar{X}_n + c_i s_n$, then as it was shown by Gurland and Dahiya (1972) (see also Watson (1957),(1958)), the statistic X_R^2 is distributed, under H_0, in the limit as $n \to \infty$ as

$$\chi_{k-3}^2 + \lambda_1\xi_1^2 + \lambda_2\xi_2^2, \tag{38}$$

where λ_1 and λ_2 do not depend on $\boldsymbol{\theta}$. Using some results of Watson (1957), Dahiya and Gurland tabulated the distribution of Roy's statistic X_R^2 for $k = 3, 4, \ldots, 15$, for the significance level α equal to 0.1,0.05,0.01, in the case when the constants c_i in (37) are chosen so that $p_i(\hat{\boldsymbol{\theta}}_n) = 1/k$.

Remark 4. It is important to remark that when $\boldsymbol{\theta}$ is unknown and we have to estimate it, the limit distribution of the Pearson's statistic $X_n^2(\boldsymbol{\theta}_n^*)$ changes in general in accordance with asymptotical properties of the estimator $\boldsymbol{\theta}_n^*$ we shall use.

For this reason it has looked reasonable to have a statistic which limit distribution is wellknown when we apply the maximum likelihood estimator or anyone BAN estimator. In the papers of Nikulin (1973) (see also, for example, Rao and Robson (1974), Moore and Spruill (1975)), is exposed how to construct a chi-squared test for a continuous distribution (in particular, for the normal distribution and for distributions with shift and scale parameters), based on the statistic $Y_n^2(\boldsymbol{\theta}_n^*)$, by using any BAN estimator $\boldsymbol{\theta}_n^*$ of $\boldsymbol{\theta}$. For example we can take $\boldsymbol{\theta}_n^* = \hat{\boldsymbol{\theta}}_n$, where $\hat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimator. We note that the technique of chi-squared tests for the exponential family of distributions of rank one, s=1, and some applications of MVUE's were exposed by Nikulin and Voinov (1989). Another modification $W_n^2(\boldsymbol{\theta}_n^*)$, which limit distribution is stable with respect to any statistical method of estimation, providing $\sqrt{n}$ - consistent estimator $\boldsymbol{\theta}_n^*$ was proposed by Dzhaparidze and Nikulin (1974). We shall apply the statistic Y_n^2 to test the hypothesis H_0 according to which the distribution of $\mathbf{X}_i$ belongs to the family of the logistic distributions. This topic is studied also by Dudley (1976), Drost(1988).

5. LOGISTIC DISTRIBUTION AND THE CHI-SQUARED GOODNESS-OF-FIT TEST

Let $\mathbb{X} = (X_1, \ldots, X_n)^T$ be a random sample, i.e. $X_1, \ldots, X_n$ are independent identically distributed random variables. In this section we consider the problem of testing the hypothesis H_0 that the distribution function of X_1 belongs to the family of logistic distributions $G\left(\frac{x-\mu}{\sigma}\right)$ depending on the shift parameter μ and the scale parameter σ:

$$(39)\ \mathbf{P}\{X_1 \leq x \mid H_0\} = G\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{1+\exp\{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)\}}, \qquad x \in \mathbb{R}^1,$$

$$\mu = \mathbf{E}\{X_1 \mid H_0\}, \qquad |\mu| < \infty, \qquad \sigma^2 = \mathbf{Var}X_1, \qquad \sigma > 0.$$

Under H_0 the density function of X_i is

$$(40)\quad \frac{1}{\sigma}g\left(\frac{x-\mu}{\sigma}\right) = G'\left(\frac{x-\mu}{\sigma}\right) = \frac{\pi}{\sqrt{3}\sigma}\frac{exp\left(-\frac{\pi}{\sqrt{3}}\frac{x-\mu}{\sigma}\right)}{\left[1+exp\left(-\frac{\pi}{\sqrt{3}}\frac{x-\mu}{\sigma}\right)\right]^2}, \qquad x \in \mathbb{R}^1.$$

We remark that $g(x)$ is symmetric.

We point out that "Handbook of the logistic distribution" edited by Balakrishnan (1992) was published recently about the theory, the methodology and some applications of the family of logistic distributions, see also, Oliver(1964), Pearl and Reed (1920), Reed and Berkson(1929).

We denote $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, and let $\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\sigma}_n^2)^T$ be the maximum of likelihood estimator of $\boldsymbol{\theta}$. Since there is no any other sufficient statistic for $\boldsymbol{\theta}$ than the trivial one $\mathbb{X} = (X_1, \ldots, X_n)^T$, the maximum likelihood equation has no explicit root. Balakrishnan and Cohen (1990) proposed an approximate solution of the maximum likelihood equations based on a "type II censured sample" of Harter and Moore (1967) (see also Grizzle (1961)). They proved that this approximate solution gives an asymptotically efficient estimator, i.e. asymptotically equivalent to $\hat{\boldsymbol{\theta}}_n$.

Let $\hat{\boldsymbol{\theta}}_n$ be such an estimator. As it follows from (21) the limit covariance matrix of the random vector $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ will be $\mathbf{I}^{-1}$, where

$$(41)\quad \mathbf{I} = \frac{1}{\sigma^2}\|I_{ij}\|_{2\times 2} = \frac{1}{9\sigma^2}\begin{Vmatrix} \pi^2 & 0 \\ 0 & \pi^2+3 \end{Vmatrix},$$

$$I_{11} = \int_{-\infty}^{+\infty}\left[\frac{g'(x)}{g(x)}\right]^2 g(x)dx = \frac{\pi^2}{9}, \quad I_{22} = \int_{-\infty}^{+\infty} x^2\left[\frac{g'(x)}{g(x)}\right]^2 g(x)dx - 1 = \frac{\pi^2+3}{9},$$

and since $g(x)$ is symmetric

$$I_{12} = I_{21} = \int_{-\infty}^{+\infty} x\left[\frac{g'(x)}{g(x)}\right]^2 g(x)dx = 0.$$

Let us fix the vector $\mathbf{p} = (p_1, p_2, \ldots, p_k)^T$ of positive probabilities such that

$$(42)\quad p_1 = \ldots = p_k = 1/k,$$

and let

$$(43)\quad y_i = G^{-1}(p_1 + \cdots + p_i) = \frac{\sqrt{3}}{\pi}\ln\left(\frac{i}{k-i}\right), \quad i = 1, \ldots, k-1, \quad y_0 = -\infty, \quad y_k = +\infty.$$

Further, let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)^T$ be the frequency vector arising from grouping $X_1, \ldots, X_n$ over the intervals with random ends

(44) $(-\infty, z_1], (z_1, z_2], \ldots, (z_{k-1}, +\infty)$, where $z_i = z_i(\hat{\boldsymbol{\theta}}_n) = \hat{\mu}_n + \hat{\sigma}_n y_i$,

and let

(45) $$\mathbf{a} = (a_1, \ldots, a_k)^T, \quad \mathbf{b} = (b_1, \ldots, b_k)^T, \quad \mathbf{W}^T = -\frac{1}{\sigma} \| \mathbf{a} \dot{:} \mathbf{b} \|,$$

where for $i = 1, 2, \ldots, k$

$$\begin{aligned} a_i &= g(y_i) - g(y_{i-1}) = \frac{\pi}{k^2\sqrt{3}}(k - 2i + 1), \\ b_i &= y_i g(y_i) - y_{i-1} g(y_{i-1}) = \\ &= \frac{1}{k^2}\left[(i-1)(k-i+1)\ln\frac{k-i+1}{i-1} - i(k-i)\ln\frac{k-i}{i}\right], \end{aligned}$$

(46) $$\alpha(\boldsymbol{\nu}) = k\sum_{i=1}^{k} a_i\nu_i = \frac{\pi}{\sqrt{3}k}\left[(k+1)n - 2\sum_{i=1}^{k} i\nu_i\right],$$

(47) $$\beta(\boldsymbol{\nu}) = k\sum_{i=1}^{k} b_i\nu_i = \frac{1}{k}\sum_{i=1}^{k-1}(\nu_{i+1} - \nu_i)i(k-i)\ln\frac{k-i}{i},$$

(48) $$\lambda_1 = I_{11} - k\sum_{i=1}^{k} a_i^2 = \frac{\pi^2}{9k^2}, \qquad \lambda_2 = I_{22} - k\sum_{i=1}^{k} b_i^2.$$

Since g is symmetric we have $a_1 + a_2 + \cdots + a_k = b_1 + b_2 + \cdots + b_k = 0$. Let

(49) $$\mathbf{B} = \mathbf{D} - \mathbf{p}^T\mathbf{p} - \mathbf{W}^T\mathbf{I}^{-1}\mathbf{W},$$

where $\mathbf{D}$ is the diagonal matrix with the elements $1/k$ on the main diagonal. The matrix $\mathbf{B}$ does not depend on $\boldsymbol{\theta}$, and $\mathrm{rank}\mathbf{B} = k - 1$, i.e. the matrix $\mathbf{B}$ is singular, while the matrix $\tilde{\mathbf{B}}$, obtained as a result of deleting the last row and column in $\mathbf{B}$, has an inverse

(50) $$\tilde{\mathbf{B}}^{-1} = \mathbf{A} + \mathbf{A}\tilde{\mathbf{W}}^T(\mathbf{I} - \tilde{\mathbf{W}}\mathbf{A}\tilde{\mathbf{W}}^T)^{-1}\tilde{\mathbf{W}}\mathbf{A},$$

where $\mathbf{A} = \tilde{\mathbf{D}}^{-1} + \mathbf{1}\mathbf{1}^T/p_k$, $\tilde{\mathbf{D}}^{-1}$ is a diagonal matrix with elements $\frac{1}{p_1}, \ldots, \frac{1}{p_{k-1}}$ on the main diagonal, $\mathbf{1} = \mathbf{1}_{k-1}$ is the vector of dimension $(k-1)$, all elements of which are equal to 1, $\tilde{\mathbf{W}}$ is a matrix obtained from $\mathbf{W}$ by deleting the last column. Since the vector $\tilde{\boldsymbol{\nu}} = (\nu_1, \ldots, \nu_{k-1})^T$ is asymptotically normally distributed with parameters

(51) $\mathbf{E}\tilde{\boldsymbol{\nu}} = n\tilde{\mathbf{p}} + O(\mathbf{1}_s)$ and $\mathbf{E}(\tilde{\boldsymbol{\nu}} - n\tilde{\mathbf{p}})^T(\tilde{\boldsymbol{\nu}} - n\tilde{\mathbf{p}}) = n\tilde{\mathbf{B}} + O(\mathbf{1}_{s\times s})$,

where $\tilde{\mathbf{p}} = (p_1, \ldots, p_{k-1})^T$, we obtain the next result

Theorem 1.

The statistic

(52) $$Y_n^2 = \frac{1}{n}(\tilde{\boldsymbol{\nu}} - n\tilde{\mathbf{p}})^T\tilde{\mathbf{B}}^{-1}(\tilde{\boldsymbol{\nu}} - n\tilde{\mathbf{p}}) = X_n^2 + \frac{\lambda_1\beta^2(\boldsymbol{\nu}) + \lambda_2\alpha^2(\boldsymbol{\nu})}{n\lambda_1\lambda_2}$$

has, as $n \to \infty$, chi-squared limit distribution with $(k-1)$ degrees of freedom, where

(53) $$X_n^2 = \sum_{i=1}^{k}\frac{(\nu_i - np_i)^2}{np_i} = \frac{k}{n}\sum_{i=1}^{k}\nu_i^2 - n.$$

Remark 5. We consider the hypothesis H_η according to which X_i follows $G(\frac{x-\mu}{\sigma}, \eta)$, where $G(x,\eta)$ is continuous, $\mid x \mid < \infty$, $\eta \in \mathbf{H} \subset \mathbb{R}^1$, $G(x,0) = G(x)$, and $\eta = 0$ is a limit point of $\mathbf{H}$. Let us assume also, that

(54) $$\frac{\partial}{\partial x}G(x,\eta) = g(x,\eta) \qquad \text{and} \qquad \frac{\partial}{\partial\eta}g(x,\eta)\mid_{\eta=0} = \Psi(x),$$

exist, where $g(x,0) = g(x) = G'(x)$. In this case if $\frac{\partial^2 g(x,\eta)}{\partial\eta^2}$ exists and is continuous for all x in the neighbourhood of the $\eta = 0$, then

(55) $$\mathbf{P}\{z_{i-1} < X_i \le z_i \mid H_\eta\} = p_i + \eta c_i + \mathrm{o}(\eta),$$

where

(56) $$c_i = \int_{z_{i-1}}^{z_i}\Psi(x)dx, \qquad i = 1, \ldots, k,$$

and finally, in the limit as $n \to \infty$ the statistic Y_n^2 has noncentral chi-squared distribution with $(k-1)$ degrees of freedom and with non-centrality parameter λ:

(57) $$\lim_{n\to\infty}\mathbf{P}\{Y_n^2 \ge x \mid H_\eta\} = \mathbf{P}\{\chi_{k-1}^2(\lambda) \ge x\},$$

where

$$\lambda = \sum_{i=1}^{k}\frac{c_i^2}{p_i} + \frac{\lambda_2\alpha^2(\mathbf{c}) + \lambda_1\beta^2(\mathbf{c})}{\lambda_1\lambda_2}, \mathbf{c} = (c_1, c_2, \ldots, c_k)^T,$$

$\mathbf{p}$, $\alpha(\mathbf{c})$, $\beta(\mathbf{c})$, λ_1, λ_2 are given by (42),(46),(47),(48) respectively.

6. EXAMPLE (ABOUT A CHOICE OF INTERVALS)

1. Simple hypotheses. Suppose that we want to test the simple hypothesis H_0 acording to which

(1) $$\mathbf{P}\{X_1 \leq x \mid H_0\} = G(x)$$

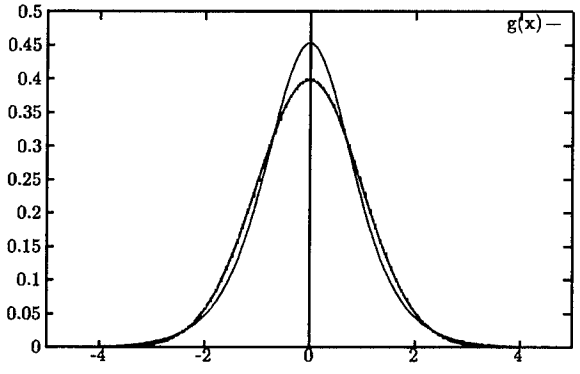
against the simple hypothesis H_1:

(2) $$\mathbf{P}\{X_1 \leq x \mid H_1\} = \Phi(x)$$

($\Phi(x)$ being the standard normal distribution function).

It would be possible to use the Neymann and Pearson test, yet it would entrail large calculations. We shall then try to adapt a chi-squared test.

Let $(X_1, ..., X_n)^T$ be a sample of mutually independent identically distributed random variables with $\mathbf{E}(X_1) = 0$, $\mathbf{Var}(X_1) = 1$. Before to construct a chi-square test for testing H_0 against H_1 we shall do one remark on the cells' choice. As the test only compares the respective frequencies on the cells, it is worthwhile to choose those when both density curves are the most distant, that is to say those got by their junctions.



The representative curves of the density function $g(x)$ of $L(0,1)$ distribution and the density function $\varphi(x)$ of the standard normal $N(0,1)$ distribution have four symmetric points of intersection:

$$x_1 = -x_6 = -\infty, \qquad x_2 = -x_5, \qquad x_3 = -x_4,$$

where $\varphi(x_i) = g(x_i)$. Let

$$I_i = \{x_{i-1} < x \leq x_i\}, \qquad i = 1, 2, 3, 4, 5,$$

be the such definite intervals of grouping the data. We can even improve the power of the test by considering the cells only differentiated by the relative positions of the two curves:

$$\begin{aligned} J_1 &= I_1 \cup I_3 \cup I_5 = \{x : g(x) > \varphi(x)\}, \quad g(x) \text{ higher that } \varphi(x), \\ J_2 &= I_2 \cup I_4 = \{x : \varphi(x) > g(x)\}, \qquad \varphi(x) \text{ higher that } g(x). \end{aligned}$$

Let us remark that if we consider, for example, an interval $-2 < x \leq 0$ for grouping the data with one point of intersection x_3 on the inside, as shown on the schema, it is clear that in this case the two probabilities

$$\mathbf{P}\{-2 < X_i \leq 0 \mid H_0\} \text{ and } \mathbf{P}\{-2 < X_i \leq 0 \mid H_1\}$$

will be approximately equal, and it will be difficult to decide which of both hypotheses is true and hence the power of the test using such interval will be low. A test using the intervals $\{-2 < x \leq x_3\}$ and $\{x_3 < x \leq 0\}$ (or $x_3 < x \leq x_4\}$) will obviously be more powerfull. From this remark we obtain the next

Proposition

To test H_0 against H_1, the chi-squared test using I_1, I_2, I_3, I_4, I_5 is less powerful as the one using two cells:

$$J_1 \text{ and } J_2.$$

a) Let $(\nu_1, \nu_2, \ldots, \nu_5)^T$ be the observed frequencies of the sample in the intervals $I_1, \ldots, I_5$ respectively. Then

$$\mathbf{P}\{X_1 \in I_i \mid H_0\} = p_i^{(0)}, \quad i = 1, 2, \ldots, 5 \tag{60}$$

(with $p_1^{(0)} = p_5^{(0)} = 0.0155$; $p_2^{(0)} = p_4^{(0)} = 0.207$; $p_3^{(0)} = 0.555$). In our case we shall use the standard statistic of Pearson:

$$X_n^2 = \sum_{i=1}^{5} \frac{(\nu_i - np_i^{(0)})^2}{np_i^{(0)}}. \tag{61}$$

Under H_0, X_n^2 is asymptotically χ_4^2 and we shall reject H_0 if $X_n^2 \geq c_{4,\alpha}$, since

$$\mathbf{P}\{X_n^2 \geq c_{4,\alpha} \mid H_0\} \approx \mathbf{P}\{\chi_4^2 \geq c_{4,\alpha}\} = \alpha.$$

The power of this test is $\mathcal{P}_5 = \mathbf{P}\{X_n^2 \geq c_{4,\alpha} \mid H_1\}$. Or,

$$\mathbf{P}\{X_1 \in I_i \mid H_1\} = p_i^{(1)}, \quad i = 1, 2, \ldots, 5, \tag{62}$$

$(p_1^{(1)} = p_5^{(1)} = 0.011; p_2^{(1)} = p_4^{(1)} = 0.234; p_3^{(1)} = 0.51)$.

Under H_1, X_n^2 has asymptotically a non-central chi-square distribution with 4 degrees of freedom and the parameter of non centrality λ:

$$\lambda = n \sum_{i=1}^{5} \frac{(p_i^{(0)} - p_i^{(1)})^2}{p_i^{(0)}}, \qquad (\lambda \sim 0.0133n).$$

Using the approximation of Patnaik we can compute

$$\mathbf{P}\{\chi_{k-1}^2(\lambda) \geq c_\alpha\} \sim \mathbf{P}\{\chi_{k-1}^2 \geq c_\alpha \left(1 - \frac{\lambda}{k-1}\right)\}. \tag{63}$$

For example, for $\alpha = 0.05$ we have $c_{4,\alpha} = 9.49$, from which it follows that

in order to have $\mathcal{P}_5$	$>$	0.5	0.74	0.9	0.99,
it is necessary to take n	$>$	195	240	269	292

respectively.

b) Let ν be the observed frequency on J_1. We have

$$\mathbf{P}\{X_1 \in J_1 \mid H_0\} = \omega^{(0)} \text{ with } \omega^{(0)} = 0.586$$

and

$$\mathbf{P}\{X_1 \in J_1 \mid H_1\} = \omega^{(1)} \text{ with } \omega^{(1)} = 0.532. \tag{64}$$

To test H_0 against H_1 we shall use again the standard statistic of Pearson

$$X_n^2 = \frac{(\nu - n\omega^{(0)})^2}{n\omega^{(0)}(1 - \omega^{(0)})}.$$

Under H_0 the statistic X_n^2 is distributed $(n \to \infty)$ asymptotically as χ_1^2. We shall reject H_0 if $X_n^2 \geq c_{1,\alpha}$, since $\mathbf{P}\{\chi_1^2 \geq c_{1,\alpha}\} = \alpha$ and the power of this test is

$$\mathcal{P}_2 = \mathbf{P}\{X_n^2 \geq c_{1,\alpha} \mid H_1\}.$$

Under H_1, X_n^2 has asymptotically $(n \to \infty)$ a non central chi-square distribution with one degree of freedom and the parameter of noncentrality

$$\lambda' = \frac{[\omega^{(0)} - \omega^{(1)}]^2}{\omega^{(0)}(1 - \omega^{(0)})}, \qquad (\lambda' \sim 0.012n).$$

Using the same approximation (63):

$$\mathbf{P}\{\chi_1^2(\lambda') \geq c_\alpha\} \sim \mathbf{P}\{\chi_1^2 \geq c_{1,\alpha}(1 - \lambda')\},$$

we have $c_{1,\alpha} = 3.84$ for $\alpha = 0.05$, from which it follows that

$$\begin{array}{lrrrrr} \text{in order to have } \mathcal{P}_2 & > & 0.5 & 0.75 & 0.9 & 0.99, \\ \text{it is necessary to take } n & > & 74 & 82 & 83 & 84 \end{array}$$

respectively. One can see that in the case b) the chi-square test based on J_1 and J_2 is more powerfull, then in the case a). The same approach may be used for composite hypotheses (see, for example, Nikulin & Voinov, 1989).

2. Composite hypotheses. Let now $\mathbb{X} = (X_1, \ldots, X_n)^T$ be a sample, $\mathbf{E}(X_1) = \mu$ and $\mathbf{Var}(X_1) = \sigma^2$, $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, $\boldsymbol{\theta}$ is unknown, and let us test H_0 according to which X_1 follows a logistic distribution (39):

$$\mathbf{P}\{X_1 \leq x \mid H_0\} = G(x, \boldsymbol{\theta}) = G(\frac{x-\mu}{\sigma}) \tag{65}$$

against the hypothesis of normality H_1 according to which

$$\mathbf{P}\{X_1 \leq x \mid H_1\} = \Phi(\frac{x-\mu}{\sigma}). \tag{66}$$

Let $\hat{\boldsymbol{\theta}}_n$ be an estimator which satisfies to (21). According to the precedent study and to §5, we shall take the two cells with random boundaries (41):

$$\begin{aligned} J_1(\hat{\boldsymbol{\theta}}_n) &=]-\infty, -\hat{\sigma}_n x_5 + \hat{\mu}_n] \cup] -\hat{\sigma}_n x_3 + \hat{\mu}_n, \hat{\sigma}_n x_3 + \hat{\mu}_n] \cup]\hat{\sigma}_n x_5 + \hat{\mu}_n, +\infty[, \\ J_2(\hat{\boldsymbol{\theta}}_n) &= \mathbb{R}^1 \setminus J_1(\hat{\boldsymbol{\theta}}_n). \end{aligned}$$

Let $\boldsymbol{\nu} = (\nu, n-\nu)^T$ be the vector of frequencies obtained in the result of the groupement of the sample $\mathbb{X} = (X_1, \ldots, X_n)^T$ into the intervals $J_1(\hat{\boldsymbol{\theta}}_n)$ and $J_2(\hat{\boldsymbol{\theta}}_n)$. According to definitions (43) to (47) we calculate

$$a'_1 = -\sigma \frac{\partial}{\partial \mu} \mathbf{P}\{X_i \in J_1(\boldsymbol{\theta}) \mid H_0\} = a_1 + a_3 + a_5 = 0 = a'_2,$$

$$\begin{aligned} b'_1 &= -\sigma \frac{\partial}{\partial \sigma} \mathbf{P}\{X_i \in J_1(\boldsymbol{\theta}) \mid H_0\} = \\ &= b_1 + b_3 + b_5 = 2[x_3 g(x_3) - x_5 g(x_5)] = -b'_2 \approx 0.296, \end{aligned}$$

$$\lambda_1 = I_{11} - \frac{a'^2_1}{\omega^{(0)}} - \frac{a'^2_2}{1-\omega^{(0)}} = I_{11} = \frac{\pi^2}{9},$$

$$\lambda_2 = I_{22} - \frac{b_1'^2}{\omega^{(0)}} - \frac{b_2'^2}{1-\omega^{(0)}} = \frac{\pi^2+3}{9} - \frac{b_1'^2}{\omega^{(0)}(1-\omega^{(0)})},$$

$$\lambda_3 = I_{12} - 0 = 0,$$

$$\alpha(\boldsymbol{\nu}) = \frac{a_1'\nu}{\omega^{(0)}} + \frac{a_2'(n-\nu)}{1-\omega^{(0)}} = 0,$$

$$\beta(\boldsymbol{\nu}) = \frac{b_1'\nu}{\omega^{(0)}} + \frac{b_2'(n-\nu)}{1-\omega^{(0)}} = b_1'\frac{(\nu - n\omega^{(0)})}{\omega^{(0)}(1-\omega^{(0)})},$$

$$X_n^2 = \frac{(\nu - n\omega^{(0)})^2}{n\omega^{(0)}(1-\omega^{(0)})},$$

and finally we obtain the statistic

$$(67) \qquad Y_n^2 = X_n^2 + \frac{\beta(\boldsymbol{\nu})}{n\lambda_2} = X_n^2 + \frac{\frac{b_1'^2}{\omega^{(0)}(1-\omega^{(0)})}}{\left[\frac{\pi^2+3}{9} - \frac{b_1'^2}{\omega^{(0)}(1-\omega^{(0)})}\right]} X_n^2 \sim 1.34 X_n^2,$$

for testing the composite hypothesis H_0, given by (39), against the hypothesis of normality H_1, and

$$\mathbf{P}\{Y_n^2 \geq x \mid H_0\} \to \mathbf{P}\{\chi_1^2 \geq x\}, \qquad (n \to \infty).$$

AKNOWLEDGEMENT

We would like to thank Professors S. Kotz, A. Rukhin, C. Huber, N. Balakrishnan and J. Antoch and referees for helpful discussion, advice and encouragement during the writing of the paper.

REFERENCES

[1] **Balakrishnan, N.** (1992). *Handbook of the logistic distribution.* Marcel Dekker, New York.

[2] **Balakrishnan, N.** and **Cohen, A.C.** (1990). *Order Statistics and Inference: Estimation Methods.* Academic Press, Boston.

[3] **Bolshev, L.N.** and **Nikulin, M.S.** (1975). "One solution of the problem of homogeneity". Serdika. *Bulgarsko Mathematichesko Spicanie,* **1**, 104-109.

[4] **Chernoff, H.** and **Lehmann, E.L.** (1954). "The use of maximum likelihood estimates in χ^2 tests for goodness of fit". *Ann. Math. Stat.,* **25**, 579-586.

[5] **Cramer, H.** (1946). *Mathematical Methods of Statistics.* Princeton University Press.

[6] **Dahiya, R.C.** and **Gurland, J.** (1970a). "Pearson chi-square test of fit with random intervals. I. null case". *MRC Technical Summary Report 1046.* University of Wisconsin.

[7] **Dahiya, R.C.** and **Gurland, J.** (1970b). "Pearson chi-square test of fit with random intervals. II. Non-null case". *MRC Technical Summary Report 1051.* University of Wisconsin.

[8] **Dahiya, R.C.** and **Gurland, J.** (1972). "Pearson chi-square test of fit with random intervals". *Biometrika,* **59**, 1, 147-153.

[9] **Drost, F.C.** (1988). "Asymptotics for generalised chi-square goodness-of-fit tests". *Amsterdam: Centre for Mathematics and Computer Sciences,* CWI tracts, **48**.

[10] **Dudley, R.M.** (1976). "Probabilities and metrics-convergence of laws on metric spaces with a view to statistical testing". *Lecture Notes, Aarhus Universitet,* **45**.

[11] **Dzhaparidze, K.O.** and **Nikulin, M.S.** (1974). "On a modification of the standard statistic of Pearson". *Theory of Probability and its Applications,* **19**, 4, 851-852.

[12] **Fisher, R.A.** (1928). "On a property connecting the χ^2 measure of discrepancy with the method of maximum likelihood". *Atti de Congresso Internazionale di Mathematici,* Bologna, **6**, 94-100.

[13] **Grizzle, J.E.** (1961). "A new method of testing hypotheses and estimating parameters for the logistic model". *Biometrics,* **17**, 372-385.

[14] **Harter, H.L.** and **Moore, A.H.** (1967). "Maximum-likelihood estimation, from censored samples, of the parameters of a logistic distribution". *J. Amer. Statist. Assoc.,* **62**, 675-684.

[15] **Le Cam, L., Mahan, C.** and **Singh, A.** (1983). "An extension of a theorem of H. Chernoff and E.L. Lehmann", in *Recent advances in statistics.* Academic Press, 303-332.

[16] **Moore, D.S.** and **Spruill, M.C.** (1975). "Unified large-sample theory of general chi-squared statistics for tests of fit". *Ann. of Statist.*, **3**, 599-616.

[17] **Nikulin M.S.** (1973). "Chi-square test for normality". *International Vilnius Conference on Probability Theory and Mathematical Statistics,* **2**, 119-122.

[18] **Nikulin M.S.** (1973). "Chi-square test for continuous distributions with shift and scale parameters". *Theory of Probability and its Applications,* **18**, #3, 559-568.

[19] **Nikulin M.S.** (1973). "Chi-square test for continuous distributions". *Theory of Probability and its Applications,* **18**, #3, 638-639.

[20] **Nikulin M.S.** and **Voinov V.G** (1989). "A chi-square goodness-of-fit test for exponential distribution of the first order". *Springer-Verlag Lecture Notes in Mathematics*, **1312**, 239-258.

[21] **Nikulin M.S.** and **Greenwood, P.E.** (1990). "A Guide to Chi-squared Testing". *Technical Report 94.* The Departement of Statistics, The University of British Columbia, Vancouver, Canada, 199p.

[22] **Oliver, F.R.** (1964). "Methods of estimating the logistic growth function". *Appl.Statist.*, **13**, 57-66.

[23] **Pearl, R.** and **Reed, L.J.** (1920). "On the rate of growth of the population of the United States since 1790 and its mathematical representation". *Proc. of National Acad. Sci.*, **6**, 275-288.

Rao, C.R. (1965). Linear Statistical Inference and its Applications, John Wiley & Sons.

[24] **Rao, K.C.** and **Robson, D.S.** (1974). "A chi-squared statistic for goodness-of-fit tests within the exponential family". *Commun. in Statistics,* **3**, 1139-1153.

[25] **Reed, L.J.** and **Berkson, J.** (1929). "The application of the logistic function to experimental data". *J.Physical Chemistry*, **33**, 760-779.

[26] **Roy, A.R.** (1956). "On χ^2-statistics with variable intervals". *Technical Report N1*, Stanford University, Statistics Departement.

[27] **Watson, G.S.** (1958). "On chi-square goodness-of-fit tests for continuous distributions". *J. Roy. Statist. Soc.*, **B20**, 1, 44-61.

[28] **Watson, G.S.** (1957). "The χ^2 goodness-of-fit test for normal distributions". *Biometrika,* **44**, 336-348.

[29] **Zacks, S.** (1979). *The Theory of Statistical Inference.* Wiley, New York.

QÜESTIIÓ, Vol 18, 3 pp. 337-368, 1994

TEST DE BONDAD DE AJUSTE DEL MODELO LINEAL GENERAL BAJO CORRELACIÓN SERIAL DE LOS ERRORES*

J.M. VILAR FERNÁNDEZ* y W. GONZÁLEZ MANTEIGA†

Dado el siguiente modelo de regresión de diseño fijo, con correlación serial en los errores: $Y_i = m(x_i) + \epsilon_i$, *donde* $x_i \in \mathcal{C}$, $i = 1, \ldots, n$, *siendo* $\mathcal{C}$ *un conjunto compacto de* $\mathbb{R}$, *con error aleatorio* ϵ_i *siguiendo una estructura lineal de tipo* $MA(\infty)$, *se propone un nuevo método para contrastar la hipótesis de que la función de regresión siga un modelo lineal, de la forma:* $m_\theta(-) = \mathcal{A}^t(-)\theta$, *con* $\theta \in \Theta \subset \mathbb{R}^q$, *y* $\mathcal{A}$ *es un funcional de* $\mathbb{R}$ *en* $\mathbb{R}^q$.

El estadístico propuesto para contrastar la hipótesis de linealidad, que denominamos d^2, *se obtiene como la distancia del tipo Cramer-von Mises entre el estimador no paramétrico de Gasser-Müller,* $\hat{m}$ *de la función de regresión y el estimador paramétrico de mínima distancia bajo la hipótesis de linealidad:* $m_{\hat{\theta}}$.

Los resultados presentados de normalidad asintótica para ambos estimadores: $\hat{\theta}_n$ *y* d^2, *y los estudios de simulación llevados a cabo sirven para ilustrar el efecto de la dependencia e indicar algunas formas de elegir el parámetro de suavización. Finalmente se incluyen ejemplos con datos reales.*

Testing the hypothesis of a General Linear Model when errors are non-independent.

Key words: Test de hipótesis paramétrica, modelos de regresión, series de tiempo, estimadores no paramétricos.

Clasificación A.M.S.: 62G07, 62G10, 62J05, 62M10.

* J.M. Vilar Fernández. Departamento de Matemáticas. Universidad de La Coruña. España.
† W. González Manteiga. Departamento de Estadística e Investigación Operativa. Universidad de Santiago de Compostela. España.
✠ Este trabajo ha sido parcialmente subvencionado por la DGYCYT (PB91-0794) y Xunta de Galicia (X4GA20701B92).
-Article rebut el setembre de 1993.
-Acceptat el febrer de 1994.

1. INTRODUCCIÓN

Tradicionalmente, a lo largo de este siglo, ha sido de interés en el campo de la Estadística, el estudio de la relación entre las variables observadas en los diversos problemas de la vida real. Siendo el objetivo principal la creación de modelos de tipo estocástico, que sirvan para interpretar dicha relación.

De los modelos diseñados para dicho análisis, destacan por su importancia, los llamados modelos de regresión. Así, si Y es la variable de interés (variable respuesta, variable predecible, etc.) y $x^t = (x_1, \ldots, x_p)$ es el conjunto de variables regresoras (variables predictoras, ..., etc.), de las que se desea estudiar su influencia sobre la variable anterior, se define un modelo de regresión, como el dado por:

$$(1.1) \qquad Y_i = m_\theta(x_{1i}, \ldots, x_{pi}) + \epsilon_i, \quad i = 1, \ldots, n; \quad \theta \in \Theta \subset \mathbb{R}^q$$

en el que $\{\epsilon_i\}_{i=1}^n$ es una sucesión de variables aleatorias de media cero y $\{((x_{1i}, \ldots, x_{pi}), Y_i)\}_{i=1}^n$ representa la muestra inicial observada.

Uno de los tópicos más estudiados en la teoría de los modelos de regresión, como el (1.1) antes descrito, es el relativo a los distintos contrastes de hipótesis acerca de la forma m_θ que mejor explica la dependencia entre las variables predictoras y la variable respuesta. Existe una extensa literatura al respecto, de la que destacamos: Seber (1977), para modelos de regresión lineal, esto es, $m_\theta(x) = \mathcal{A}^t(x)\theta$, con $\mathcal{A}(x) \in \mathbb{R}^q$ y $x \in \mathbb{R}^p$; y, Seber y Wild (1989), para situaciones en las que, en general, m_θ es una función no lineal de θ. En cualquiera de estos casos, el planteamiento general del contraste de hipótesis viene dado por la hipótesis nula:

$$\mathrm{H_o}: Y_i = m_\theta(x_{1i}, \ldots, x_{pi}) + \epsilon_i,\ i = 1, \ldots, n;\ \theta \in \Theta_1 \subset \Theta, \text{con } \dim(\Theta_1) = q_1 < q$$

frente a la alternativa general $\mathrm{H_1}$, que representa la validez del modelo (1.1).

Una perspectiva más general que la descrita en los párrafos anteriores consistiría en suponer que el espacio asociado a $\mathrm{H_1}$ no estuviese restringido a una familia paramétrica de funciones, tratándose de contrastar:

$$(1.2) \qquad \mathrm{H_o}: m \in \{m_\theta(-): \theta \in \Theta \subset \mathbb{R}^q\}$$

frente a la alternativa:

$$\mathrm{H_1} = \text{“}m \text{ es una función con cierto grado de suavidad.”}$$

De este modo estaríamos realmente contrastando la bondad de ajuste de una familia paramétrica de funciones de regresión, haciéndose, por tanto, necesaria

la estimación no paramétrica de la función de regresión m en la hipótesis general H_1. El estudio de estos contrastes es bastante reciente, apareciendo los primeros desarrollos formales fundamentalmente en los tres últimos años. Raz (1990), Firth-Glosup-Hinkley (1991), Staniswalis-Severini (1991), Eubank-Spiegelman (1990), Kozek (1991), Hart-Wehrly (1992) y Eubank-Hart (1992 y 1993) entre otros, son algunos ejemplos del trabajo desarrollado en esa línea cuando se toman, principalmente, como estimadores no paramétricos los de tipo núcleo o spline.

Una extensión natural del modelo (1.1) viene dada por la suposición de que los errores $\{\epsilon_i\}_{i=1}^n$ siguen una estructura de dependencia. Lo que sucede frecuentemente, en el estudio de muestras de datos económicos, ya que en muchos casos estos son recogidos secuencialmente a lo largo del tiempo (índices de desempleo, precios de ciertos productos, ..., etc.) mostrando una clara correlación serial.

Como consecuencia de la presencia de correlación entre los errores en el modelo, los clásicos estimadores $\hat{\theta}$ de máxima verosimilitud, cuando no se considera la dependencia de las observaciones, o los de mínimos cuadrados, pueden llegar a ser muy ineficientes. (Ver el capítulo 6 de Seber y Wild (1989) para un análisis detallado del efecto de la correlación en la estimación de los parámetros). Además la dependencia en los errores del modelo da lugar a estimaciones infra o sobresuavizadas de m, en la hipótesis general H_1, cuando la ventana (o parámetro de suavización) es elegida adecuadamente para situaciones de errores incorrelados (ver Altman (1990) o Chu-Marron (1991) para más detalles).

En lo que sigue se tratará el contraste de la hipótesis general (1.2) bajo la hipótesis de que ϵ_i, el error aleatorio en el modelo (1.1), sigue una estructura de tipo lineal: "medias móviles de orden infinito". Es decir, $\epsilon_i = \sum_{j=0}^{\infty} b_j e_{i-j}$, con $\{e_i\}$ una sucesión de ruido blanco, o sea, variables aleatorias independientes de media cero y varianza σ_e^2.

El contraste de este tipo de hipótesis con alternativas generales no paramétricas y con suposición de dependencia en los errores es, desde nuestra perspectiva, un planteamiento más general que no había sido abordado hasta la actualidad. La suposición anterior, del tipo de medias móviles, incluye, entre otros, a los modelos denominados ARMA.

En este trabajo se estudia un contraste de la hipótesis nula, cuando ésta es de tipo lineal, esto es, $H_o: m \in \{m_\theta(-) = \mathcal{A}^t(-)\theta : \theta \in \Theta \subset \mathbb{R}^q\}$, siendo $\mathcal{A}$ un funcional de $\mathbb{R}$ en $\mathbb{R}^q$. Para ello se mide la distancia entre estimador paramétrico bajo H_o y un estimador no paramétrico, tipo núcleo.

En la sección 2 se define el estimador de mínima distancia, $\hat{\theta}_n$, del parámetro del modelo, y el estadístico del contraste de bondad de ajuste, d^2, obteniéndose la distribución asintótica de ambos estimadores bajo la hipótesis nula H_o. En el apartado 3 se analizan estudios de simulación que ilustran el efecto de la dependencia e indican una forma razonable y eficiente de elegir el parámetro de suavización. En el apartado 4, se estudian algunos ejemplos con datos reales. Finalmente, en el Apéndice se presenta la demostración de los resultados obtenidos.

2. DEFINICIONES Y RESULTADOS ASINTÓTICOS

2.1. Definición de los estimadores

En lo que sigue consideraremos el siguiente modelo de diseño fijo:

$$Y_i = m(x_i) + \epsilon_i \tag{2.1}$$

donde m es una función suave, que sin pérdida de generalidad, suponemos definida en [0,1], los $\{x_i\}_{i=1}^n$ están equiespaciados, es decir, $x_i = i/n$, los errores $\{\epsilon_i\}_{i=1}^n$ constituyen un proceso causal como se comentó en el apartado anterior: $\epsilon_i = \sum_{j=0}^{\infty} b_j e_{i-j}$, con $\{e_t\}$ ruido blanco de varianza σ_e^2. (Ver Brockwell y Davis (1991), Def. 3.1.3 para más detalles) y la muestra $\{Y_i\}_{i=1}^n$ representa a las observaciones obtenidas en los puntos de diseño.

Para contrastar la hipótesis nula de linealidad

$$H_o: m \in \left\{ m_\theta(-) = \mathcal{A}^t\theta : \theta \in \Theta \subset \mathbb{R}^q \right\},\ \mathcal{A} \text{ un funcional de } \mathbb{R} \text{ en } \mathbb{R}^q$$

frente a la alternativa:

H_1: "m es una función con cierto grado de suavidad."

Se calculará la distancia entre una estimación no paramétrica, $\hat{m}_n(x)$, de la función de regresión, $m(x)$, y un estimador paramétrico de la misma función, bajo el supuesto de que se verifique H_o. En particular, se tomará una distancia del tipo Cramer-von Mises:

$$d^2(\hat{m}_n, H_o) = \min_{\theta\in\Theta} \int \left(\hat{m}_n(x) - \mathcal{A}^t(x)\theta\right)^2 \omega(x)\, d\Omega_n(x) = d^2\left(\hat{m}_n, \mathcal{A}^t(x)\hat{\theta}_n\right) \tag{2.2}$$

donde ω es una función de ponderación que se utiliza para evitar el efecto frontera y Ω_n es la distribución empírica de los puntos del diseño $\{x_i\}_{i=1}^n$.

Para estimar la función de regresión utilizaremos estimadores no paramétricos núcleo del tipo Gasser-Müller (1979):

$$(2.3) \qquad \hat{m}_n(x) = \sum_{j=1}^{n} h^{-1} \left(\int_{s_{j-1}}^{s_j} K\left(\frac{x-s}{h}\right) \mathrm{d}s \right) Y_j = \sum_{j=1}^{n} W_{n,j}(x) Y_j$$

con $s_0 = 0$, $s_{j-1} \leq x_j \leq s_j$, $s_n = 1$, donde K es una función núcleo: una función de densidad simétrica y de soporte compacto.

Bajo la hipótesis nula $\mathrm{H_0}$ de linealidad, es posible estimar θ con diversos criterios "**root$-n$**", esto es, que la convergencia en probabilidad del estimador al parámetro sea de orden $n^{-\frac{1}{2}}$ $\left(\sqrt{n}(\hat{\theta}-\theta) = O_{\mathbb{P}}(1)\right)$, por ejemplo, el mínimo cuadrático bajo ciertas condiciones de regularidad (Seber-Wild (1989), cap. 6). Pero hemos considerado más conveniente utilizar un criterio de mínima distancia, que resulta de la optimización de (2.2). Una estimación de este tipo parece más natural para un posterior constraste de la hipótesis $\mathrm{H_0}$.

Por tanto, el estimador $\hat{\theta}_n$ en (2.2) representa el estimador de mínima distancia asociado, y viene dado por la expresión:

$$(2.4) \quad \hat{\theta}_n = \mathbf{B}^{-1} \left(\sum_{\tau=1}^{n} \mathcal{A}(x_\tau)\omega(x_\tau)\hat{m}_n(x_\tau) \right) \quad \text{con } \mathbf{B} = \sum_{j=1}^{n} \mathcal{A}(x_j)\mathcal{A}^t(x_j)\omega(x_j)$$

obtenida de minimizar la función $\Psi(\theta) = \int \left(\hat{m}_n(x) - \mathcal{A}^t(x)\theta\right)^2 \omega(x)\, \mathrm{d}\Omega_n(x)$ en el conjunto $\Theta \subset \mathbb{R}^q$, siendo el valor de esta expresión en el mínimo, el estadístico del contraste de bondad de ajuste, d^2. Además, como se prueba en el teorema 1, abajo indicado, la estimación de mínima distancia también es root$-n$ bajo condiciones no muy restrictivas.

2.2. Resultados asintóticos

Para el estudio asintótico de los estimadores de mínima distancia y del estadístico d^2, consideraremos las siguientes hipótesis además de las ya indicadas:

A1. La función m es dos veces diferenciable, con derivadas acotadas, y ω tiene soporte compacto contenido en (0,1).

A2. La función núcleo K es derivable continuamente.

A3. La serie de errores $\{\epsilon_i\}$ tiene función de autocovarianza:

$\gamma(|i-j|) = \sigma_e^2 \sum_{\ell=0}^{\infty} b_\ell b_{\ell+|i-j|}$ absolutamente sumable, esto es, $\sum_{k=-\infty}^{+\infty} |\gamma(k)| < \infty$

A4. Para algún $\delta > 0$ se verifica que: $\mathrm{E}\left[|e|^{4+2\delta}\right] < \infty$.

A5. El coeficiente de curtosis de las variables del ruido blanco es cero.

A6. $nh \longrightarrow \infty$ y $h = o\left(n^{-\frac{\delta+2}{2\delta+2}}\right)$ cuando $n \longrightarrow \infty$.

Utilizando estas hipótesis y el Teorema Central del Límite para sucesiones con parte principal $m(n)$–dependiente debido a Nieuwenhuis (1992) se ha obtenido la distribución asintótica del estimador de mínima distancia, $\hat{\theta}_n$ y del estadístico d^2.

Teorema 1

Bajo la hipótesis nula, $\mathrm{H_0}$, y si se cumplen las hipótesis anteriores, se verifica:

a) $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathrm{N}\left(0, \sigma_\epsilon^2 \mathbf{Q}^{-1}\mathbf{Q}^*\mathbf{Q}^{-1}\right)$ donde $\mathbf{Q} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathcal{A}(x_i)\mathcal{A}^t(x_i)\omega(x_i)$

y $\mathbf{Q}^* = \sum_{u=-\infty}^{\infty} \gamma(u)R(u)$, con $R(u) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n-u} \mathcal{A}(x_j)\mathcal{A}^t(x_{j+u})\omega(x_j)\omega(x_{j+u})$
(límites que se supone que existen).

b) $\sqrt{n^2 h}\left(d^2\left(\hat{m}, \mathcal{A}^t(-)\hat{\theta}_n\right) - \frac{\overline{\omega}\sigma_\epsilon^2 \int K^2}{nh}\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_d^2\right)$ con $\sigma_\epsilon^2 = \mathrm{E}(\epsilon^2)$;

$\sigma_d^2 = 2\left(\sum_{k=-\infty}^{\infty} \gamma(k)\right)^2 \int (K*K)^2 \int \omega^2$; y $\overline{\omega} = \frac{1}{n}\sum_{i=1}^{n} \omega(x_i)$ siendo $K*K$ la convolución de la función núcleo consigo misma.

2.3. Comentarios

Comentario 1. Las hipótesis A1. y A2. son clásicas en la literatura relativa a la estimación no paramétrica. El carácter absolutamente sumable de la función de autocovarianzas es verificado, por ejemplo, por todos los modelos ARMA (ver

(Brockwell-Davis (1991), pág. 219). La hipótesis A5. se incluye únicamente para una simplificación de los cálculos de la parte b) del teorema que se expone más adelante. Finalmente la hipótesis A6. indica que la ventana h ha de ser elegida con un orden que fluctúa entre n^{-1} y $n^{-\frac{\delta+2}{2\delta+2}}$. En particular, esta hipótesis implica que $nh^4 \longrightarrow 0$, lo que es necesario para la cancelación de sesgos.

Comentario 2. Obsérvese que bajo la ausencia de correlación serial, $\gamma(k) = 0$ para todo $k = 0$ y, por tanto, $\sigma_d^2 = 2\sigma_\epsilon^4 \int (K * K)^2 \int \omega^2$, extendiéndose el resultado de González Manteiga y Cao Abad (1991) obtenido en el contexto de datos independientes. En este trabajo se utilizan hipótesis análogas a las utilizadas aquí, siendo las diferencias en el orden de elección del parámetro de suavización y el orden del momento del error muestral que tiene que ser acotado, y sobre todo, que allí se trabaja bajo hipótesis de independencia de las observaciones, lo que hace que la demostración de los resultados obtenidos en este trabajo sea más compleja.

Comentario 3. El estimador mínimo cuadrático, $\hat{\theta}_{mc}$, es un caso particular del estimador de mínima distancia, $\hat{\theta}$, que corresponde a una elección degenerada del parámetro de suavización: $h \cong 1/n$ en (2.3), para el estimador piloto usado en (2.2), y produce una mayor frecuencia de rechazo de las hipótesis cuando $d^2(\hat{m}_n, \mathrm{H_o})$ es usado como estadístico para el contraste de $\mathrm{H_o}$. Además, existen elecciones del parámetro de suavización no degeneradas para las que se verifica:

$$d^2\left(\hat{m}_n, \mathcal{A}^t(-)\hat{\theta}_n\right) < d^2\left(\hat{m}_n, \mathcal{A}^t(-)\hat{\theta}_{mc}\right)$$

lo que justifica la utilización del estimador $\hat{\theta}$ en lugar de $\hat{\theta}_{mc}$ para el cálculo de d^2. Por otra parte, una elección del parámetro de suavización del orden $1/n$ convierte a d^2 en el clásico test de bondad de ajuste basado en los residuos (F−test).

3. ESTUDIOS DE SIMULACIÓN

3.1. Influencia de la dependencia. Un ejemplo de simulación

La consideración de la dependencia en los errores asociados a los modelos (1.1) o (2.1) cuando ésta existe es de notable importancia. Por ejemplo, si se utiliza el resultado asintótico dado en b) para contrastar la hipótesis nula de linealidad $\mathrm{H_o}$ se rechaza al nivel α si:

$$(3.1) \qquad \sqrt{n^2 h}\left(d^2\left(\hat{m}_n, \mathcal{A}^t(-)\hat{\theta}_n\right) - \frac{\overline{\omega}\sigma_\epsilon^2 \int K^2}{nh}\right) > z_\alpha \sigma_d$$

donde z_α es tal que $\Phi(z_\alpha) = 1 - \alpha$ (Φ es la función de distribución de una N(0,1)). Es evidente, por tanto, que la omisión de la posible estructura de dependencia de los errores puede hacer que se rechace la hipótesis más veces que las correspondientes al nivel que si, por ejemplo, hay correlación serial positiva.

Tabla 1

$\sigma_\epsilon^2 = 0.1$								
	$\alpha = 0.10$				$\alpha = 0.05$			
h	.018	.020	.024	.025	.0150	.020	.025	.028
$\rho = 0$	.9426	.8662	.8116	.7872	1.0000	.9490	.8742	.8014
$\rho = .2$	.8588	.7362	.6580	.6254	1.0000	.8768	.7462	.6500
$\rho = .4$	.7278	.5394	.4672	.4386	1.0000	.7294	.5660	.4654
$\rho = .6$	.5938	.3702	.2980	.2716	1.0000	.5438	.3790	.2938
$\rho = .8$	.5646	.2484	.1754	.1540	1.0000	.3646	.2220	.1602

$\sigma_\epsilon^2 = 1$								
	$\alpha = 0.10$				$\alpha = 0.05$			
h	.0100	.0200	.0500	.0600	.0300	.0350	.0500	.0600
$\rho = 0$	1.0000	.9918	.9726	.9714	.9962	.9932	.9874	.9882
$\rho = .2$	1.0000	.9494	.9076	.9190	.9742	.9598	.9480	.9530
$\rho = .4$	.9998	.8040	.7572	.7882	.8826	.8512	.8454	.8658
$\rho = .6$	.9994	.5456	.5248	.5834	.6616	.6176	.6356	.6786
$\rho = .8$	.9992	.3290	.3066	.3652	.4048	.3724	.3964	.4540

Este efecto se observa en el siguiente estudio de simulación: se han generado 5.000 muestras de tamaño $n = 25$ del modelo $Y_1 = m(x_i) + \epsilon_i = 1 + 2x_i + x_i^2 + \epsilon_i$, $i = 1, \ldots, n$; en el que $x_i = i/n$ y $\epsilon_i = \rho\epsilon_{i-1} + e_i$ con $e_i \in \mathrm{N}(0, \sigma_e^2)$. Es decir, ϵ_i sigue una estructura del tipo $AR(1)$ con $|\rho| < 1$. Para cada una de las muestras se contrasta a nivel α, suponiendo que los errores son independientes, la hipótesis

nula H_0: "m es un polinomio de grado menor o igual que dos", presentándose en la Tabla 1 adjunta la frecuencia de aceptación de la hipótesis para distintos valores de α, ρ, h y σ^2. Aquí y en lo que sigue $\hat{m}_n$ es un estimador tipo Priestley-Chao (1972) con núcleo gaussiano y función de ponderación, $\omega(x) = 1$.

En esta tabla se observa claramente que al aumentar la dependencia de los errores del modelo disminuye progresivamente la frecuencia de aceptación, ocurriendo de forma drástica para valores altos de la correlación, $\rho = 0.6$ ó 0.8, lo que confirma lo expuesto anteriormente. También destacar la gran influencia en los resultados del parámetro de suavización, h, elegido, problema que tratamos en el apartado siguiente. Para la obtención de la Tabla 1, se ha elegido h de forma empírica, después de distintas pruebas se han tomado los h que mejores resultados proporcionaban y se exponen resultados con otros valores de h para comparar.

En todo caso debe de tenerse en cuenta que se están calculando proporciones de rechazo utilizando la distribución asintótica y estimando parámetros (las covarianzas) lo que hace que trabajando con muestras finitas los resultados no coincidan con los teóricos.

3.2. Consideraciones en base a estudios de simulación sobre la selección del parámetro de suavización

En este apartado se exponen los resultados de un amplio estudio de simulación realizado con un objetivo doble. En primer lugar analizar el buen comportamiento del estadístico introducido en el apartado 2 y en segundo lugar buscar orientaciones sobre la forma de elegir el parámetro de suavización, h.

En la primera parte del estudio de simulación se han generado 2.000 muestras de tamaño 50 de los modelos siguientes:

$$Y_i = m_j(x_i) + \epsilon_i, \quad x_i = i/50, \quad i = 1, \ldots, 50 \quad \text{para } j = 1, 2, 3$$

siendo

$$\begin{array}{lcll} m_1(x) & = & 1 + 2x + x^2 & \text{(verifica la hipótesis } H_0) \\ m_2(x) & = & 1 + 2x + x^2 + 5x^3 & \text{(no se verifica la hipótesis } H_0) \\ m_3(x) & = & 2 + 2\text{sen}\,(3\pi x/2) & \text{(no se verifica la hipótesis } H_0) \end{array}$$

y $\{\epsilon_i\}$ es el ruido con una correlación serial siguiendo una estructura del tipo MA con varianza $\sigma^2 = 0.10$ y 0.50. Como modelos de dependencia MA se han utilizado los siguientes:

$$\begin{array}{llcll} \textbf{DEP.1.:} & \epsilon_t & = & e_t + 0.9e_{t-1}, & \text{por tanto} \quad \rho_1 = 0.497 \\ \textbf{DEP.2.:} & \epsilon_t & = & e_t - 1.5e_{t-1} + 0.9e_{t-2}, & \text{por tanto} \quad \rho_1 = -0.7; \quad \rho_2 = 0.22 \end{array}$$

siendo $\{\epsilon_t\}$ ruido blanco gaussiano y $\{\rho_1, \rho_2\}$ los coeficientes de correlación a uno y dos retardos.

Se contrasta la hipótesis nula H_0: "m es un polinomio de grado menor o igual que dos". El contraste utilizado es el asintótico dado por (3.1) donde σ_d es el valor teórico dado por el Teorema 1 y $\hat{\theta}_n$ es el estimador de mínima distancia. En las Tablas 2 y 3 se presentan las proporciones de aceptación de H_0 para distintos h, los dos tipos de dependencias, y niveles de significación $\alpha = 0.10$ y 0.01. También se ha obtenido la media del nivel crítico del test.

Tabla 2

Dependencia 1

$\sigma^2 = 0.10$		Niv. Sign. $\alpha = 0.10$	Niv. Sign. $\alpha = 0.01$	Media Niv. **Crít.**
$h = .005$	m_1	0.103	0.316	0.0328
	m_2	0.053	0.208	0.0188
	m_3	0.000	0.000	0.0000
$h = .009$	m_1	0.759	0.967	0.2874
	m_2	0.591	0.891	0.2005
	m_3	0.001	0.012	0.0006
$h = .013$	m_1	0.627	0.934	0.2071
	m_2	0.053	0.358	0.0208
	m_3	0.000	0.005	0.0002
$h = .017$	m_1	0.392	0.786	0.1102
	m_2	0.000	0.003	0.0001
	m_3	0.000	0.000	0.0000

$\sigma^2 = 0.50$		Niv. Sign. $\alpha = 0.10$	Niv. Sign. $\alpha = 0.01$	Media Niv. **Crít.**
$h = .006$	m_1	0.303	0.665	0.0988
	m_2	0.273	0.625	0.0902
	m_3	0.068	0.240	0.0243
$h = .010$	m_1	0.809	0.980	0.3180
	m_2	0.759	0.968	0.2858
	m_3	0.368	0.728	0.1154
$h = .014$	m_1	0.774	0.972	0.2900
	m_2	0.586	0.911	0.1867
	m_3	0.295	0.640	0.0925
$h = .018$	m_1	0.732	0.960	0.2620
	m_2	0.392	0.781	0.1073
	m_3	0.231	0.567	0.0746

Tabla 3

Dependencia 2

$\sigma^2 = \mathbf{0.10}$		Niv. Sign. $\alpha = \mathbf{0.10}$	Niv. Sign. $\alpha = \mathbf{0.01}$	Media Niv. **Crít.**
$h = .008$	$\boldsymbol{m}_1$	0.273	0.284	0.2588
	$\boldsymbol{m}_2$	0.149	0.150	0.1345
	$\boldsymbol{m}_3$	0.000	0.000	0.0000
$h = .012$	$\boldsymbol{m}_1$	0.994	0.995	0.9926
	$\boldsymbol{m}_2$	0.117	0.176	0.1556
	$\boldsymbol{m}_3$	0.000	0.000	0.0000
$h = .016$	$\boldsymbol{m}_1$	0.995	0.997	0.9935
	$\boldsymbol{m}_2$	0.000	0.000	0.0000
	$\boldsymbol{m}_3$	0.000	0.000	0.0000
$h = .017$	$\boldsymbol{m}_1$	0.726	0.756	0.6871
	$\boldsymbol{m}_2$	0.000	0.000	0.0000
	$\boldsymbol{m}_3$	0.000	0.000	0.0000

$\sigma^2 = \mathbf{0.50}$		Niv. Sign. $\alpha = \mathbf{0.10}$	Niv. Sign. $\alpha = \mathbf{0.01}$	Media Niv. **Crít.**
$h = .008$	$\boldsymbol{m}_1$	0.281	0.293	0.2658
	$\boldsymbol{m}_2$	0.248	0.272	0.2360
	$\boldsymbol{m}_3$	0.009	0.010	0.0080
$h = .023$	$\boldsymbol{m}_1$	1.000	1.000	1.0000
	$\boldsymbol{m}_2$	0.236	0.255	0.1952
	$\boldsymbol{m}_3$	0.006	0.008	0.0037
$h = .038$	$\boldsymbol{m}_1$	1.000	1.000	1.0000
	$\boldsymbol{m}_2$	0.000	0.000	0.0000
	$\boldsymbol{m}_3$	0.000	0.000	0.0000
$h = .053$	$\boldsymbol{m}_1$	0.772	0.828	0.6919
	$\boldsymbol{m}_2$	0.000	0.000	0.0000
	$\boldsymbol{m}_3$	0.000	0.000	0.0000

En los resultados obtenidos en el estudio de simulación y detallados en las tablas anteriores se aprecia gran sensibilidad del test descrito respecto al parámetro de

suavización, h, ya que pequeñas variaciones de este parámetro llevan a resultados diferentes. Nuevamente, en las tablas obtenidas, se ha elegido el h de forma empírica, utilizando el que hemos considerado más adecuado, después de distintas pruebas, y otros valores próximos de forma que se pueda observar cómo influyen en los resultados.

Para estudiar con mayor detalle la influencia de este parámetro se ha hecho un segundo estudio de simulación en el que utilizando los mismos modelos de funciones y de dependencia de los errores que en la simulación anterior, se ha calculado la media del estadístico, d^2, en función de h, para 100 muestras simuladas de tamaño 50, que se denota por $\overline{d}^2(h)$ y los niveles críticos medios del test en dos situaciones: primero, utilizando las autocovarianzas teóricas de los errores, obteniendo la función CLT(h). Y, en segundo lugar, estimando las autocovarianzas de los errores a partir de los datos de cada muestra siguiendo la técnica de Müller-Stadmüller (1988), obteniendo la función CLE(h).

En las figuras 1-2-3-4-5 se han representado las funciones $\overline{d}^2(h)$, CLT(h) y CLE(h) bajo distintas situaciones y se observa que una buena elección de h para la hipótesis nula puede consistir en tomar el mínimo local más pequeño resultante de optimizar:

$$\min_h \left\{ d^2 \left(\hat{m}_{n,h}, \mathcal{A}^t(-)\hat{\theta}_{n,h} \right) \right\}$$

ya que a este valor de h le corresponde el mayor nivel crítico. Obsérvese que el mínimo es local ya que cuando $h \longrightarrow \infty$, se verifica que $d^2 \longrightarrow 0$, pues en ese caso $\hat{m}(x)$ es una curva sobresuavizada que tiende a ser constante. También de las gráficas representadas se deduce que cuando se utiliza el test con autocovarianzas estimadas los resultados son bastantes similares a los obtenidos con autocovarianzas teóricas.

En hipótesis alternativas, como la dada por los modelos $m_2(x)$ y $m_3(x)$ la h elegida por el criterio anterior puede no ser la adecuada como se puede apreciar en las Figuras 4–5, donde se observa una tendencia de aceptación de la hipótesis para dicho valor de h.

En todo caso, las conclusiones anteriores se han obtenido en base a un estudio de simulación, que aunque amplio, no es exhaustivo, y permanecen abiertas muchas cuestiones sobre el tema. Por ejemplo, si se puede asegurar la existencia de un mínimo local en la curva $d^2(h)$. Por todo ello y teniendo en cuenta que una correcta elección de h es crucial, será necesaria una amplia investigación sobre este particular, principalmente, bajo hipótesis alternativas, ya que bajo H_0 la sugerencia hecha parece funcionar adecuadamente, aunque es costosa desde un punto de vista computacional.

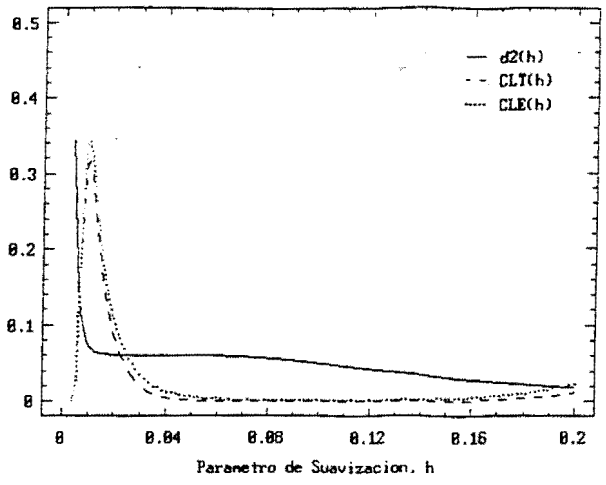


Figura 1.

Curvas $\overline{d}^2(h)$ (media de las d^2 para 100 muestras), CLT(h) (nivel crítico medio con autocovarianzas de los errores conocidas) y CLE(h) (autocovar. estimadas) para $m_1(x) = 1 + 2x + x^2$, con **DEP.1.** y $\sigma^2 = 0.10$.

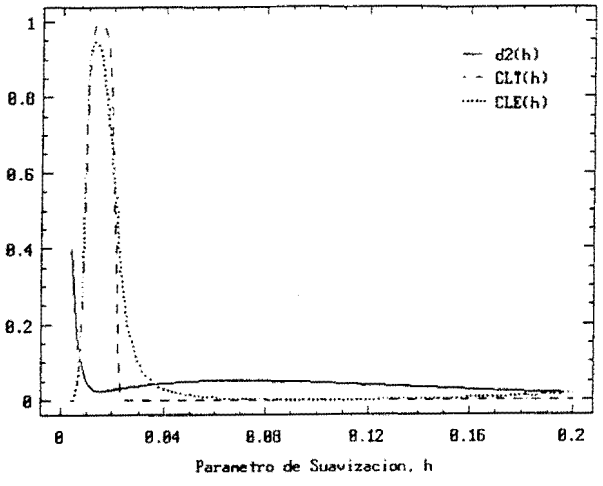


Figura 2.

Curvas $\overline{d}^2(h)$ (media de las d^2 para 100 muestras), CLT(h) (nivel crítico medio con autocovarianzas de los errores conocidas) y CLE(h) (autocovar. estimadas) para $m_1(x) = 1 + 2x + x^2$, con **DEP.2.** y $\sigma^2 = 0.10$.

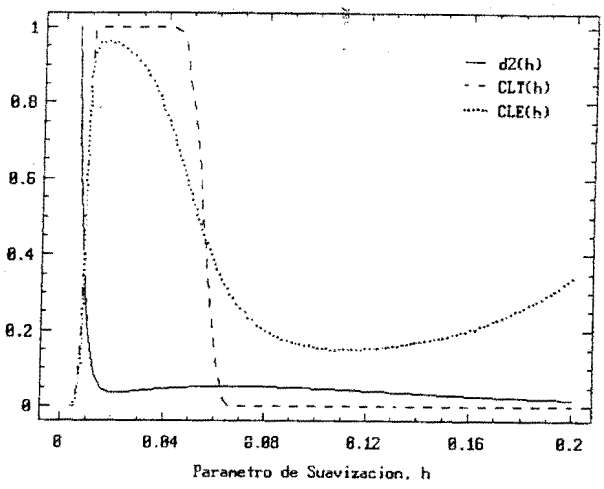


Figura 3.

Curvas $\overline{d}^2(h)$ (media de las d^2 para 100 muestras), CLT(h) (nivel crítico medio con autocovarianzas de los errores conocidas) y CLE(h) (autocovar. estimadas) para $m_1(x) = 1 + 2x + x^2$, con **DEP.2.** y $\sigma^2 = 0.50$.

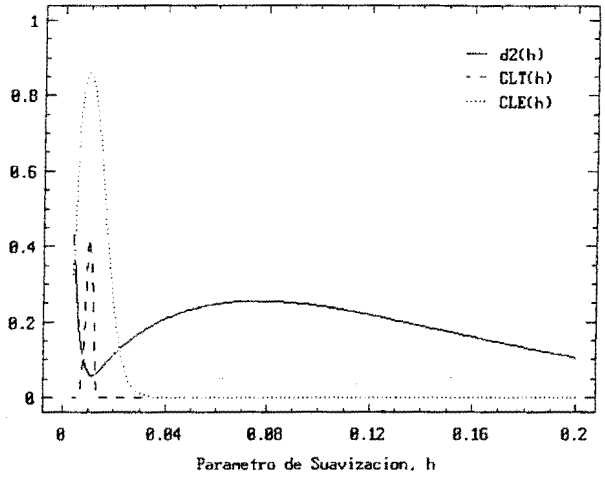


Figura 4.

Curvas $\overline{d}^2(h)$ (media de las d^2 para 100 muestras), CLT(h) (nivel crítico medio con autocovarianzas de los errores conocidas) y CLE(h) (autocovar. estimadas) para $m_2(x) = 1 + 2x + x^2 + 5x^3$, con **DEP.2.** y $\sigma^2 = 0.10$.

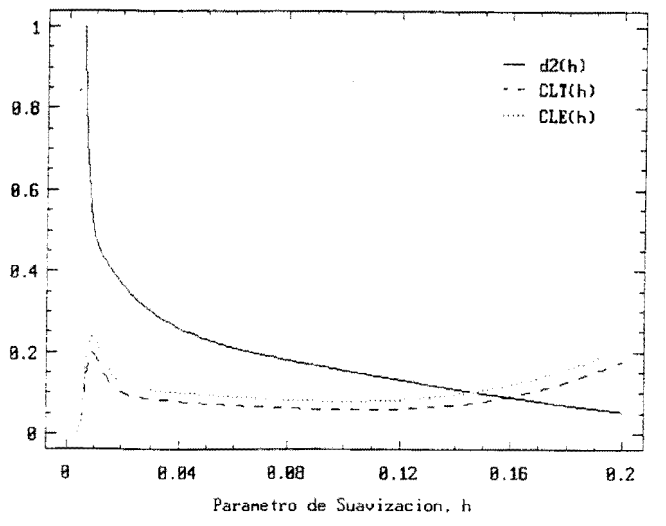


Figura 5.

Curvas $\overline{d}^2(h)$ (media de las d^2 para 100 muestras), CLT(h) (nivel crítico medio con autocovarianzas de los errores conocidas) y CLE(h) (autocovar. estimadas) para $m_3(x) = 2 + 2\text{sen}\,(3\pi x/2)$, con **DEP.1.** y $\sigma^2 = 0.50$.

4. ESTUDIO DE UN CASO CON DATOS REALES

Se estudiará en este apartado un ejemplo en el que a partir de una muestra de datos reales se aplica el contraste diseñado para modelos con posible correlación serial entre los errores. Como variable respuesta Y_i se considera "la tasa de cobertura bruta de las prestaciones por desempleo total en Galicia", definida como sigue:

$$\text{T.C.B.} = \frac{\begin{array}{c}\textit{número de beneficiarios de prestaciones por} \\ \textit{desempleo total y asistencias económicas}\end{array}}{\textit{Paro registrado}} \times 100$$

La muestra utilizada abarca el período de Enero de 1985 a Abril de 1992 (88 datos, facilitados por el Instituto Galego de Estadística, Xunta de Galicia y publicados en los Boletines de Series Estadísticas de Galicia). La variable regresora, $x_i = i/88$, $i = 1, \ldots, 88$, es el tiempo normalizado en el intervalo [0,1]. En la figura 6 se ha representado la serie $\{Y_i\}_{i=1}^{88}$ de datos observados a lo largo del tiempo. Y del análisis de las funciones de autocorrelación y autocorrelación parcial de la muestra se deduce utilizando la metodología Box-Jenkins que se puede realizar una buena aproximación por un modelo ARIMA con componente estacional y, por tanto, por un modelo MA(∞).

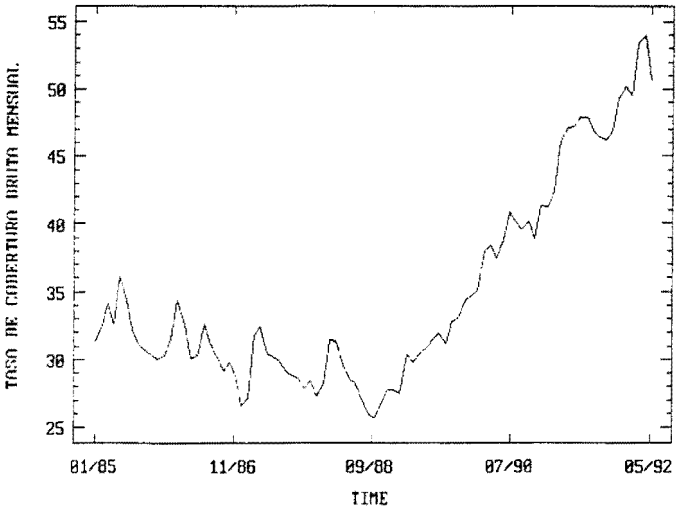


Figura 6.

Datos de la Tasa de Cobertura Bruta mensual en Galicia de Enero de 1985 a Abril de 1992.

Si se realiza un ajuste mediante regresión polinómica de grado cinco entre la variable Y_i y el tiempo $x_i = i/88$, $i = 1, \ldots, 88$:

$$Y_i = m_\theta(x_i) + \epsilon_i = \theta_1 + \theta_2 x_i^2 + \theta_3 x_i^3 + \theta_4 x_i^4 + \theta_5 x_i^5 + \epsilon_i, \quad i = 1, \ldots, 88$$

suponiendo una estructura de independencia y homocedasticidad para los errores con estructura gaussiana, resulta la siguiente tabla adjunta:

Tabla 4

Variables Independientes	Estimación	$t-$valor	Nivel de Significación
Término Constante	32.804	28.12	0.0000
x_i	-1.438	-0.06	0.9500
x_i^2	-33.380	-0.24	0.8095
x_i^3	-93.376	-0.27	0.7866
x_i^4	379.177	1.01	0.3135
x_i^5	-232.475	-1.57	0.1180
Coeficiente de correlación Ajustado = 0.9508		Durbin-Watson: 0.863	

De los resultados expuestos en esta tabla se deduce que el modelo explica de forma significativa la variable respuesta aunque los contrastes individuales (test t) no sean significativos, lo que se debe a que tienen varianzas muy altas. Esto ocurre con frecuencia en los modelos polinómicos ya que las variables $x, x^2, \ldots$ son fuertemente dependientes y provoca problemas de multicolinealidad. (Ver Peña, 1987).

Además tampoco es recomendable en la teoría de regresión polinómica utilizar modelos de grado elevado y en este ejemplo también se puede obtener un ajuste adecuado utilizando un polinomio de grado dos, esto es, haciendo un ajuste del tipo: $Y_i = m_\theta(x_i) + \epsilon_i = \theta_1 + \theta_2 x_i + \theta_3 x_i^2$ se obtienen los siguientes resultados (Tabla 5):

Tabla 5

Variables Independientes	Estimación	$t-$valor	Nivel de Significación
Término Constante	35.828	56.25	0.0000
x_i	-43.204	-14.86	0.0000
x_i^2	61.436	22.06	0.0000
Coeficiente de correlación Ajustado = 0.9312		Durbin-Watson: 0.615	

Como indica el Coeficiente de Correlación Ajustado el modelo de grado dos proporciona un ajuste casi tan bueno como el de grado cinco pero con la ventaja de que los parámetros del modelo son significativos ya que el problema de multicolinealidad no se presenta como al ajustar un polinomio de grado mayor. Por otra parte, del resultado obtenido para el contraste de Durbin-Watson se obtiene 0.615 valor muy inferior al 1.60 donde se acepta la hipótesis de independencia de los residuos al nivel de significación del 0,05. Además el contraste de Box-Ljung, cuyo valor es 118.58 nos confirma la dependencia de los residuos, lo que gráficamente puede observarse en la figura 7 en la que se ha representado la gráfica de la función de autocorrelación de los residuos.

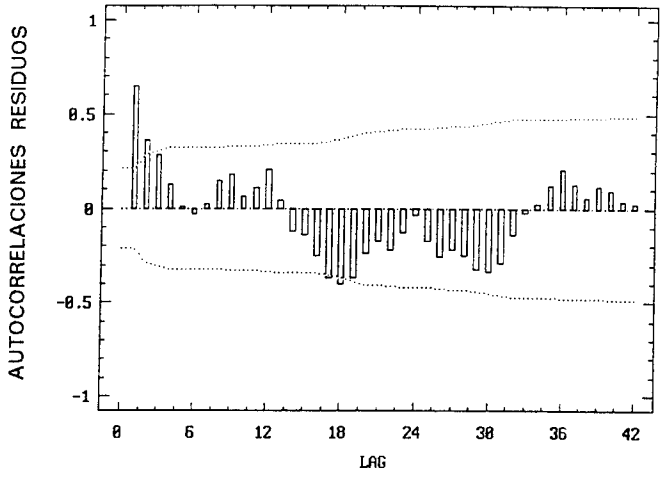


Figura 7.

Función de autocorrelación de los residuos al ajustar un polinomio de grado dos a la Tasa de Cobertura Bruta.

Para realizar un análisis mediante el test asintótico (3.1), con el que se contrasta la hipótesis:

H_0: "m_θ es un polinomio de grado menor o igual que dos"

se ha calculado el valor del estadístico $d^2(h)$ para valores del parámetro de suavizado, h, desde 0.0002 hasta 0.0300 con incrementos de 0.0002 suponiendo que los residuos siguen un modelo MA(20) y se ha calculado el nivel crítico (CL(h)) asociado a los valores del estadístico.

En la figura 8 se han representado las dos funciones obtenidas, $d^2(h)$ reescalada de manera adecuada y CL(h).

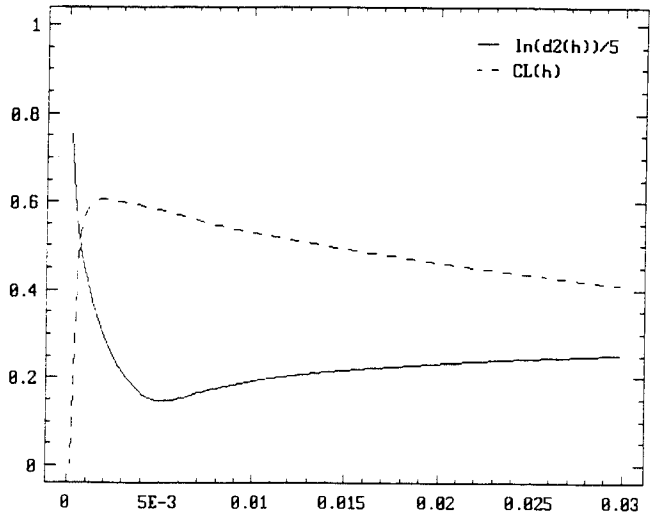


Figura 8.

Funciones $\ln(d^2(h))/5$ ($d^2(h)$: valor del estadístico) y CL(h) (Nivel crítico asociado al valor del estadístico).

De los resultados obtenidos se deduce que para un amplio intervalo de valores de h se aceptaría la hipótesis H_0 con un alto nivel de significación y si deseamos apoyar la hipótesis H_0 elegiríamos el h que minimiza la función $d^2(h)$ que coincide, aproximadamente, con el máximo de la función CL(h). En nuestro estudio esto ocurre para el valor de $h = 0.0052$. Para este valor de h y otros próximos se ha calculado el valor del estadístico, el del estadístico estandarizado en el supuesto de dependencia de los residuos, estimando las autocovarianzas según Müller-Stadmüller y el valor del estadístico estandarizado suponiendo estructura de independencia en los errores y en el que σ_ϵ^2 es estimada de forma estándar mediante la suma de residuos al cuadrado ajustada por el número de parámetros. Finalmente, se han calculado los niveles de significación asociados a los dos estadísticos tipificados. Los resultados obtenidos se recogen en la tabla 6.

Tabla 6

H_0: "m_θ es un polinomio de grado ≤ 2"

		Dependencia			Independencia	
Banda: h	Estadístico: d^2	Estad. tip.	Niv. Sign.		Est. tip.	Niv. Sig.
0.0010	75.215	−0.199	0.579		−14.102	1
0.0030	8.387	−0.254	0.601		−7.963	1
0.0052	**4.315**	**−0.199**	**0.579**		**2.905**	**0.002**
0.0070	5.165	−0.146	0.559		10.458	0
0.0090	6.332	−0.095	0.538		16.028	0
0.0110	7.262	−0.053	0.522		20.449	0

De los resultados obtenidos se deduce que bajo la hipótesis de dependencia de los residuos se aceptaría la hipótesis de que m_θ es un polinomio de grado menor o igual que dos para distintos valores de h y claramente para el h que minimiza el estadístico d^2, sin embargo si utilizásemos la hipótesis de independencia de los residuos se rechazaría la hipótesis H_0 para la mayoría de los valores de h aceptándose solamente para valores muy pequeños de h. Esto pone de manifiesto la importancia de considerar la posible estructura de dependencia de los residuos así como la sensibilidad del estadístico respecto al parámetro h, donde mucha teoría sobre su elección queda todavía por desarrollar.

Finalmente, comentar que el polinomio estimado, de grado dos, por mínima distancia, para $h = 0.0052$ viene dado por:

$$m(x) = 34.89 - 36.06x + 52.67x^2$$

los residuos de este ajuste pueden modelarse adecuadamente, según la metodología de Box-Jenkins por un AR(1) de ecuación:

$$\epsilon_t = 0.29\epsilon_{t-1} + e_t$$

5. ANEXO. DEMOSTRACIONES

En este apartado se realiza una descripción, no detallada, de las demostraciones de los apartados a) y b) del teorema 1, que son bastante laboriosas y

técnicas. Se basan en la utilización del siguiente teorema de Nieuwenhuis (1992) sobre la normalidad asintótica de disposiciones triangulares con parte principal $m(n)-$dependiente.

Teorema 2

Considérese la disposición triangular $\{X_{\tau n}\}_{\tau=1}^{n}$ que admite la siguiente descomposición: $X_{\tau n} = X_{\tau n,m(n)} + \overline{X}_{\tau n,m(n)}$.

Siendo la parte principal $\left(X_{\tau n,m(n)}\right)$ $m(n)-$dependiente Y $\left(\overline{X}_{\tau n,m(n)}\right)$ es la parte residual.

Denominemos $B_n^2 = \text{Var}\left(\sum_{\tau=1}^{n} X_{\tau n}\right)$. Y supongamos que las disposiciones triangulares $\left(\frac{X_{\tau n}}{B_n}\right)$ y $\left(\frac{\overline{X}_{\tau n,m(n)}}{B_n}\right)$ verifican, respectivamente, las condiciones C_1 y C_1^*:

Condición C_1: $\max_{i<j\leq n} \frac{1}{j-i} \text{Var}\left(\frac{1}{B_n}\sum_{k=i+1}^{j} X_{kn}\right) = O\left(\frac{1}{n}\right)$ cuando $n \longrightarrow \infty$

Condición C_1^*: $\max_{i<j\leq n} \frac{1}{j-i} \text{Var}\left(\frac{1}{B_n}\sum_{k=i+1}^{j} \overline{X}_{kn,m(n)}\right) = o\left(\frac{1}{n}\right)$ cuando $n \longrightarrow \infty$

Y ambas satisfacen la condición C_2 para algún $\delta > 0$.

Condición C_2: $\max_{i<\tau\leq n} \frac{1}{B_n^{2+\delta}} \text{E}\,|Y_{\tau n}|^{2+\delta} = O\left(\frac{1}{n^{1+(\delta/2)}}\right)$ cuando $n \longrightarrow \infty$

(siendo $Y_{\tau n} = x_{\tau n}$ o $Y_{\tau n} = x_{\tau n,m(n)}$). Entonces:

$$\frac{1}{B_n}\sum_{\tau=1}^{n}(X_{\tau n} - \text{E}\,X_{\tau,n}) \xrightarrow{d} \text{N}(0,1) \quad \text{cuando } n \longrightarrow \infty$$

Demostración del Teorema 1.a)

Dado que el estimador $\hat{\theta}_n$ asociado a (2.2) viene dado por:

$$\hat{\theta}_n = B^{-1}\left(\sum_{\tau=1}^{n}\mathcal{A}(x_\tau)\omega(x_\tau)\hat{m}_n(x_\tau)\right) \quad \text{con } B = \sum_{j=1}^{n}\mathcal{A}(x_j)\mathcal{A}^t(x_j)\omega(x_j)$$

bajo la hipótesis H_o se verifica que:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \sqrt{n}\left\{B^{-1}\left(\sum_{i=1}^{n}\sum_{\tau=1}^{n}\mathcal{A}(x_i)\mathcal{A}^t(x_\tau)\omega(x_i)W_{n,\tau}(x_i)\theta\right) - \theta\right\} +$$
$$+ \sqrt{n}\left\{B^{-1}\left(\sum_{i=1}^{n}\sum_{\tau=1}^{n}\mathcal{A}(x_i)\omega(x_i)W_{n,\tau}(x_i)\epsilon_\tau\right)\right\} = \Delta_1 + \Delta_2$$

donde Δ_1 es el término debido al sesgo.

Bajo las hipótesis A1–A2 y A6. y usando las expresiones para el sesgo del estimador Gasser-Müller:

$\sum_{\tau=1}^{n}\mathcal{A}^t(x_\tau)W_{n,\tau}(x_i) - \mathcal{A}^t(x_i) = O(h^2)$, de forma uniforme en i, en el soporte de ω, y, por tanto:

$$\Delta_1 = \sqrt{n}\left\{nB^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{A}(x_i)\omega(x_i)\right)\right\}O(h^2) = O\left(\sqrt{n}h^2\right) \longrightarrow 0$$

ya que $nh^4 \longrightarrow 0$.

Por otro lado para Δ_2 se obtiene utilizando A6.,

$$\begin{aligned}\Delta_2 &= \sqrt{b}B^{-1}\sum_{\tau=1}^{n}\left(\sum_{i=1}^{n}\mathcal{A}(x_i)\omega(x_i)W_{n,\tau}(x_i)\right)\epsilon_\tau = \\ &= \sqrt{n}B^{-1}\sum_{\tau=1}^{n}\left(\mathcal{A}(x_\tau)\omega(x_\tau) + O(h^2) + O\left(\frac{1}{nh}\right)\right)\theta_\tau = \\ &= \sqrt{n}B^{-1}\sum_{\tau=1}^{n}\mathcal{A}(x_\tau)\omega(x_\tau)\epsilon_\tau + o_{\mathbb{P}}(1)\end{aligned}$$

Por tanto basta con probar la normalidad asintótica de

$$\sqrt{n}B^{-1}\sum_{\tau=1}^{n}\mathcal{A}(x_\tau)\omega(x_\tau)\theta_\tau$$

Para ello utilizando el Teorema 2 se demuestra la normalidad asintótica de

$$\sum_{\tau=1}^{n}c^t\mathcal{A}(x_\tau)\omega(x_\tau)\epsilon_\tau = \sum_{\tau=1}^{n}d_{\tau n}\epsilon_\tau = \sum_{\tau=1}^{n}d_{\tau n}\left(\sum_{j=0}^{\infty}b_j e_{\tau-j}\right) =$$

$$= \sum_{\tau=1}^{n}X_{\tau n}, \quad \forall c \in \mathbb{R}^q \tag{5.1}$$

Realicemos la siguiente descomposición $X_{\tau n} = X_{\tau n,m(n)} + \overline{X}_{\tau n,m(n)}$, siendo

$$\begin{aligned} X_{\tau n,m(n)} &= \sum_{j=0}^{m(n)} d_{\tau n} b_j e_{\tau-j} \qquad \text{(parte principal)} \\ \overline{X}_{\tau n,m(n)} &= \sum_{j=m(n)+1}^{\infty} d_{\tau n} b_j e_{\tau-j} \quad \text{(parte residual)} \end{aligned}$$

Calculemos, en primer lugar, B_n^2

$$B_n^2 = \text{Var}\left(\sum_{\tau=1}^{n} X_{\tau n}\right) = \sum_{\tau=1}^{n}\sum_{\tau=1}^{n} d_\tau d_\tau,\ \gamma(\tau-\tau') = c^t \sum_{j=1}^{n}\sum_{\ell=j-n}^{n-1} a_{j\ell} c$$

Y utilizando A3. y el teorema de la convergencia dominada se obtiene,

$$\frac{B_n^2}{n} \longrightarrow c^t \sum_{n=-\infty}^{\infty} \gamma(n) R(n) c \tag{5.2}$$

Se prueba a continuación que las sucesiones $\left\{\frac{X_{\tau n}}{B_n}\right\}$ y $\left\{\frac{\overline{X}_{\tau n,m(n)}}{B_n}\right\}$ verifican las condiciones del teorema 2.

$\boxed{(C_1)}$ $\dfrac{1}{j-i}\,\text{Var}\left(\sum_{k=i+1}^{j}\dfrac{X_{\tau n}}{B_n}\right) \leq \dfrac{1}{j-i}\sum_{k=i+1}^{j}\sum_{k'=i+1}^{j}|\gamma(k-k')|\,O\left(\dfrac{1}{B_n^2}\right) \leq$

$$\leq \sum_{h=-(j-i-1)}^{j-(i+1)}\left[1-\frac{|h|}{j-i}\right]|\gamma(h)|O\left(\frac{1}{B_n^2}\right)$$

y dado que cuando $(j-i) \longrightarrow \infty$ se sigue que

$$\sum_{h=-(j-i-1)}^{j-(i+1)}\left[1-\frac{|h|}{j-i}\right]|\gamma(h)| \longrightarrow \sum_{h=-\infty}^{\infty}|\gamma(h)|$$

se obtiene que cuando $n \longrightarrow \infty$

$$\max_{1<j\leq n}\frac{1}{j-i}\,\text{Var}\left(\sum_{k=i+1}^{j}\frac{X_{kn}}{b_n}\right) = O\left(\frac{1}{B_n^2}\right) = O\left(\frac{1}{n}\right)$$

$\boxed{(C_1^*)}$ $\left|\frac{1}{j-i}\,\mathrm{Var}\left(\sum_{k=i+1}^{j}\frac{\overline{X}_{kn,m(n)}}{B_n}\right)\right| \leq$

$$\leq \sum_{h\in\mathbb{Z}}\left|\mathrm{Cov}\left(\sum_{j=m(n)+1}^{\infty} b_j e_{1-j},\ \sum_{j=m(n)+1}^{\infty} b_j e_{1+h-j}\right)\right| O(1) \leq$$

$$\leq \sigma_e^2 O(1)\sum_{h\in\mathbb{Z}}\sum_{\substack{j>m(n)\ \ell>m(n)\\ \{\ell-j=h\}}}\sum |b_j||b_\ell| = \sigma_e^2 O(1)\left(\sum_{k>m(n)}|b_k|\right)^2 \longrightarrow 0$$

cuando $n \longrightarrow \infty$, , uniformemente en $i < j \leq n$.

$\boxed{(C_2)}$ Aplicando el Lema de Fatou, la desigualdad de Jensen y A4., se obtiene

$$\mathrm{E}|X_{\tau n}|^{2+\delta} \leq O(1)\ \underline{\lim}\ \mathrm{E}\left(\left|\sum_{j=0}^{m(n)}|b_j||e_{\tau-j}|\right|^{2+\delta}\right) < \infty$$

Y, dado que, $\frac{1}{B_n^{2+\delta}} = O\left(\frac{1}{n^{1+\delta/2}}\right)$ se obtiene que la sucesión $\left\{\frac{X_{\tau n}}{B_n}\right\}$ verifica la condición C_2. Y utilizando la desigualdad de Minkwosky se obtiene que dicha condición también la verifica la sucesión $\left\{\frac{\overline{X}_{\tau n,m(n)}}{B_n}\right\}$.

De todo lo anterior se sigue que $\frac{1}{B}\sum_{\tau=1}^{n} X_{\tau n}$ converge en distribución a una N(0,1). Ahora, aplicando el teorema de Cramer-Wold se sigue la conclusión del teorema 1.a).

■

Demostración del Teorema 1.b)

Por el resultado anterior, Teorema 1.a) se verifica:

$$d^2\left(\hat{m}_n, \mathcal{A}^t(-)\hat{\theta}_n\right) = \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n} W_{ni}(x_j)Y_i - \mathcal{A}^t(x_j)\theta + O_{\mathbb{P}}\left(n^{-1/2}\right)\right)^2 \omega(x_j) =$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n} W_{ni}(x_j)Y_i - \mathcal{A}^t(x_j)\theta\right)^2 \omega(x_j) +$$

$$+\frac{2}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}W_{ni}(x_j)Y_i - \mathcal{A}^t(x_j)\theta\right)^2 \omega(x_j)O_{\mathbb{P}}\left(n^{-1/2}\right) + O_{\mathbb{P}}\left(n^{-1/2}\right) =$$

$$= \Delta_1 + \Delta$$

donde Δ engloba a los dos últimos sumandos y será un término despreciable para el estudio asintótico.

Por lo que respecta a Δ_1:

$$\begin{aligned}\Delta_1 &= \frac{1}{n}\sum_{j=1}^{n}b^2(\theta,x_j)\omega(x_j) + \sum_{i=1}^{n}b_{nii}\epsilon_i^2 + 2\sum_{\tau<k}b_{n\tau k}\epsilon_\tau\epsilon_k + \\ &+ \frac{2}{n}\sum_{j=1}^{n}b(\theta,x_j)\left(\sum_{i=1}^{n}W_{ni}(x_j)\epsilon_i\right)\omega(x_j) = \Delta_{11}+\Delta_{12}+\Delta_{13}+\Delta_{14}\end{aligned}$$

donde

$$b(\theta,x_j) = \sum_{i=1}^{n}W_{ni}(x_j)\mathcal{A}^t(x_i)\theta - \mathcal{A}^t\theta$$

y

$$b_{n\tau k} = \frac{1}{n}\sum_{j=1}^{n}W_{n\tau}(x_j)W_{nk}(x_j)\omega(x_j)$$

Bajo las hipótesis A1., A2. y A6.,

$$\sqrt{n^2h}\,\Delta_{11} = O\left(nh^{9/2}\right)\longrightarrow 0,$$

(5.3)

$$\mathrm{E}(\Delta_{12}) = \frac{\sigma_\epsilon^2}{nh}\int K^2\overline{\omega} + O\left(\frac{1}{n}\right)$$

Además, por A5.,

$$\mathrm{Var}(\Delta_{12}) = 2\sum_i\sum_j b_{nii}b_{njj}\gamma^2(i-j)$$

Por otro lado y teniendo en cuenta que las hipótesis sobre K permiten acotar las diferencias entre los pesos del tipo Gasser-Müller y Priestley-Chao y utilizando sumas de Riemann se obtiene:

$$\mathrm{Var}(\Delta_{12}) \simeq 2\cdot\frac{1}{n^4h^2}\sum_i\sum_j\omega(x_i)\omega(x_j)\gamma^2(i-j)\left(\int K^2\right)^2 \simeq O\left(\frac{1}{n^3h^2}\right)$$

De esto y como consecuencia de (5.3), se sigue:

$$\Delta_{12} = \frac{\sigma_\epsilon^2}{nh}\int K^2\overline{\omega} + O_{\mathbb{P}}\left(\frac{1}{n^{(3/2)}h}\right) \tag{5.4}$$

Por lo que respecta a Δ_{13}, con razonamientos similares a los utilizados anteriormente se obtiene:

$$\mathrm{Var}(\Delta_{13}) \simeq \frac{2}{n^2h}\left(\sum_{k=-\infty}^{\infty}\gamma(k)\right)^2 (K*K)^2\int\omega^2 = \sigma_d^2 \tag{5.5}$$

Finalmente para Δ_{14}, aplicando la desigualdad de Hölder se obtiene:

$$\mathrm{Var}\,(\Delta_{14}) = \mathrm{E}\left(\frac{2}{n}\sum_{j=1}^{n} b(\theta,x_j)\left(\sum_{i=1}^{n} W_{ni}(x_j)\epsilon_i\right)\omega(x_j)\right)^2 \leq O(h^4)\sum_i\sum_\tau b_{ni\tau}\gamma(i-\tau)$$

Y utilizando aproximaciones de Riemann se obtiene:

$$\begin{aligned}\mathrm{Var}\,(\Delta_{14}) &\simeq O(h^4)\sum_i\sum_\tau\frac{1}{n^2h}\left((K*K)\left(\frac{x_\tau - x_i}{h}\right)\right)\gamma(i-\tau)\omega(x_i) \leq \\ &\leq O(h^4)\sum_i\sum_\tau\frac{1}{n^2h}\cdot\gamma(i-\tau) = O\left(\frac{h^4}{nh}\right)\end{aligned}$$

por la sumabilidad de γ y acotación de K y ω. De este modo:

$$\sqrt{n^2h}\Delta_{14} = O_{\mathbb{P}}\left(\sqrt{n}h^2\right) \longrightarrow 0 \quad \text{dado que } nh^4 \longrightarrow 0$$

En lo que sigue se obtiene la distribución asintótica de

$$\sqrt{n^2h}\left[d^2\left(\hat{m}_n, \mathcal{A}^t(-)\hat{\theta}_n\right) - \frac{\overline{\omega}\sigma_\epsilon^2\int K^2}{nh}\right]$$

que por los cálculos realizados es la misma que la de $\sqrt{n^2h}\Delta_{13}$.

Obsérvese además que $\sqrt{n^2h}\Delta = O_{\mathbb{P}}\left(\sqrt{h}\right) = o_{\mathbb{P}}(1)$.

Teniendo en cuenta que

$$\sum_{j=1}^{n} W_{n\tau}(x_j)W_{nk}(x_j)\omega(x_j) \simeq \frac{1}{n^2h}\left((K*K)\left(\frac{x_\tau - x_k}{h}\right)\right)\omega(x_\tau)$$

sin más que aplicar aproximaciones de Riemann, se deduce que la distribución asintótica de Δ_{13} es la misma que la de:

$$\frac{1}{n^2h}\sum_{\tau\neq k}\left((K*K)\left(\frac{x_\tau - x_k}{h}\right)\right)\omega(x_\tau)\epsilon_\tau\epsilon_k =$$

$$= \frac{1}{n^2h}\sum_{\tau=1}^{n}\left\{\sum_{\substack{k\\ \tau\neq k}}\left((K*K)\left(\frac{x_\tau - x_k}{h}\right)\epsilon_k\right)\omega(x_\tau)\epsilon_\tau\right\} = \sum_{\tau=1}^{n}X_{\tau n}$$

Nuevamente, se obtiene la distribución asintótica de expresión utilizando el resultado de Nieuwenhuis (1992) (Teorema 2).

Para ello, denotemos $\tilde{\epsilon}_i = \sum_{j=0}^{\tilde{m}(n)} b_j e_{i-j}$, entonces $X_{\tau n}$ se puede descomponer en $X_{\tau n,m(n)} + \overline{X}_{\tau n,m(n)}$ siendo:

$$X_{\tau n,m(n)} = \frac{1}{n^2h}\sum_{\tau\neq k}\left((K*K)\left(\frac{x_\tau - x_k}{h}\right)\right)\omega(x_\tau)\tilde{\epsilon}_\tau\tilde{\epsilon}_k \qquad \text{(parte principal)}$$

$$\overline{X}_{\tau n,m(n)} = \frac{1}{n}\sum_{\tau\neq k}\left((K*K)\left(\frac{x_\tau - x_k}{h}\right)\right)\omega(x_\tau)\left(\epsilon_\tau\epsilon_k - \tilde{\epsilon}_\tau\tilde{\epsilon}_k\right) \quad \text{(parte residual)}$$

Eligiendo $\tilde{m}(n) = o(nh)$, se tiene que

$$\frac{[nh + \tilde{m}(n)]}{n} \longrightarrow 0 \iff n^{1+(2/j)}h^{2+(2/j)} \longrightarrow 0$$

condición que se cumple por la hipótesis A6.

La demostración de que las disposiciones triangulares así definidas verifican las condiciones C_1, C_1^* y C_2 del teorema 2, concluyen la demostración del teorema 1.b).

■

5. BIBLIOGRAFÍA

[1] **Altman, N.S.** (1990). "Kernel smoothing of data with correlated errors". *J.A.S.A.*, **85, 411**, 739–749.

[2] **Brockwell, P.J.** & **Davis, R.A.** (1991). *Time series theory and methods.* 2ª ed. Springer.

[3] **Chu, C.K.** & **Marron, S.** (1991). "Comparison of two bandwidth selectors with dependent errors". *The Annals of Stat.*, **19, 4**, 1906–1918.

[4] **Eubank, R.L.** & **Spiegelman, C.H.** (1990). "Testing the goodness of fit of a linear model via nonparametric regression techniques". *J.A.S.A.*, **85, 410**, 387–392.

[5] **Eubank, R.L.** & **Hart, J.D.** (1992). "Commonality of cusum, Von Neumann and smoothing based goodness of fit tests". *Biometrika*, **80, 1**, 89–98.

[6] **Eubank, R.L.** & **Hart, J.D.** (1993). "Testing goodness-of-fit in regression via order selection criterio". *Annals of Stat.*, **20, 3**, 1412–1425.

[7] **Firth, D., Glosup, J.** & **Hinkley, D.** (1991). "Model checking with nonparametric curves". *Biometrika*, **78, 2**, 245–252.

[8] **Gasser, T.** & **Muller, H.G.** (1979). "Kernel estimation of regression functions in Smoothing Techniques for Curve Estimation". *Lect. Notes in Math Springer*, 23–68.

[9] **González Manteiga, W.** & **Cai Abad, R.** (1991). "Testing hypothesis of general linear model using nonparametric regression estimation". (Aparecerá en *Test.*)

[10] **Hart, J.D.** & **Wehrly, T.E.** (1992). "Kernel regression when the boundery region is large, with an application to testing the adquacy of polynomial models". *J.A.S.A.*, **87, 420**, 1018–1024.

[11] **Kozek, A.S.** (1991). "A nonparametric test of fit of a parametric model". *Journal of Multiv. Analysis*, **37**, 66–75.

[12] **Muller, H.G.** & **Statmuller, U.** (1988). "Detecting dependencies in smooth regression models". *Biometrika*, **75, 4**, 639–650.

[13] **Nieuwenhuis, G.** (1992). "Central limit theorems for sequences with $m(n)-$dependent main part". *Journal of Stat. Planning and Inference*, **32**, 229–241.

[14] **Peña Sánchez de Rivera, D.** (1987). *Estadística, Métodos y Modelos.* Vol. 2, Alianza Universidad Textos.

[15] **Priestley, M.B.** & **Chao, M.T.** (1972). "Nonparametric function fitting". *J. Roy. Stat. Soc. B*, 385–392.

[16] **Raz, J.** (1990). "Testing for no effect when estimating a smooth function by nonparametric regressión: A randomization approach". *J.A.S.A.*, **85, 409**, 132–138.

[17] **Seber, G.** (1977). *Linear Regression Analysis.* John Wiley.
[18] **Seber, G.** & **Wild** (1989). *Nonlinear Regression.* John Wiley.
[19] **Staniswalis, J.** & **Severini, T.** (1991). "Diagnostics for essesing regression models". *J.A.S.A.*, **86, 415**, 684–692.

ENGLISH SUMMARY:

TESTING THE HYPOTHESIS OF A GENERAL LINEAR MODEL WHEN ERRORS ARE NON-INDEPENDENT

Vilar Fernández, J.M. and González Manteiga, W.

1. INTRODUCTION

One of the most extensively studied questions in the theory of parametric regression models:

$$Y_i = m_\theta(x_{1i}, \ldots, x_{pi}) + \epsilon_i, \quad i = 1, \ldots, n; \quad \theta \in \Theta \subset \mathbb{R}^q \tag{1.1}$$

with $x_i \in \mathcal{C}$, $i = 1, \ldots, n$, a compact set in $\mathbb{R}$, and $\{\epsilon_i\}_{i=1}^n$ a sequence of random variables with zero mean, is how to determine whether a model m_θ is better supported by the data than another.

In this paper we have designed a new method for testing the hypothesis that the regression function follows a general linear model,

$$\mathrm{H_o}: m \in \left\{ m_\theta(\bullet) = \mathcal{A}^t(\bullet)\theta : \theta \in \Theta \subset \mathbb{R}^q \right\},$$

with $\mathcal{A}$ a function from $\mathbb{R}$ to $\mathbb{R}^q$, against the alternative

$$\mathrm{H_1} = \text{"}m \text{ belongs to one class of smooth functions."} \tag{1.2}$$

The statistic used for testing the given hypothesis (d^2) is defined to be the difference between the average squared errors (ASE) associated with the non-parametric estimator $\hat{m}$ of m the minimum distance estimators $m_{\hat{\theta}}$ of m.

The asympotic normality of both d^2 and $\hat{\theta}_n$ are obtained under general conditions and the random errors $\{\epsilon_i\}$ have an MA(∞) structure. Simulations and examples with real data show how the non independence of errors can affect decision on H_o.

2. DEFINITIONS AND ASYMPTOTIC RESULTS

We consider the fixed design model $Y_i = m(x_i) + \epsilon_i$, $(i = 1, \ldots, n)$ where m is a smooth function defined in [0,1], the $\{x_i\}_{i=1}^n$ are equally spaced $(x_i = i/n)$, the Y_i are observations obtained at the x_i, and the errors ϵ_i are generated by an infinite order moving average: $\epsilon_i = \sum_{j=0}^{\infty} b_j e_{i-j}$, where $\{\epsilon_i\}$ is zero mean white noise of variance σ_e^2.

The functional distance we shall use to test $H_o: m \in \{m_\theta = \mathcal{A}^t(\cdot)\theta | \theta \in \Theta\}$ is a Crámer-von-Mises type distance given by:

$$
\begin{aligned}
d^2\left(\hat{m}_n, H_o\right) &= \min_{\theta \in \Theta} \int \left(\hat{m}_n(x) - \mathcal{A}^t(x)\theta\right)^2 \omega(x)\, d\Omega_n(x) = \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\hat{m}_n(x_i) - \mathcal{A}^t(x_j)\hat{\theta}_n\right)^2 \omega(x_i) \qquad (2.1)
\end{aligned}
$$

where ω is a weighing function and Ω_n empirical distribution of the points of the design. As non-parametric estimator $\hat{m}_n$ we use a Gasser-Müller estimator:

$$
\hat{m}_n(x) = \sum_{j=1}^{n} h^{-1} \left\{ \int_{s_{j-1}}^{s_j} K\left(\frac{x-s}{h}\right) ds \right\} Y_j = \sum_{j=1}^{n} W_{n,j}(x) Y_j
$$

The estimator $\hat{\theta}_n$ obtained by the minimization of the function: $\Psi(\theta) = \int \left(\hat{m}_n(x) - \mathcal{A}^t(x)\theta\right)^2 \omega(x)\, d\Omega_n(x)$ becomes the classical least squares estimator $\hat{\theta}_{ls}$ if a degenerate smoothing parameter (h_n) is used, being, in general, the probability of rejecting H_o smaller for $\hat{\theta}_n$ than for $\hat{\theta}_{ls}$, since as a rule

$$
d^2\left(\hat{m}_n, \mathcal{A}^t(-)\hat{\theta}_n\right) < d^2\left(\hat{m}_n, \mathcal{A}^t(-)\hat{\theta}_{ls}\right)
$$

Like $\hat{\theta}_{ls}$, $\hat{\theta}_n$ is a "**root**$-n$" estimator for θ if H_o holds, i.e.

$$
n^{1/2}\left(\hat{\theta}_n - \theta\right) = O_{\mathbb{P}}(1).
$$

This is a consequence (part a) of the following theorem, which we prove with the help of the Central Limit Theorems for sequences with m_n-dependent main part due to Nieuwenhuis (1992).

Theorem 1

Under assumption H_0 and subject to general premises

a) $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \sigma_\epsilon^2 \mathbf{Q}^{-1}\mathbf{Q}^*\mathbf{Q}^{-1}\right)$

where $\mathbf{Q} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathcal{A}(x_i)\mathcal{A}^t(x_i)\omega(x_i)$

and $\mathbf{Q}^* = \sum_{u=-\infty}^{\infty} \gamma(u)R(u)$,

with $R(u) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n-u} \mathcal{A}(x_j)\mathcal{A}^t(x_{j+u})\omega^2(x_j)\omega^2(x_{j+u})$

b) $\sqrt{n^2h}\left(d^2\left(\hat{m}, \mathcal{A}^t(\cdot)\hat{\theta}_n\right) - \frac{\overline{\omega}\sigma_\epsilon^2 \int K^2}{nh}\right) \xrightarrow{d} N\left(0, \sigma_d^2\right)$

where $\sigma_d^2 = 2\left(\sum_{k=-\infty}^{\infty} \gamma(k)\right)^2 \int (K * K)^2 \int \omega^2$

and γ is the autocovariance function of the errors.

Note that for independent errors (i.e. when $\gamma(k) = 0$ for all non-zero k, $\sigma_d^2 = 2\sigma_\epsilon^4 \int (K * K)^2 \int \omega^2$, which is González-Manteiga and Cao Abad's (1994) result.

3. SIMULATION STUDIES

The Theorem 1.a) shows that the variance of the least distance estimator $\hat{\theta}_n$ is asymptotically larger if the sample errors are positively correlated than if they are uncorrelated. It is also clear from Th.1.b) that ignoring positive correlation when it exists can lead to H_0 being unduly rejected. This behaviour is illustrated by the results of a first simulation study.

To study the behaviour of test d^2 as smoothing parameter h is varied, we performed a second simulation experiments, and we observed that the choice of bandwidth is crucial and the simulation study suggest that a good selection for h, under H_o consists in taking the first minimum of the optimization:

$$\min_h \left[d^2 \left(\hat{m}_{n,h}, \mathcal{A}^t(-)\hat{\theta}_{n,h} \right) \right]$$

suggestion made above works well under H_o. But this is not a good selection under other alternatives. More research is necessary in order to investigate the selection of h under other alternatives.

The influence of correlation among the errors is also illustrated by analysis of a real example studied in the section 4. And, finally, the Appendix is devoted to a sketch of the proof of theorem 1.

QÜESTIIÓ, Vol 18, 3 pp. 369-375, 1994

COMPARISON OF SIX ALMOST UNBIASED RATIO ESTIMATORS

M. DALABEHERA* AND L.N. SAHOO†

In this paper, we compare six almost unbiased ratio estimators with respect to bias and efficiency for (i) finite populations, and (ii) infinite populations in which the joint distribution of the characters under study is bivariate normal.

Key words: Almost unbiased estimator, bias, efficiency, mean square error, ratio estimator.

AMS classification: 62 DO5

1. INTRODUCTION

Let y and x be real variates taking values y_i and x_i $(1 \leq i \leq N)$ for the ith unit of a population of size N with means μ_y and μ_x respectively. Suppose that a sample of n units is selected from the population by simple random sampling without replacement (SRSWOR) for the purpose of estimating the ratio $R = \mu_y/\mu_x \quad (\mu_x \neq 0)$. Under the assumption that the two variates are positively correlated the usual ratio estimator $r = \overline{y}/\overline{x}$ $(\overline{y}, \overline{x}$ denoting the sample means of y and $x)$ for R is well known in the literature. In general, the estimator is

* M. Dalabehera. Department of Statistics. Orissa University of Agriculture and Technology, Bhubaneswar. India.

† L.N. Sahoo. Department of Statistics. Utkal University, Bhubaneswar. India.

–Article rebut l'octubre de 1993.

–Acceptat l'abril de 1994.

biased and though for large n the bias is negligible as compared to the standard deviation, several "ratio-type" estimators have been put forward that are unbiased or almost unbiased (unbiased upto terms of order n^{-1}).

Beale (1962) and Tin (1965) proposed equivalent almost unbiased ratio (AUR) estimators

$$\lambda_1 = r\left(1+\theta_1 c_{xy}\right)/\left(1+\theta_1 c_x^2\right)$$

and

$$\lambda_2 = r\left[1+\theta_1\left(c_{xy}-c_x^2\right)\right]$$

respectively, where

$$\begin{aligned}
\theta_1 &= n^{-1}-N^{-1}, \\
c_{xy} &= s_{xy}/\overline{x}\overline{y}, \\
c_x^2 &= s_x^2/\overline{x}^2, \\
s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right),
\end{aligned}$$

and

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2.$$

Another AUR estimator, similar to those of Beale and Tin suggested by Sahoo (1983) is

$$\lambda_3 = r/\left[1+\theta_1\left(c_x^2-c_{xy}\right)\right].$$

Sahoo (1987) generated a class of AUr estimators considering a function of c_x^2 and c_{xy} satisfying some regularity conditions, which includes λ_1, λ_2 and λ_3 as particular cases. Three new AUR estimators as members of the class were also proposed by Sahoo (1987). These estimators are defined by

$$\begin{aligned}
\lambda_4 &= r\left(1+\theta_1 c_{xy}\right)\left(1-\theta_1 c_x^2\right) \\
\lambda_5 &= r\left(1-\theta_1 c_x^2\right)/\left(1-\theta_1 c_{xy}\right)
\end{aligned}$$

and

$$\lambda_6 = r/\left[\left(1-\theta_1 c_{xy}\right)\left(1+\theta_1 c_x^2\right)\right].$$

But, no attempts have been made to study the design based properties of these estimators.

In this paper, a comparison of these six AUR estimators is perfomed on the basis of biasedness and efficiency for

(i) finite populations considering bias and mean square error to terms of $O(n^{-2})$, and

(ii) infinite populations assuming that the distribution of the two-dimensional variable (x, y) is bivariate normal and considering mean square error upto $O(n^{-3})$.

The comparative study made by Tin (1965) leads to the conclusion that for (i) λ_1 has a smaller bias than λ_2, whereas λ_2 is more efficient than λ_1 for (ii). Sahoo (1983) also judged the efficiency of λ_3 for (ii) and shown that it is more efficient than λ_1 but more efficient than λ_2 when $C_{11} > C_{20}$, where we define $C_{pq} = K_{pq}/\mu_x^p \mu_y^q$, K_{pq} being the (p, q)th cumulant of x and y.

In the present study, the biases and mean square errors of the estimators are obtained by using the same notations and approximations used by Tin (1965) [also see Kendall, Stuart and Ord, vol. III, p. 242].

2. COMPARISON OF BIASES

To terms of order n^{-2}, the expected values of the AUR estimators are given as follows:

$$E(\lambda_1) = R\left[1 - H - 2\theta_1^2 C_{20}(C_{20} - C_{11})\right]$$

$$E(\lambda_2) = R\left[1 - H - 3\theta_1^2 C_{20}(C_{20} - C_{11})\right]$$

$$E(\lambda_3) = R\left[1 - H - \theta_1^2 (2C_{20} + C_{11})(C_{20} - C_{11})\right]$$

$$E(\lambda_4) = R\left[1 - H - \theta_1^2 C_{20}(3C_{20} - 2C_{11})\right]$$

$$E(\lambda_5) = R\left[1 - H - \theta_1^2 (3C_{20} + C_{11})(C_{20} - C_{11})\right]$$

$$E(\lambda_6) = R\left[1 - H - \theta_1^2 (2C_{20} - C_{11})^2\right]$$

where $H = (2\theta_2 - 3\theta_1/N)(C_{21} - C_{30})$ and $\theta_2 = n^{-2} - N^{-2}$.

From the above results no general conclusions can be expected. However, taking into account the signs of the biases the following tentative conclusions may be drawn:

(a) When $C_{21} > C_{30}$,

$$|B(\lambda_1)| < |B(\lambda_3)| < |B(\lambda_2)| < |B(\lambda_5)| < |B(\lambda_4)| < |B(\lambda_6)|$$

if $C_{11} < C_{20}$, showing that Beale's estimator is the least biased.

(b) When $C_{21} < C_{30}$,

$$|B(\lambda_1)| < |B(\lambda_2)| < |B(\lambda_3)| < |B(\lambda_5)| \quad \text{if } C_{11} > C_{20}$$

and

$$|B(\lambda_4)| < |B(\lambda_1)| < |B(\lambda_2)| < |B(\lambda_3)| < |B(\lambda_5)| \quad \text{if } C_{11} > 1.5C_{20}$$

(c) If the distribution of (x, y) is bivariate normal, $C_{21} = C_{30} = 0$ and we have

$$\begin{aligned}
&|B(\lambda_1)| < |B(\lambda_2)| < |B(\lambda_5)|, \\
&|B(\lambda_1)| < |B(\lambda_3)| < |B(\lambda_5)|, \\
&|B(\lambda_2)| \gtrless |B(\lambda_3)| \quad \text{according as } C_{11} \gtrless C_{20}, \\
&|B(\lambda_5)| < |B(\lambda_4)| < |B(\lambda_6)| \quad \text{for } C_{11} < C_{20}
\end{aligned}$$

and

$$|B(\lambda_4)| < |B(\lambda_5)| \quad \text{for } C_{11} > 1.5C_{20}.$$

(d) If the regression line of y on x is linear passing through the origin, $C_{11} = C_{20}, C_{21} = C_{30}$ and we have

$$B(\lambda_i) = 0, \quad i = 1, 2, 3 \text{ and } 5$$

and

$$B(\lambda_4) = B(\lambda_6) \neq 0.$$

3. COMPARISON OF MEAN SQUARE ERRORS

To terms of order n^{-2}, the mean square errors of the six AUR estimators are identical and given by the following expressions:

$$\begin{aligned}
M(\lambda_i) &= R^2 \left[\theta_1 (C_{20} - 2C_{11} + C_{02}) + \right. \\
&+ \theta_1^2 \left(2C_{20}^2 - 4C_{20}C_{11} + C_{11}^2 + C_{20}C_{02}\right) + \\
&+ \left. (2\theta_1/N)(C_{30} - 2C_{21} + C_{12})\right], \quad i = 1, 2, \ldots, 6. \qquad (3.1)
\end{aligned}$$

Using the characteristics C_{pq}, Tin (1965) also expressed the mean square error of r to $O(N^{-2})$ as

$$\begin{aligned} M(r) &= R^2 \left[\theta_1 \left(C_{20} - 2C_{11} + C_{02}\right) + \right. \\ &+ \theta_1^2 \left(9C_{20}^2 - 18C_{20}C_{11} + 6C_{11}^2 + 3C_{20}C_{02}\right) - \\ &- \left. 2\left(\theta_2 - 3\theta_1/N\right)\left(C_{30} - 2C_{21} + C_{12}\right)\right] \end{aligned} \tag{3.2}$$

Thus, from (3.1) and (3.2) we see that the inequality $M(r) \gtrless M(\lambda_i)$ holds whenever we have

$$7C_{20}^2 - 14C_{20}C_{11} + 5C_{11}^2 + 2C_{20}C_{02} \gtrless 2\left(C_{30} - 2C_{21} + C_{12}\right)$$

or after adjusting

$$3.5\left(C_{20} - C_{11}\right)^2 + \left(1 - \lambda^2\right) C_{20}C_{02} \gtrless \delta, \tag{3.3}$$

where λ is correlation coefficient between x and y; $\delta = C_{30} - 2C_{21} + C_{12}$.

In practice, one can not check any of the conditions given in (3.3) to be able to make an appropriate decision, since the value of δ is not known or rarely known. Tin (1965) while comparing the precisions of λ_1 and λ_2 with r, did not point out anything on the contribution of δ to the mean square errors. However, when the distribution of (x, y) is bivariate normal, $\delta = 0$, and six AUR estimators have smaller mean squares than the usual ratio estimator r.

A choice among the six AUR estimators naturally depends on the comparison of mean square errors to $O(n^{-3})$. Unfortunately, the results are too complicated to lead to a useful guide for practical purposes. But, assuming the distribution of (x, y) to be bivariate normal, the results simplify considerably and we get,

$$M(\lambda_1) = \frac{R^2}{n}\left[V_1 + \frac{C_{20}}{n}\left(1 + \frac{1}{n}\right)\left(V_1 + A^2C_{20}\right) - \frac{2C_{20}^2}{n^2}\left(V_1 + 6A^2C_{20}\right)\right]$$

$$M(\lambda_2) = \frac{R^2}{n}\left[V_1 + \frac{C_{20}}{n}\left(1 + \frac{1}{n}\right)\left(V_1 + A^2C_{20}\right) - \frac{4C_{20}^2}{n^2}\left(V_1 + 6A^2C_{20}\right)\right]$$

$$\begin{aligned} M(\lambda_3) = \frac{R^2}{n}\left[V_1 + \frac{C_{20}}{n}\left(1 + \frac{1}{n}\right)\left(V_1 + A^2C_{20}\right) - \frac{2C_{20}^2}{n^2}\left(V_1 + 6A^2C_{20}\right) - \right. \\ \left. - \frac{2C_{11}}{n^2}\left(V_1C_{11} + 4A^2C_{20}^2\right)\right] \end{aligned}$$

$$M(\lambda_4) = \frac{R^2}{n}\left[V_1 + \frac{C_{20}}{n}\left(1+\frac{1}{n}\right)(V_1 + A^2C_{20}) - \frac{4C_{20}^2}{n^2}(V_1 + 5A^2C_{20} + \right.$$
$$\left. + \ 2AC_{11})\right]$$

$$M(\lambda_5) = \frac{R^2}{n}\left[V_1 + \frac{C_{20}}{n}\left(1+\frac{1}{n}\right)(V_1 + A^2C_{20}) - \frac{2C_{20}^2}{n^2}(3V_1 - 2AV_1 + \right.$$
$$\left. + \ A^2V_1 + 14A^2C_{20} - 4A^3C_{20})\right]$$

$$M(\lambda_6) = \frac{R^2}{n}\left[V_1 + \frac{C_{20}}{n}\left(1+\frac{1}{n}\right)(V_1 + A^2C_{20}) - \frac{2C_{20}^2}{n^2}(2V_1 - 2AV_1 + \right.$$
$$\left. + \ A^2V_1 - 4AC_{20} + 14A^2C_{20} - 4A^3C_{20})\right]$$

where $A = (1 - C_{11}/C_{20})$, $\quad V_1 = A^2C_{20} + V_2$, $\quad V_2 = C_{02}(1 - \lambda^2)$.

On comparison of the mean squares to $O(n^{-3})$ the following conclusions may be drawn:

(a) If $C_{11} > C_{20}$,

$$M(\lambda_5) < M(\lambda_2) < M(\lambda_1)$$
$$M(\lambda_6) < M(\lambda_3) < M(\lambda_2)$$

and

$$M(\lambda_6) < M(\lambda_5) < M(\lambda_4).$$

(b) If $C_{11} < C_{20}$,

$$M(\lambda_4) < M(\lambda_2) < M(\lambda_1)$$
$$M(\lambda_4) < M(\lambda_3) < M(\lambda_1)$$

and

$$M(\lambda_5) < M(\lambda_3).$$

(c) If $C_{11} < 2C_{20}$,

$$M(\lambda_6) < M(\lambda_1).$$

(d) If $C_{20} < C_{11} < 2C_{20}$, then λ_6 is uniformly better than other AUR estimators under comparison.

(e) If there is a linear regression of y on x passing through the origin,

$$M(\lambda_5) < M(\lambda_2) = M(\lambda_3) = M(\lambda_4) = M(\lambda_6) < M(\lambda_1).$$

4. CONCLUDING REMARKS

We have seen that, when the regression line of y on x is linear passing through the origin ($C_{11} = C_{20}$), which is of course an optimality condition for ratio estimate to be fruitfully employed in practice, the performance of λ_5 seems to be better than its competitors. But for other situations ($C_{11} \neq C_{20}$) no general conclusions can be obtained and performance of one estimator over other depends mainly on the population parameters. Because, the most part of the resulting expressions is only in an asymptotical form. The asymptotic theory, however, can not be expected to help when very small samples are taken without replacement especially from populations of small and moderate size. Thus, the above comparison clearly indicates the need for an extensive study mathematical as well as empirical, using various models and wide varieties of actual data.

REFERENCES

[1] **Belae, E.M.L.** (1962). "Some uses of computers in operational research". *Industrielle Organisation,* **31**, 51–52.

[2] **Kendall, M., Stuart, A.** and **Ord, J.K.** (1983). *The Adavanced Theory of Statistics.* (vol. III). Charles Griffin and Company Limited. London and High Wycombe.

[3] **Sahoo, L.N.** (1983). "On a method of bias reduction in ratio estimation". *Jour. Stat. Res.,* **17**, 1–6.

[4] **Sahoo, L.N.** (1987). "On a class of almost unbiased estimators for population ratio". *Statistics,* **18**, 119–121.

[5] **Tin, M.** (1965). "Comparison of some ratio estimators". *Jour. Amer. Stat. Assoc.,* **60**, 294–307.

QÜESTIIÓ, Vol 18, 3 pp. 377-383, 1994

FORMATION OF AN INDEX OF ECONOMIC CAPACITY IN BARCELONA✠

TOMAS ALUJA-BANET*

Universitat Politècnica de Catalunya

This study presents an evaluation of the "economic capacity" for the year 1988 of each of the 1919 censual sections that Barcelona is divided into. In view of the confidential nature of all information concerning incomes and resources, we have employed an indirect method of estimation using all the territorial information available al the highest level o disaggregation, that is the censual sections. This evaluation gathers variable indicators of income distribution and wealth, and without being exactly one or the other, incorporates elements of both aspects. Taking these variables into account, through Generalised Principal Components Analysis we obtain the economic capacity index of the average household in the censual section. Then it just remains to scale this factor according to the variability of family income by districts in Barcelona, giving a value of 100 for the whole city.

Key words: Economic capacity, quantification, principal components, correspondence analysis.

✠This study was promoted by the Barcelona City Council and the "Caixa d'Estalvis i Pensions de Barcelona", with the collaboration of "Telefónica de España, S.A.".

* Tomas Aluja-Banet. Dept. Statistics and Operations Research. Universitat Politècnica de Catalunya. C/Pau Gargallo, 5. 08028 Barcelona. E-mail: aluja@eio.upc.es.

–Article rebut el novembre de 1992.

–Acceptat l'abril de 1994.

1. INTRODUCTION

The purpose of this study is to build an index of economic capacity at the most disaggregated level in Barcelona, which is in our case the censual section (1919 censual sections of approximately 300 households on average). This index can be used for a wide range of purposes, economic studies with territorial reference of any kind, electoral sociology, planning in the public administration, and business and services such as marketing and banking.

Due to the fact that all information concerning incomes and economic resources is confidential, statistical, techniques must be used to create this index. There are different ways of measuring the economic capacity per household, each having its advantages and disavantages. First of all, there is the possibility of making a specific survey; this would have to be very carefully carried out due to the sensitive nature of the subject, which makes it particularly difficult to obtain reliable information about it. The advantage of this method is that it estimates the economic capacity directly from every statistical unit (household) selected. Another possibility consists in estimating economic capacity indirectly from expenditures. In both cases it is necessary to carry out a survey which would inevitably be time-consuming and costly. However, economic capacity would be measured at an individual level (in our case, the statistical individual is the family), and would therefore allow us to characterise the families according to their socio-demographic characteristics, and to obtain average figures for the different territorial units defined in Barcelona (the censual sections, quarters and districts mentioned above), if the selected sample will allow it.

Here, we have employed an indirect methodology for estimating economic capacity, which obviates the need to carry out a survey and consists in using all the territorial information available at the highest level of disaggregation (the censual sections). In fact, this information has already been compiled, especially at a municipal level (to be exact, from the register of inhabitants, the register of vehicles and the land registry), or from services, such as the telephone company. This information has been traditionally taken as an indicator of income or wealth, depending of the cases, by censual section. Therefore, the statistical methodology consisted in extracting the common factor of all the available indicators as the closest approximation to financial capacity. Note that the issued index is not a traditional index of income or wealth, since it incorporates elements of both; nor is it strictly speaking a family index. Indeed it is an average family index by censual section. The main advantage of such a methodology is that it can provide a quick estimate of the economic capacity per censual section at very low cost.

2. AVAILABLE SOURCES OF INFORMATION

Barcelona City Council has provided the following information from its municipal registers of inhabitants (1988), vehicles (1987) and land (1988), aggregated by censual section. From "Telefónica de España, S.A." we received the average annual telephone consumption of its private clients for 1989, also aggregated by censual section:

It is first worth pointing out the (relative) heterogeneity of the indicators provided: they do not all express quite the same. In particular, we can say that the socio-professional category represents a "potential" income in a censual section, or what we might call the "intellectual estate" of the censual section. The size of vehicles and motorcycles, and specially the rateable value of construction and services, are indicative of estate. Land value is more difficult to characterise as it is linked to speculation, but it is an indirect indicator of income spent as it reflects the level of prices in a given area. The size of premises can also be considered an indicator or estate, while the age of vehicles or premises is difficult to characterise, though they can be taken more as indicators of income spent, a lower age indicating higher purchasing power. Telephone consumption is certainly a traditional indicator of income spent, while the connection charge gives some idea of the telephone equipment installed by clients and the number of lines can be indicative of development in a given area.

All the available information has a clear definition in expressing an average value per censual section, taking the most suitable denominator in each case. On the contrary, the quantification of the socio-professional category obtained from the register of inhabitants is not evident, although this information is decisive for any evaluation of income distribution, because the corresponding value in a given censual section is not obvious. Traditionally, this information is presented in the form of a compound index of occupational variables, level of education, type of accommodation, etc. In the present case, we have built a specific socio-professional index of the city of Barcelona, only taking into consideration the occupation of working-age population. Indeed, we known from previous studies that the non-active population expresses an orthogonal dimension to the active one, linked to the age of censual sections; at the same time, the educational level does not provide any complementary information to that of the occupations, relative to censual sections. Therefore, we used the distribution of the working-age population to get a classification of the active population according to their declared or last declared occupation, (in cases where this information is missing, the educational level is used instead), forming a frequency table of 1919 rows (one per each censual section) and 24 columns (the categories of the socio-professional variable). The problem lies in passing from this variable (de-

fined as a multidimensional one) to a synthetic numerical index. We do so by means of the Correspondence Analysis of the above table, obtaining a first factor that summarises 51.63% of the information in the original table. Thus we will use this factor as a numerical quantification of the socio-professional value of each censual section. This factor orders the censual sections according to their professional profile. It reflects the fact that residence is chosen according to the social prestige of occupations. We can therefore name the first axis obtained as a social status axis, which is not directly related to income or wealth (a construction worker can have a very high income) but is related with lifestyle and the possibilities life has to offer. It thus defines potential economic capacity by censual sections.

Finally, after detecting the colinealities among the indicators provided, we used the following ones, which represent different pieces of information for the index of economic capacity:

- Social status (first factor of the CA of the socio-professional table)
- Average power of motors cars per household
- Average age of motor cars
- Average rateable value of construction and services of premises
- Average rateable of sites of premises
- Average monthly spending on telephone per household

3. FORMATION OF THE INDEX

Moving on, so as to extract the common factor most correlated to all the original variables (indicators), we carried out a Normalised Principal Component Analysis weighing each section according to the number of registered households. The matrix of correlations of indicators is as follows:

Social Status	1.00					
Age of motor cars	0.57	1.00				
Rateable of sites	0.82	0.41	1.00			
Power of motor cars	0.67	0.45	0.71	1.00		
Rateable construction services	0.71	0.65	0.74	0.71	1.00	
Telephone consumption	0.74	0.49	0.75	0.71	0.76	1.00

Note that the maximum correlation is that of social status with site value, while the lowest correlations are in the age of vehicles. The singular value decomposition of this matrix gives a first factor which summarises 72.06% of the information of all the original variables together. The expression of the first factor found on the basis of standardised original variables is as follows:

$$\begin{aligned} e.c.i. \;=\; & 0.89\ Social\ Status + 0.84\ Power\ of\ motor\ cars + \\ & + 0.68\ Age\ of\ motor\ cars + 0.90\ Rateable\ construction\ servies + \\ & + 0.88\ Rateable\ of\ sites + 0.88\ Telephone\ consumption \end{aligned}$$

(where e.c.i. stands for economic capacity index).

The correlations between variables and the first factor are given directly by the coefficients of the model. Note that high value of correlations (from 0.84 to 0.90, except for the age of motor cars, which is slightly lower), indicating that this first factor represents in fact a new (unobserved) variable, a compromise between all the variables (indicators) observed. The similarity of the coefficients suggest that the index formed is an approximate average of the selected set of standardised indicators. Finally, all that remains is to express this first factor as an index on the basis 100 for the whole Barcelona. The first factor as the best linear summary of the indicators observed, gives the ordering of the censual sections according to the level of their economic capacity, the mean of this factor corresponding to the average economic capacity for the whole of Barcelona. It does not, however, give the relative difference between censual sections. To do this we need to estimate the variance in the e.c.i. by censual section; we do so using the obtained variance of family income by districts in Barcelona from previous studies of the subject. Finally, expressing the index according to the dispersion found, we obtain an index centered on 100, with a deviation of 34.03; its range of variation goes from 32 to 292, in other words, it goes from approximately a third to approximately three times as much, giving a relative difference between the extremes of nine times as much. The resulting distribution curve is asymmetrical, as we can see in Fig. 1, the histogram of the index produced.

This characteristic, common to all functions of income distribution, becomes clearly evident in the concentration curve (or Lorentz curve, Fig. 2), which is one of the most widely used instruments in the study of inequality.

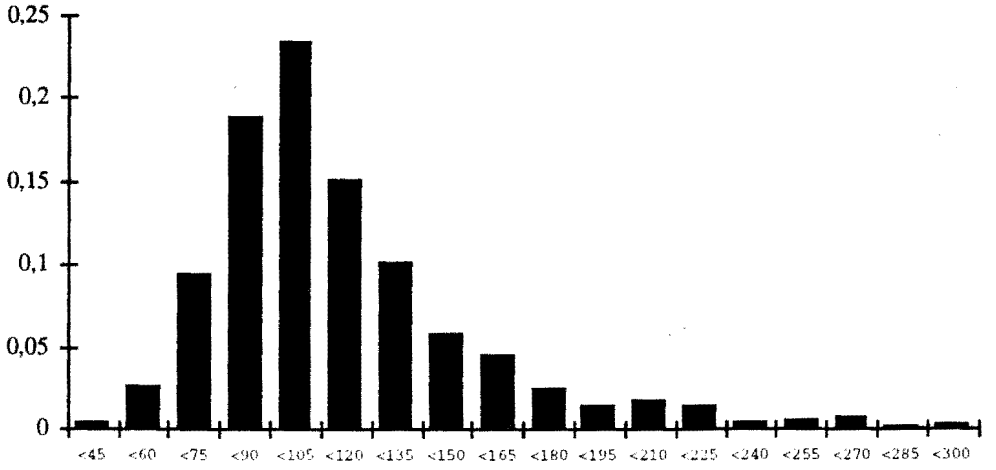


Figure 1.
Histogram of economic capacity index.

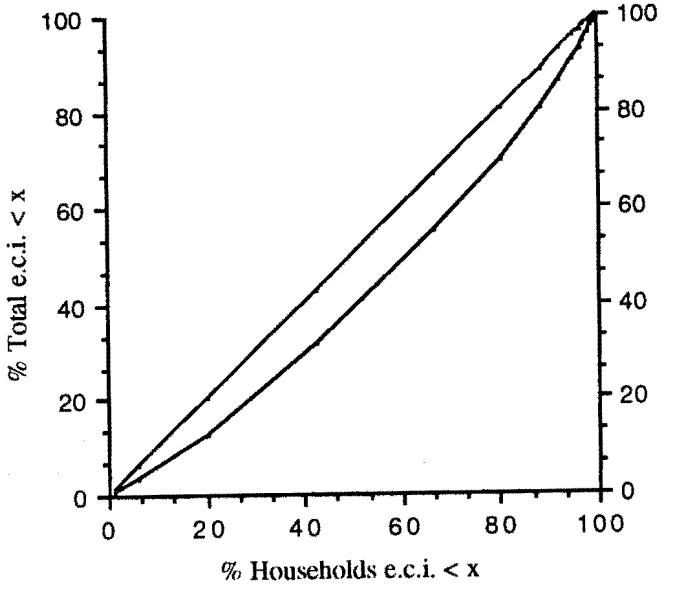


Figure 2.
Lorentz curve of the economic capacity index.

This index can be obtained from the Service of Statistics of the Barcelona City Council; here we give the value of the index for the 10 districts of Barcelona, which allows their comparison with other related studies.

	DISTRICT	E.C.I.	Number of families	% E.C.I. cumulative	% of families cumulative
1	Ciutat Vella	62.70	39067	6.20	6.94
8	Nou Barris	77.35	60264	13.86	17.65
3	Sants-Montjuïc	86.14	62042	22.38	28.68
10	Sant Martí	86.94	69597	30.99	41.05
9	Sant Andreu	88.84	46726	39.78	49.36
7	Horta-Guinardó	91.71	59792	48.86	59.98
6	Gràcia	100.98	47058	58.85	68.35
2	Eixample	114.36	102338	70.17	86.54
4	Les Corts	138.80	27978	83.90	91.51
5	Sarrià-Sant Gervasi	162.67	47762	100.00	100.00
	BARCELONA	100.00	562624		

REFERENCIAS

[1] **Aluja, T.** & **Ventura, A.** (1991). *Index de capacitat econòmica familiar a la ciutat de Barcelona.* Ajuntament de Barcelona i Caixa d'Estalvis i Pensions de Barcelona.

[2] **Oliver, J.; Busom, I.** & **Trullén, J.** (1989). *Estimació de la renda familiar disponible per càpita de Barcelona, els seus districtes i els 27 municipis de la Corporació Metropolitana de Barcelona (1985).* Ajuntament de Barcelona.

QÜESTIIÓ, Vol 18, 3 pp. 385-395, 1994

A MODEL FOR CREDIT SCORING: AN APPLICATION OF DISCRIMINANT ANALYSIS

MANUEL ARTÍS, MONTSERRAT GUILLÉN and JOSÉ Mª MARTÍNEZ*

Universitat de Barcelona

The application of statistical techniques in decision making, and more specifically for classification requirements, has proved to be adequate in the context of financial problems. In this study, we present the methodology used and the results obtained in the elaboration of a decision-support system for credit assignment. The problem was to provide an automatic tool for a Spanish financial institution that needed to quantify and analyse credit applications from clients. Firstly, we shall present the statistical techniques. Secondly, we shall describe the characteristics of the data set used for estimating the discrimination function and, finally, we show the results obtained when the model is used to discriminate among clients in the data set, whose history of financial behaviour is reflected in the data by means of a variable counting the number of unpaid instalments. Essentially, every individual asking the bank for a loan is assigned a certain score. This score is directly related to the probability he or she has of returning the money, but also to the risk of the institution lending that amount. Some comments about the application of the model are given and results concerning the optimal level of risk are also discussed, in order to give clear patterns for implementation.

Key words: Multivariate Analysis, Discriminant functions, Classification, Credit Scoring.

---

* Manuel Artís, Montserrat Guillén and José Mª Martínez. Departament d'Econometria, Estadística i Economia Espanyola. Universitat de Barcelona. Avgda. Diagonal, 690. 08034 Barcelona. Spain.

We acknowledge the support received from DGICYT, grant PB92-0545.

–Article rebut el novembre de 1993.

1. INTRODUCTION

The application of statistical techniques for the analysis of decision problems entailing an element of classification has experimented a great development in the last decade. Specially, those techniques have proven to be very useful in the context of business and finance (see Altman, Eisenbeis and Sinkey, 1981).

This work is a real application of the elaboration of a model for decision making, based on discriminant analysis, that has been developed to provide automatic decisions for credit assigment to private clients in a finance institution in the Spanish territory. More explicitely, we present a methodological approach that leads to the elaboration of a model. Afterwards, we present the characteristics of the data base used for the estimation and, finally, we analyze the results obtained for the simulation of the model performance as a tool for the evaluation of the risk adquired by the institution in the acceptation of loans.

The main interest of this paper is to present a practical application of the discriminant analysis techniques and, also, in the original use of the data in the context of a Spanish bank.

2. METHODOLOGY

Discriminant analysis concerns two different types of techniques, Factorial Discriminant Analysis and Discriminants Functions. The first one is considered in the analysis of differences among groups of individuals, it reveals the characteristics that make their relevance clear and also shows their relative importance. Discriminant functions have the aim of finding an optimal way of classifying units in groups. This latter problem has been approached in many different ways.

One of the applications of discriminant analysis is we studied in this paper, and it deals with the classification of individuals applying for a credit into two groups depending on whether they will pay the credit back on not. This procedure is based on assigning a probability to every individual, it is the probability of paying the money back or otherway, the probability of not paying it back. The probability may easily be transformed into a score that will be the essential information to make the final decision about the acceptation of the application for credit. Somehow, the probability of not returning the money is a measure of the risk of the financial institution when the client will be given credit, and

it may serve as a necessary information for the global estimation of risk of the financial institution in a given period of time.

In such a situation, a data base is normally available, containing information about clients that have applied for credit, and whose behaviour is known throughtout the period of credit payment. Moreover, usually, a batch of characteristics of the individuals, private characteristics, socio-economic status and financial behaviour in the past are available so that a typical portrait can be established for the analysis of future applications. Let us remark that the sample is truncated. This means that it has been chosen in the subpopulation of customers that already have credit, and it could be non-representative of the larger population of people that have applied for it. Nevertheless that was the only available information in the data bases and that limitation was not relevant for the practical aims of our study.

Let us call $\mathbf{x}$ the vector of the values for the explanatory variables measured in the individual whose application is to be evaluated. Let us assume that $\mathbf{x_M}$ is a vector of averages for the explanatory variables for the group of sampled individuals that have effectively payed the credit back, and $\mathbf{x_{NM}}$ the corresponding vector for the group that have not payed the money. Let us denote by $\mathbf{S_M}$ the covariance matrix for the first group and $\mathbf{S_{NM}}$ for the second. Afterwards, we may establish a measure that calculates the distance from individual to group that will be different for every group. The idea is that an individual will be classified into the group that is nearest to him, using the following distance:

$$D^2_M(x) = g_1(\mathbf{x}, M) + g_2(M)$$

for the individuals paying back,

$$D^2_{NM}(x) = g_1(\mathbf{x}, NM) + g_2(NM)$$

for the individuals not paying back.

The form of the functions is as follows:

$$g_1(\mathbf{x}, i) = (\mathbf{x} - \mathbf{x_i})^T \mathbf{S_i}^{-1} (\mathbf{x} - \mathbf{x_i}) + \log |\mathbf{S_i}| \quad i = M, NM$$

and,

$$g_2(i) = -2 \log(q_i) \quad i = M, NM$$

where q_i is the probability *a priori* for every group.

Sometimes, it is possible to simplify the above mentioned formula, due to the fact that the *a priori* probabilities may be consedered equal. Another source of simplification is the use of the global variance and covariance matrix, calculated

for all the individuals in the sample, that will be used instead of the two matrices above.

This latter aspect is a common practice when there are small differences in the dispersion and correlation between the variables in the studied groups. The specialised literature states that, if a test is performed, it may suggest to use the global matrix, but it is necessary to use the hypothesis of normality to ensure the adequate statistical properties of the estimators. Finally, the probability of belonging to each group is given by the Bayes' rule and it will provide the scoring:

$$p(\mathbf{x}) = \frac{\exp(-\frac{1}{2}D^2_{NM}(\mathbf{x}))}{\exp(-\frac{1}{2}D^2_{NM}(\mathbf{x})) + \exp(-\frac{1}{2}D^2_{M}(\mathbf{x}))}$$

From the early works of Fisher (1936), a great deal of studies have been published about Discriminant Analysis Applications, the classification problems have been viewed from very different points of view, including non parametric techniques.

3. THE DATA

The elaboration of a data base has the objective of obtaining reliable and detailed informationo about a number of clients that have been given credit, so that they would be represented randomly in Spain. Initially, a sample size of 5000 individuals was suggested, the information of their characteristics when the application was done had to be complemented with the incidences at payment time.

With a considerable effort, due to the difficulties in the data collection process, it was possible to obtain 4961 individuals, and the variables of interest were the year of birth, gender, marital status, average anual income, account level, studies, number of children, property of the household, job, credit destination, ... We quantified the qualitative variables although some aspects might depend on that quantification.

Afterwards, a simple statistical analysis was performed for a first analysis and detection of errors was carried out. Next, we proceeded to the model selection.

4. FINAL MODEL FOR CREDIT SCORING AND EVALUATION OF ITS DISCRIMINANT PERFORMANCE

In the first step, all the variables were introduced in the model so that a backstep algorithm might be used to find out those with greater capacity for discrimination. This search was done step by step. The objective was to obtain the best model, that is to say the model which provided the best proportions of correct classification when applied to the individuals in the sample used for estimation. We will consider a good model for credit scoring, a model that, when evaluating the applications in the sample leads to a greater percentage of correct classification of the individuals according to their posterior behaviour concerning the return of the amount. That means, finding out whether the applications belong to the group of paying or not paying back. This statistical procedure will lead to a model that is not necessarily the optimal, but it is the best when compared to the different trials. The search for a good model depends on the number of variables that the user allows into the model, or it depends on the restrictions about the presence or absence of certain variables. The first difficulty appears when deciding the maximum number of variables to be present in the model, e.g. the degrees of freedom.

Finally, after the different approximations and models, the percentages of correct classification for both the entire sample and the two populations were calculated. Simultaneously, some sensibility studies were performed in order to decide the inclusion or exclusion of some variables, in fact, we were trying to see the changes produced by the elimination of some variables that are difficult to measure in front of the inclusion of much more accessible variables.

The variables included in the final model are divided into different groups according to the source of the information that they provided, whether they are items responded by the individual applying for credit, or if those items are information that although it may be provided by the applicant, the financial institution is able to check in its archives.

The variables in the model may be found in one of the following three groups:

- Personal variables —three— (date of birth, marital status, number of children,...)
- Socio-economic variables —four— (net monthly income, ownership of house,...)
- Financial variables —five— (Monthly mortgage, availability of credit card,...)

The main innovation about the variables that are used in the model for credit scoring is that the above variables provide the information that is needed to create new variables finnally used in the model. The modifications are made in two different senses and have two well established objectives:

a) On the one hand, the financial institution is generally interested in preserving the confidentiality of the discriminant function that is finally being implemented. Since the scoring for a particular applicant must be calculated using special purpose software, it might well be the case that someone could have access to the coefficients for the discriminant function. We have seen that many credit managers sometimes distort the information in order to obtain a certain value of the scoring because they know the influence of a variable in the final result. The construction of new variables tries to avoid the above mentioned practice, or at least make it more difficult.

b) On the other hand, by using some new variables, the model can cope with the interactions which are produced among the different variables. Therefore, some new variables, interactions, have been introduced to detect very specific groups of defaulters. The interaction variables lead to a much better performance of the final model.

Once the variables that are to be present in the discriminant function are decided, we may proceed to the evaluation of the discriminating power of the model, or the capacity and operativity of the automatic decision tool. So, the discriminant function is used for the classification of the sampled applications, which are known to have complete information and known posterior behaviour, since the bank has in the files the number of payments that they have not been paid. Table I. shows five different columns, every row corresponds to a different prior probability for the two populations, this is equivalent to say that each row corresponds to a different threshold value for the posterior probability . The threshold establishes the minimum value of the scoring for the individual to be classified into the group of non defaulters, and for the credit to be granted.

Starting with a value of 50, and until 80 points (note that the posterior probability has been multiplied by 100), this table shows the different sceneries. Column one contains the different thresholds. The values in columns two and three correspond to the credits granted, and are the number of good credits granted and the number of bad credits granted, respectively. The next two columns correspond to the denied credits, first the number of good credits denied and finally bad credits denied.

Some graphics are presented below, showing a picture of the information in Table 1. Figure 1. shows that when the threshold increases, that is the minimum

score for granting, then the percentage of applications granted decreases, as well as the number of bad applications accepted.

Table I

THRESHOLD	GRANTED		DENIED	
	GOOD	BAD	GOOD	BAD
50	2486	186	1343	676
51	2470	183	1359	679
52	2449	177	1380	685
53	2429	174	1400	688
54	2411	170	1418	692
55	2388	165	1441	697
56	2365	160	1464	702
57	2343	154	1486	708
58	2329	151	1500	711
59	2306	150	1523	712
60	2290	149	1539	713
61	2257	147	1572	715
62	2232	141	1597	721
63	2212	134	1617	728
64	2180	129	1649	733
65	2149	125	1680	737
66	2104	115	1725	747
67	2062	108	1767	754
68	2026	106	1803	756
69	1956	94	1873	768
70	1866	80	1963	782
71	1799	79	2030	783
72	1707	71	2122	791
73	1608	66	2221	796
74	1534	65	2295	797
75	1441	57	2388	805
76	1319	49	2510	813
77	1231	44	2598	818
78	1097	34	2732	828
79	987	29	2842	833
80	897	22	2932	840

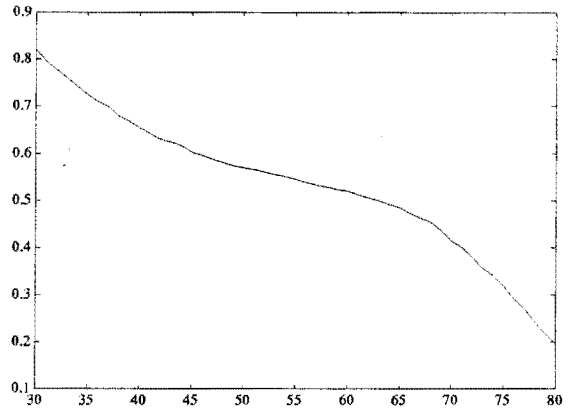


Figure 1

Percentage of granted credits for differents minimum score levels.

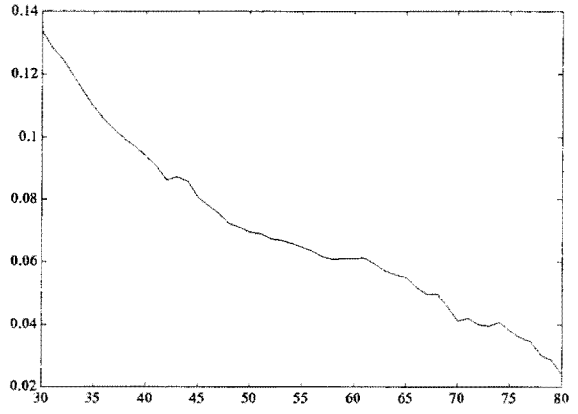


Figure 2

Percentage of bad applicants among those granted.

Looking at Figure 2, we see that it is obvious that between threshold 60 and 70, there is a zone of fluctuations. Therefore we suggested that applications receiving a score between those two levels should be regarded as doubtful, and should be analysed more carefully.

Let us make an example, if we decide that 80 will be the minimum score for granting credit, then only 919 credits would be granted of the inicial 4691 in the sample. 22 bad credits and 897 good credits. On the other hand 3772 would be denied , from those 2932 would be good and 840 bad. If we believe that this minimum score is too high then we can set the threshold to 79 points,

with this new minimum, 97 more applications are accepted and of those only 7 are bad credits. Using this strategy, we establish the optimal threshold. The best minimum point is about 60, according to the previous criterium of the financial institution based on profitability and risk. Therefore, the bad group is correctly classified (denied) with a percentage of 83%. 61% of the good clients are correctly classified (accepted). 52.36% of the applications would be accepted. Among those accepted only 6.1% would be bad credits, which means about 1 in every 20 credits would be delinquent.

5. STATISTICAL PROCEDURES AND SOFTWARE FOR THE AUTOMATIC EVALUATION

5.1. Discriminant function estimation

All the statistical procedures applied used the sample of almost 5000 applications and were performed using the computer statistical analysis system SAS. At the begining the basic procedures were used, for data description and depuration. For the elaboration of the decison model, we used STEPDISC and DISCRIM, that allow the estimation of the discrimant function of the model. Finally, we developed a tool in SAS to evaluate the results of the estimated models. Once, we found the final model, the following step was to develop a special purpose software to automatize the evaluation of new applications according to the model found.

5.2. Software for the automatic evaluation of new credit applications

For the practical implementation, we have designed a new autonomous system for the evaluation of applications on the basis of the classification model selected. Moreover, with the use of this software, the calculation of the final score remains as a black box, not the credit manager nor the applicant know the specific weights of the different items of information and their contribution to the final result. Thus, manipulations would be reduced and biased sores would be avoided.

The computer software elaborated by the authors has the objective of implementing the final model. The system has been prepared for IBM compatibles under DOS. It has been written in the programming language FORTRAN. The

software has two different modules. Firstly, the interactive module reads the items of information required by asking the applicant or the credit manager. Secondly, the scoring module uses the information to create new variables and calculates the final score that will be between 0 and 100, according to the posterior probability of the applicant to pay back.

A remarkable characteristic of the system is the facility to adjust for new estimations of the model that may vary along time.

6. FINAL REMARKS

The evaluation of credit applications, for loans of moderate quantities, may be supported by a quantitative tool based on the historical behaviour of applicants that were granted credit.

The main characteristic of the selected discriminant model is the powerful evaluating facility, the simplicity of the information required to give the final score. The applicant will normally inform only about a limited number of information items, which the bank must verify to ensure its reliability.

The statistical methodology applied to the context of finantial problems has lead to the development of the decision tool. This entails an innovation in speed and computerisation of the problem of granting credit, that will undoubtedly contribute to provide a dinamicity specially adequate for this kind of transactions, faster and providing better service to the client.

The last conclusion has to do with the validation of the presented model for classification. On the one hand, we must take into account that there must be a continuous reviewing of the evolution of the portfolio, analysing the behaviour and its changes. The percentages of defaulters should be frequently supervised to see whether the politics of granting must be altered so that the global risk acquired by the institution is not increased.

BIBLIOGRAPHY

[1] **Anderson , T.W.** (1958). *An Introduction to Multivariate Statistical Methods.* John Wiley and Sons.

[2] **Altman E.R., Avery R., Eisenbeis R. and J. Sinkey** (1981). *Classification Techniques with Applications to Business and Finance.* JAI Press.

[3] **Boyes, W.J. Hoffman, D.L. and S.A. Low** (1989). "An Econometric Analysis of the Bank Credit Scoring Problem". *Journal of Econometrics*, **40**, 3–14.

[4] **Fisher, R. A.** (1936). "The use of multiple measurements in taxonomic problems". *Ann. Eugen.*, **7**, 179–188.

[5] **Hand, D.J.** (1981). *Discrimination and Classification.* John Wiley and Sons.

[6] SAS User's Guide (1983). Sas Institue, Cary , U.S.

QÜESTIIÓ, Vol 18, 3 pp. 397-415, 1994

ANÁLISIS INTERACTIVO DE LAS SOLUCIONES DEL PROBLEMA LINEAL MÚLTIPLE ORDENADO

E. CONDE SÁNCHEZ, F.R. FERNÁNDEZ GARCÍA
y J. PUERTO ALBANDOZ

Departamento de Estadística e I.O.
Universidad de Sevilla

En este trabajo se estudian procedimientos para la elección entre las soluciones eficientes del Problema de Programación Lineal con Objetivos Múltiples, cuando el decisor manifiesta preferencias sobre ciertas ordenaciones de las valoraciones de las funciones objetivo, utilizándose como criterio de valoración global funciones basadas en el k−ésimo valor del vector de los objetivos ordenado en cada punto.

Se desarrollan procedimientos para generar ordenaciones compatibles con la información del decisor. Estos se incorporan en un proceso interactivo que proporciona una solución eficiente final con un mayor compromiso, a juicio del decisor, en las valoraciones de todos los objetivos.

Interactive analysis of the multiple ordered Lineal Problems solutions.

Key words: Programación Lineal con Objetivos Múltiples, Teoría de Decisión, Análisis de Sensibilidad.

Clasificación A.M.S.: 90C05, 62C99.

Autor para correspondencia: J. Puerto Albandoz

–Article rebut el febrer de 1993.
–Acceptat el febrer de 1994.

1. INTRODUCCIÓN

El problema de Programación Lineal con Objetivos Múltiples (MOLP) ha sido ampliamente estudiado, ver por ejemplo Yu, Zeleny (1976), Gal (1977, 1979), Steuer (1986), y consiste en encontrar un vector x del conjunto factible $X = \{x \in \mathbb{R}^n / Ax \leq b,\ x \geq 0,\ b \in \mathbb{R}^m\}$ que maximice las funciones lineales $c^1x, c^2x, \ldots, c^px$.

El conjunto de soluciones eficientes representa el concepto de solución de este problema. La determinación de dicho conjunto es difícil puesto que incluso el problema de determinar el número de vértices de un poliedro es un problema NP−duro, véase Dyer (1983) presentándose además la dificultad adicional de la elección de una solución de entre todas las eficientes cuando se desea implantar una política, ver por ejemplo Steuer (1986), White (1982) y Schrijver (1986).

Para destacar una solución eficiente acorde con las preferencias del decisor, es necesario que éste suministre alguna información adicional que será incorporada al problema original, véase Rios, French (1991). Por ejemplo en el caso en el que se acepte que existe una función de valor tipo lineal, $U(x) = \sum_{i=1}^{p} w_i c^i x$, si el decisor proporciona las ponderaciones de los distintos criterios se obtendrá un orden débil sobre X. Ahora bien, cuando la valoración de las ponderaciones no pueda realizarse de modo tan preciso, el decisor podrá optar por introducir relaciones de importancia entre las ponderaciones o entre los criterios, llamándose a esta información *aproximación cualitativa*, ver por ejemplo Paelinck (1975), Claessen *et al.* (1991).

La información suministrada restringe el conjunto de soluciones del problema, pero no asegura en todos los casos la unicidad de la solución. Es por tanto necesario proporcionar al decisor una metodología que le permita realizar la "elección" de la solución de acuerdo con dicha información. En esta metodología la información puede utilizarse en la reducción del conjunto factible y/o en la búsqueda de una función de valor.

La reducción del conjunto factible puede realizarse, entre otras formas, utilizando la ordenación de las valoraciones entre las funciones objetivo.

En lo referente al segundo punto, se han propuesto en la literatura diversas funciones de valoración, ver por ejemplo Wald (1950), destacando aquellas que hacen uso de la ordenación de los elementos $c^1x, c^2x, \ldots, c^px$, y consideran las valoraciones extremas: maximax y maximin. Estas ordenaciones entre los objetivos son aplicables cuando éstos estén valorados en una misma escala, o puedan

ser transformados a dicha escala, algunos de estos métodos de transformación pueden ser consultados en Hwang y Yoon (1981). En lo que sigue se supone que estamos en este caso.

Los criterios de valoración anteriormente expuestos pueden generalizarse a cualquier posición del orden de las valoraciones, dando lugar a criterios que equilibran el optimismo y el pesimismo de los objetivos extremos. Siendo la función de valoración asociada al lugar $k-$ésimo en un punto factible $x \in X$

$$U_k(x) = c^{\sigma_k} x \quad \text{donde} \quad c^{\sigma_1} x \leq c^{\sigma_2} x \leq \cdots \leq c^{\sigma_p} x$$

y $\sigma_1, \sigma_2, \ldots, \sigma_p$ es una permutación de $1, 2, \ldots, p$, véase Puerto (1990). Esto plantea la resolución del problema

$$\max U_k(x) \quad \text{sujeto a} \quad Ax \leq b,\ x \geq 0 \quad (P_k)$$

para obtener la decisión óptima con este criterio.

El problema (P_k) es un problema de programación no convexa, salvo para los casos extremos, $k = 1$ y $k = p$, maximin y maximax, por lo que presenta mayor dificultad de resolución que los dos problemas extremos. Ahora bien, el problema (P_k) es lineal sobre ciertas particiones de $\mathbb{R}^n$, lo que permite dar algoritmos para su resolución y estudiar la sensibilidad de las soluciones, véase Fernández, Puerto (1992).

En este trabajo se utilizan las dos vías de explotación de la información (búsqueda de función de valor y reducción del conjunto factible) en el desarrollo de un algoritmo interactivo, en el cual la información se suministra mediante ordenaciones entre las funciones objetivo. Este algoritmo requiere un orden completo entre los objetivos para proporcionar una solución inicial. Además tiene flexibilidad para permitir efectuar cambios en la función globalizadora y en la ordenación de objetivos mediante el análisis de cada nueva solución.

El resto del artículo se compone de tres secciones. En la primera se estudian procedimientos para generar ordenaciones completas entre objetivos, compatibles con la información suministrada por el decisor. En la segunda se proponen métodos que, basándose en el conjunto factible modificado (por las ordenaciones), destacan una solución eficiente. Se utilizan para este cometido funciones globalizadoras con sistemas de ponderaciones que reflejan la importancia de los diversos objetivos. Por último, se da un procedimiento interactivo que, partiendo de una solución inicial con respecto a un orden dado y mediante análisis de sensibilidad de la misma, nos conduzca a una solución que sea aceptada por el decisor.

2. DETERMINACIÓN DE UN ORDEN ENTRE LOS OBJETIVOS

La determinación de un orden acorde con la información suministrada por el decisor es una labor delicada, por lo cual ésta es una fase clave en el proceso de resolución del problema planteado y depende del nivel de información de que disponga aquel. Distinguimos fundamentalmente dos casos:

a) posee información de orden completa o incompleta, siendo capaz de estructurarla en relaciones de tipo requerido y

b) no posee información de orden o no es capaz de estructurarla en relaciones de tipo requerido.

En el primer caso, estas relaciones se incorporan directamente al problema. En otro caso, nosotros proponemos unos procedimientos para estructurar las preferencias del decisor (total o parcialmente) en relaciones que generan un orden de compromiso.

En primer lugar vamos a considerar las ordenaciones naturales obtenidas al elegir (muestrear) p soluciones eficientes $x^*(i)$, $i = 1, \ldots, p$, siendo el orden natural asociado a $x^*(i)$

$$\left(c^{\sigma(i)_1}, c^{\sigma(i)_2}, \ldots, c^{\sigma(i)_p}\right) \text{ tal que: } c^{\sigma(i)_1} x^*(i) \leq c^{\sigma(i)_2} x^*(i) \leq \cdots \leq c^{\sigma(i)_p} x^*(i)$$

traduciéndose los empates en relaciones de igualdad entre los objetivos en las condiciones del orden.

Nótese que en el caso de información incompleta, si el decisor es capaz de escoger una de entre estas soluciones que verifique la información parcial, su orden natural asociado proporciona una primera ordenación de compromiso. En cualquier caso debemos resaltar que esta elección no es definitiva, ya que este orden puede ser modificado con ayuda de un análisis de sensibilidad que será abordado en una sección posterior.

Una forma coherente (pues trata igualmente a todos los objetivos) y sencilla que proponemos para realizar este muestreo consiste en optimizar todas las funciones objetivo de modo individual, y obtener las ordenaciones asociadas a las soluciones óptimas. Esto es $x^*(i)$ $i = 1, \ldots, p$, solución de

$$\max c^i x \quad \text{sujeto a:} \quad Ax \leq b,\ x \geq 0 \quad P(i)$$

Estos problemas se resuelven secuencialmente, pues la región factible de todos ellos es la misma, y el vértice solución óptima de uno de ellos sirve como solución inicial del problema siguiente, al cambiar sólo la función objetivo.

En el caso de que ninguno de los órdenes naturales asociados a las soluciones $x^*(i)$ $i = 1, \ldots, p$, sea compatible con la información de orden del decisor, proponemos el siguiente esquema para la elección del orden inicial:

Conocidas las soluciones eficientes $x^*(i)$, $i = 1, \ldots, p$, valoramos todas las funciones objetivo en estos puntos:

Tabla 1

	$x^*_{(1)}$		$x^*_{(i)}$		$x^*_{(p)}$
c^1x	z_{11}	$\cdots$	z_{1i}	$\cdots$	z_{1p}
	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
c^ix	z_{i1}	$\cdots$	z_{ii}	$\cdots$	z_{ip}
	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
c^px	z_{p1}	$\cdots$	z_{pi}	$\cdots$	z_{pp}

A partir de esta tabla, ordenamos cada solución para los diferentes objetivos, esto es, reordenamos cada columna de la matriz anterior dando paso a:

Tabla 2

	1		i		p
PRIMERO	$z_1^{(1)}$	$\cdots$	$z_i^{(1)}$	$\cdots$	$z_p^{(1)}$
	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
I − ÉSIMO	$z_1^{(i)}$	$\cdots$	$z_i^{(i)}$	$\cdots$	$z_i^{(i)}$
	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
ÚLTIMO	$z_1^{(p)}$	$\cdots$	$z_i^{(p)}$	$\cdots$	$z_p^{(p)}$

Definiendo n_{ij} como el número de veces que el objetivo i−ésimo está en posición j, tenemos la frecuencia absoluta de aparición del objetivo i−ésimo en el lugar j. Este valor puede ser considerado como un índice objetivo de la aportación del objetivo c^i a la ordenación en posición j. Por tanto una forma de establecer el orden relativo entre los objetivos, sin intervención subjetiva del decisor, es considerar la asignación óptima dada por la solución del siguiente problema de asignación:

$$\max \sum_{i,j}^{p} w_{ij} x_{ij}$$

$$\sum_{i=1}^{p} x_{ij} = 1 \quad j = 1, \ldots, p$$

$$\sum_{j=1}^{p} x_{ij} = 1 \quad i = 1, \ldots, p$$

siendo $w_{ij} = \dfrac{n_{ij}}{\sum_{i,j}^{p} n_{ij}} \quad \forall i, j.$

La solución de este problema nos da la ordenación inicial, que representaremos por ρ, siendo el objetivo i ordenado en posición j, si la variable x_{ij} toma el valor 1.

Nótese que en los casos de información incompleta en los que se fijan algunos objetivos en determinadas posiciones, el esquema propuesto es válido. Simplemente se procede eliminando en la tabla 2, las filas correspondientes a los objetivos fijados y las columnas correspondientes a las posiciones que les asignaron a priori. Lo que se traduce en reducir las dimensiones del problema de asignación.

Si no existe ningún punto factible verificando el orden ρ, escogeremos el orden natural más "similar" a dicho orden.

Para ello, definimos una medida de disimilaridad entre dos ordenaciones σ, π, que identificamos por los índices de sus objetivos ordenados:

$$\text{dis}\,(\sigma, \pi) = \frac{\begin{array}{c}\text{mínimo número de intercambios entre índices consecutivos} \\ \text{para transformar la ordenación } \sigma \text{ en } \pi.\end{array}}{\text{máximo número de intercambios entre dos ordenaciones}}$$

Si definimos el número de inversiones de un índice π_i de π frente a σ como:

$$\text{inv}\,(\pi_i) = \text{ card } (\{\pi_j : \pi_j = \sigma_r;\ j > i \quad \text{y} \quad r < k\}) \quad \text{siendo } \pi_i = \sigma_k$$

dado que el mínimo número de intercambios entre índices consecutivos coincide con el número de inversiones de cada uno de los índices de la ordenación π frente a la σ y que, el máximo número de inversiones entre dos ordenaciones de p índices

viene dado por $\binom{p}{2}$, tenemos la disimilaridad entre dos ordenaciones dada por

$$\text{dis}\,(\sigma,\pi) = \frac{\sum_{i=1}^{p-1} \text{inv}\,(\pi_i)}{\binom{p}{2}} \quad \text{si } \sigma,\pi \in \mathcal{P}erm(1,\ldots,p)$$

Esta definición verifica propiedades que concuerdan con el sentido de la disimilaridad entre regiones ordenadas:

1) $\text{dis}\,(\sigma,\sigma) = 0 \quad \forall\sigma \in \mathcal{P}erm(1,\ldots,p)$

2) $0 \leq \text{dis}\,(\sigma,\pi) \leq 1 \quad \forall\sigma,\pi \in \mathcal{P}erm(1,\ldots,p)$

3) $\min_{\sigma\neq\pi} \text{dis}\,(\sigma,\pi) = \frac{1}{\binom{p}{2}}$ y se alcanza en todas aquellas ordenaciones que difieren en un único índice.

El orden que proponemos se obtiene:

Partiendo del orden ρ, elegir como decisión la solución del problema $P(i^*)$, si:

$$\text{dis}\,(s, s(i^*)) = \min_{i\in\mathbb{I}} \text{dis}\,(s, s(i))$$

siendo $s(i)$ la permutación asociada a la ordenación de los objetivos en $x^*(i)$ e $\mathbb{I}$ el conjunto de índices asociados a las ordenaciones que verifican las condiciones impuestas por el decisor.

En el caso en que el conjunto de índices $\mathbb{I}$ sea vacío podemos optar por realizar un nuevo muestreo de soluciones eficientes, o bien podemos hacer $\mathbb{I} = \{1,\ldots,p\}$ y dar una ordenación inicial de entre las existentes, dejando para el proceso interactivo la posibilidad de cambios en la misma.

3. ESTRATEGIA DE LOS VÉRTICES ÓPTIMOS

Una vez conocida la ordenación entre objetivos: $c^{\sigma_1}x \leq c^{\sigma_2}x \leq \cdots \leq c^{\sigma_p}x$, proporcionada por el procedimiento descrito en la sección anterior, vamos a determinar una solución eficiente inicial que verifique las condiciones anteriores.

Para ello añadimos al conjunto de restricciones del problema original (P) las referidas a la ordenación, obteniendo un nuevo problema múltiple que denotaremos por $(P; c^{\sigma_1}, c^{\sigma_2}, \ldots, c^{\sigma_p})$:

$$\text{"Maximizar"} \{c^1x, c^2x, \ldots, c^px\}$$

$$\begin{aligned} \text{sujeto a } & Ax \leq b \\ & c^{\sigma_i}x \leq c^{\sigma_{i+1}}x,\ i = 1, 2, \ldots, p-1 \\ & x \geq 0 \end{aligned}$$

La elección de una solución eficiente puede realizarse por cualquier método. Ahora bien, cuanta mayor información del decisor se utilice en este proceso, menor será el esfuerzo que se tendrá que realizar en el proceso iterativo, puesto que la solución aquí obtenida es el punto de partida del algoritmo que describiremos en la sección siguiente.

Para dar una solución a este problema se suelen emplear criterios de tipo lexicográfico, pero esta aproximación tiene el inconveniente de trabajar sólo con los objetivos extremos; por lo que es recomendable emplear todas las funciones objetivo.

El proceso se puede llevar a cabo partiendo de p soluciones eficientes cualesquiera, pudiendo ser las obtenidas al resolver los problemas unidimensionales con el nuevo conjunto factible.

$$\max\{c^{\sigma_k}x : Ax \leq b,\ c^{\sigma_1}x \leq c^{\sigma_2}x \leq \cdots \leq c^{\sigma_p}x,\ x \geq 0\} \quad (P; c^{\sigma_1}, \ldots, c^{\sigma_p}; c^{\sigma_k})$$

Representaremos a estas soluciones por $x(P; c^{\sigma_1}, \ldots, c^{\sigma_p}; c^{\sigma_k})$, (y abreviadamente por $x^{(\sigma_k)}$). Posteriormente podemos escoger entre ellas la solución inicial (estrategia de los vértices óptimos). Los empates en las soluciones óptimas en el criterio c^{σ_k}, se dilucidan empleando secuencialmente los otros criterios $c^{\sigma_{k+1}}, \ldots, c^{\sigma_p}, c^{\sigma_{k-1}}, \ldots, c^{\sigma_1}$.

En la elección entre las p soluciones $\left(x^{(\sigma_1)}, x^{(\sigma_2)}, \ldots, x^{(\sigma_p)}\right)$ obtenidas en este caso, se utilizará la información de todos los criterios, pues cada una de estas soluciones será valorada por las p funciones objetivo del problema lineal con objetivos múltiples. Así este proceso de decisión queda descrito por la matriz $\mathbf{D} = (d_{ij})_{p\times p}$ cuyos elementos d_{ij} son $c^{\sigma_i}x^{(\sigma_j)}\ \forall 1 \leq i, j \leq p$, que corresponden a las valoraciones de cada solución con todas las funciones objetivo. Por construcción se verifica que

$$c^{\sigma_1}x^{(\sigma_j)} \leq \cdots \leq c^{\sigma_k}x^{(\sigma_j)} \leq \cdots \leq c^{\sigma_p}x^{(\sigma_j)}.$$

Para la elección entre las diversas alternativas se puede utilizar sobre $\mathbf{D}$ cualquier criterio de decisión multiatributo, ver por ejemplo Huang, Yoon (1981).

Otra metodología diferente de elección se basa en suponer la existencia de una función de valor lineal sobre el problema. De esta forma si el decisor aporta información incompleta sobre las ponderaciones de dicha función de valor, se puede recomendar el uso de ciertos criterios. Estudiamos dos sistemas de ponderaciones propuestos por Kmietowicz, Pearman (1981).

En el primero de ellos el decisor supone que sus ponderaciones verifican

$$W_1 = \left\{ w \in \mathbb{R}^p : \sum_{i=1}^{p} w_i = 1,\ w_i \geq 0\ \forall i,\ w_1 \geq w_2 \geq \cdots \geq w_p \right\}$$

ponderando en mayor medida las valoraciones menores para no incurrir en decisiones ultraoptimistas. Entonces se verifica el siguiente teorema.

Teorema 1

La función $F\left(Cx_{(w)}^{(\sigma_j)}\right) = \sum_{k=1}^{p} w_k c^{\sigma_k} x^{(\sigma_j)}, w \in W_1,$ verifica

$$c^{\sigma_1} x^{(\sigma_j)} \leq F\left(Cx_{(w)}^{(\sigma_j)}\right) \leq \frac{1}{p} \sum_{k=1}^{p} c^{\sigma_k} x^{(\sigma_j)} \quad \forall w \in W_1$$

Demostración

El teorema de Paelinck, véase Claessen *et al.* (1991) establece que el conjunto de vectores $(w_1, \ldots, w_n) \in \mathbb{R}^p, w \in W_1$ es el cierre convexo de los vectores

$$\begin{aligned} s^1 &= (1, 0, \ldots, 0) \\ s^2 &= (\tfrac{1}{2}, \tfrac{1}{2}, \ldots, 0) \\ &\vdots \\ s^p &= (\tfrac{1}{p}, \tfrac{1}{p}, \ldots, \tfrac{1}{p}) \end{aligned}$$

Entonces por ser el objetivo lineal, los valores superiores e inferiores están en los extremos, de donde se obtiene el resultado.

■

Este teorema indica que el criterio suma es optimista y el criterio mínimo es pesimista, pues acotan la función anterior para cualquier valor de $w \in W_1$, superior e inferiormente. Así, proponemos como criterio de decisión una combinación

convexa de ambos criterios para decidir la acción óptima sobre la matriz **D**. El parámetro de esta combinación convexa, lo determinará el decisor de acuerdo con su compromiso ante las valoraciones de la función $F(\cdot)$.

En otras situaciones de decisión, es posible no sólo disponer de la ordenación del sistema de ponderaciones, sino además controlar las diferencias entre las mismas (escala de intervalo) a través de unas constantes $\{k_i\}$, que representan la importancia relativa de un peso respecto a los restantes. Este constituye el segundo conjunto de ponderaciones de los citados anteriormente que denominamos W_2:

$$W_2 = \left\{ w \in \mathbb{R}^p / w_1 - w_2 \geq k_1,\ w_2 - w_3 \geq k_2, \ldots, \right.$$
$$\left. \ldots, w_{p-1} - w_p \geq k_{p-1},\ w_p \geq k_p,\ k_p \geq 0, \sum_{i=1}^{p} w_i = 1 \right\}$$

La caracterización semejante a la del teorema de Paelinck para este conjunto de pesos viene dada por el siguiente resultado.

Teorema 2

Para todo $w \in W_2$ si $\sum_{i=1}^{p} i k_i \leq 1$ se verifica

$$c^{\sigma_1} x^{(\sigma_j)} \left(1 - \sum_{j=1}^{p} j k_j\right) + K \leq F(C x_{(w)}^{(\sigma_j)}) \leq \frac{1}{p} \sum_{k=1}^{p} c^{\sigma_k} x^{(\sigma_j)} \left(1 - \sum_{j=1}^{p} j k_j\right) + K,$$

$$\forall x \in X, \quad \text{con } K = \sum_{i=1}^{p} w_i^{\circ} c^{\sigma_i} x^{(\sigma_j)}$$

$$\text{y} \quad w^{\circ} = (k_1 + k_2 + \cdots + k_p,\ k_2 + \cdots + k_p, \ldots, k_{p-1} + k_p, k_p)$$

Demostración

Sea **A** la matriz $\begin{pmatrix} 1 & -1 & \cdot & \cdots & 0 \\ \cdot & 1 & -1 & \cdots & 0 \\ & & & \ddots & \\ \cdot & \cdot & \cdot & \ddots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$ que define las restricciones de W_2, por lo que $\mathbf{A}^{-1}$ es la matriz $\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$.

Si llamamos $y = \mathbf{A}w$, W_2 puede definirse de forma equivalente por $\sum_{i=1}^{p} jy_j = 1$, $y_j \geq k_j$ $j = 1, \ldots, p$. Este conjunto tiene por vectores extremos:

$$\hat{y}^i = (k_1, \ldots, \beta_i, \ldots, k_p), \quad i = 1, \ldots, p.$$

con $\beta_i = 1/i \left(1 - \sum_{j=1}^{p} jk_j + ik_i\right)$ $\quad i = 1, \ldots, p.$

Por tanto los puntos extremos de W_2 vendrán dados por $w^i = \mathbf{A}^{-1}\hat{y}^i$, es decir, $w^i = w^\circ + \left(1 - \sum_{j=1}^{p} jk_j\right) s^i$, $i = 1, \ldots, p$.

Usando estos resultados podemos determinar las cotas superiores e inferiores de $F(Cx_{(w)}^{(\sigma_j)})$ en W_2 con lo que se prueba el resultado.

■

Nótese que si $\sum_{i=1}^{p} ik_i > 1$, el cono W_2 es vacío, lo que se interpreta como una incoherencia en la información proporcionada.

Corolario

Bajo las condiciones del teorema anterior y siendo $\lambda \in [0, 1]$ se verifica:

a) Si tomamos $k_1 = 1 - \lambda$, y $k_i = 0$, $i = 2, \ldots, p$

$$c^{\sigma_1} x^{(\sigma_j)} \leq F\left(Cx_{(w)}^{(\sigma_j)}\right) \leq \lambda \cdot \frac{1}{p} \sum_{k=1}^{p} c^{\sigma_k} x^{(\sigma_j)} + (1 - \lambda) c^{\sigma_1} x^{(\sigma_j)}$$

b) Si tomamos $k_i = 0$, $i = 1, 2, \ldots, p-2$, $k_{p-1} = -\lambda$, $k_p = \lambda$

$$(1-\lambda)c^{\sigma_1} x^{(\sigma_j)} + \lambda c^{\sigma_p} x^{(\sigma_j)} \leq F\left(Cx_{(w)}^{\sigma_j)}\right) \leq (1-\lambda) \cdot \frac{1}{p} \sum_{k=1}^{p} c^{\sigma_k} x^{(\sigma_j)} + \lambda c^{\sigma_p} x^{(\sigma_j)}$$

Por tanto la conclusión de este resultado es que el criterio de Hurwicz es un criterio pesimista y el cent-dian, véase Halpern (1978), optimista para el conjunto de pesos que define b). Y para el conjunto definido en a) el criterio del mínimo

es nuevamente pesimista y como criterio optimista tenemos un criterio anticentdian, obtenido sustituyendo el criterio máximo por el mínimo.

Según el cono que el decisor desee emplear, en función de la información que pueda aportar, se elegirá la decisión óptima. A veces se emplea el criterio que toma por pesos el valor medio de todos los extremos del cono lo que se conoce como *criterio del centroide*, ver por ejemplo Van Huylebroeck (1990). Formas alternativas de seleccionar una ponderación pueden encontrarse en Barron, Schmidt (1988).

4. ESTRATEGIA INTERACTIVA

La estrategia de vértices descrita en el apartado anterior nos ha permitido destacar una solución eficiente con un orden de objetivos. Este orden es compatible con la información que el decisor suministró o bien ha sido generado por un procedimiento que incorpora la mayor información de orden intrínseca al problema.

Sin embargo esta solución pudiera ser mejorada al compararla con otras eficientes adyacentes a la misma respecto al conjunto factible, pues el decisor no conoce si disminuciones en el objetivo que fue destacado para encontrar la solución eficiente, pueden conducir a aumentos de otros objetivos.

Proponemos en este apartado utilizar los resultados de Fernández, Puerto (1992) para desarrollar un algoritmo interactivo, que manteniendo la eficiencia adquiera un mayor compromiso en las valoraciones y en el orden de los objetivos.

Para el estudio interactivo de $(P; c^{\sigma_1}, c^{\sigma_2}, \ldots, c^{\sigma_p})$ partimos de la solución óptima aceptada según el apartado anterior, que suponemos optimiza el criterio c^{σ_k}.

Consideraremos que la matriz $\overline{\mathbf{A}}$ de restricciones del problema, posee una descomposición $\overline{\mathbf{A}} = [B, N]$; donde B son las columnas correspondientes a la base óptima del problema y N las correspondientes a las columnas no básicas. De igual forma descomponemos $C = [C_B, C_N]$. Asimismo, sea x_{B_r} la $r-$ésima variable de la base óptima del problema $(P; c^{\sigma_1}, \ldots, c^{\sigma_p}; c^{\sigma_k})$ y $\overline{b}$ el vector de términos independientes (RHS). Denotemos por $\hat{b} = B^{-1}\overline{b}$ el término independiente en la solución óptima y por $y_{.j}$ la columna $j-$ésima de $\overline{\mathbf{A}}$, que se obtiene al ser multiplicada por la inversa de la base óptima.

Una representación tabular pormenorizada de todos los elementos de este problema es:

	Variables del problema	Holguras de restricción $\mathbf{A}x \leq b$	Holguras de restricción del orden	
Coeficiente de objetivo	C	0	0	RHS
Restricciones originales del problema	$\mathbf{A}$	I	0	b
Restricción de orden	$C^i - C^j$	0	I	0
Precios Sombra de Objetivos	$Z - C$			

Nótese que $\overline{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & I & 0 \\ C^i - C^j & 0 & I \end{bmatrix}$ y que $\overline{b} = \begin{bmatrix} b \\ 0 \end{bmatrix}$. Además, denotaremos por $\Theta_{rj} = \dfrac{\hat{b}_r}{y_{rj}}$, $C_{.j}$ la columna j–ésima de C y $Z_{.j} = C_B y_{.j}$.

Si x' es una solución adyacente a $x = x(P; c^{\sigma_1}, \ldots, c^{\sigma_p}; c^{\sigma_k})$, que se obtiene introduciendo en la base de x, una variable x_j y sacando de la base de x, la variable que ocupa el lugar r, x_{B_r}; x' conserva el orden entre los objetivos si y sólo si se verifica que:

$$0 < \min_s \left\{ \overline{b}'_s / \overline{b}'_s = \hat{b}_s - \Theta_{rj} y_{sj}, \quad y_{rj} \neq 0 \right\}$$

(véase Fernández, Puerto (1992)).

Además la diferencia entre los valores de los objetivos en ambas soluciones x, x' viene dada por:

$$\Delta Cx = Cx - Cx' = \Theta_{rj}(Z_{.j} - C_{.j})$$

Dado un cambio de base la expresión anterior nos muestra el cambio de todos los objetivos, pudiendo destacar la mejora de algunos de ellos y la cantidad en que

disminuye el objetivo $k-$ésimo. La información de los costos sombra (positivos o negativos) en cada cambio de base, servirá de ayuda al decisor para comparar x' con x.

El proceso iterativo que proponemos se basa en implicar al decisor en la selección entre soluciones eficientes adyacentes, a través de la información que señalábamos anteriormente. Además difiere de otros métodos (ver capítulo 13 de Steuer (1986)) en que se desarrolla sobre un problema con restricciones de orden sobre los objetivos.

Partiendo de $x^{(\sigma_k)}$ que optimiza el problema $(P; c^{\sigma_1}, \ldots, c^{\sigma_p}; c^{\sigma_k})$ el decisor escoge entre las variables no básicas $(j \in N)$ aquellas que a su juicio desea considerar a la luz de las valoraciones de los precios sombra. A este conjunto de variables seleccionadas por el decisor lo denominamos $\mathcal{J}$.

Un modo de elección de este conjunto $\mathcal{J}$ puede ser escoger en el conjunto

$$\{j/j \in N \quad \text{y} \quad \#\{j/Z_j - C_j < 0\} > h\}$$

siendo h un nivel fijado por el decisor. Es decir que el número de objetivos que podrán mejorar con este cambio de base sea mayor que la constante h. El decisor propone un orden de selección de estas variables no básicas, que puede ser en orden decreciente del número de objetivos mejorados.

Para cada variable no básica considerada j, se comprueba si el cambio de base factible conduce a una nueva solución eficiente, a través del (E-test) de Gal (1977):

Dado $j \in N$, resolvemos el problema (E-test)

$$\text{Min } \xi_j \quad \text{s.t. } -w'(Z-C)+\xi = 0 \quad \text{y} \quad \sum_{i=1}^{p} w_i = 1 \quad \text{con } w \text{ y } \xi \geq 0$$

Si la solución óptima ξ_j^* es 0, nos indica que el cambio sobre la variable j nos lleva a una base eficiente, ya que ambas bases tienen en común al menos el valor del parámetro solución del anterior problema, w^*.

Si se supera este test, se calculan las variaciones de las funciones objetivo $Cx - Cx'$ dadas anteriormente, en caso contrario se escoge otro $j \in N$.

Los aumentos y disminuciones que se produzcan conducirán a que el decisor acepte esta nueva solución o mantenga la antigua. La no aceptación de la solución también puede ser motivada porque la ordenación entre objetivos no satisface plenamente al decisor. En este caso el algoritmo puede modificar el

orden, conociendo previamente la factibilidad de la nueva ordenación y las mejoras que ésta aportará a los objetivos. Para ello utiliza el siguiente resultado de Fernández y Puerto (1992):

Si P' es el problema cuya ordenación intercambia el orden de los objetivos k y $(k+1)$–ésimo respecto del actual $(P; c^{\sigma_1}, \ldots, c^{\sigma_p}; c^{\sigma_k})$, es decir, P' es $(P; c^{\sigma_1}, \ldots, c^{\sigma_{k-1}}, c^{\sigma_{k+1}}, c^{\sigma_k}, \ldots, c^{\sigma_p}; c^{\sigma_k})$ entonces se verifica:

a) Si $c^{\sigma_{k+1}} x^{(\sigma_k)} - c^{\sigma_k} x^{(\sigma_k)} > 0$, o P' no tiene solución o el óptimo de P' es menor que el de $(P; c^{\sigma_1}, \ldots, c^{\sigma_p}; c^{\sigma_k})$.

b) Si $c^{\sigma_{k+1}} x^{(\sigma_k)} - c^{\sigma_k} x^{(\sigma_k)} = 0$, el óptimo de P' mejora estrictamente el valor anterior si y sólo la variable de holgura de la restricción $(A1)$ es no básica en toda solución óptima de $(P; c^{\sigma_1}, \ldots, c^{\sigma_p}; c^{\sigma_k})$.

El proceso se detiene cuando con una ordenación satisfactoria para el decisor, no existan variables a considerar o cuando los valores de los objetivos alcancen niveles admisibles según las preferencias del decisor (criterio de parada).

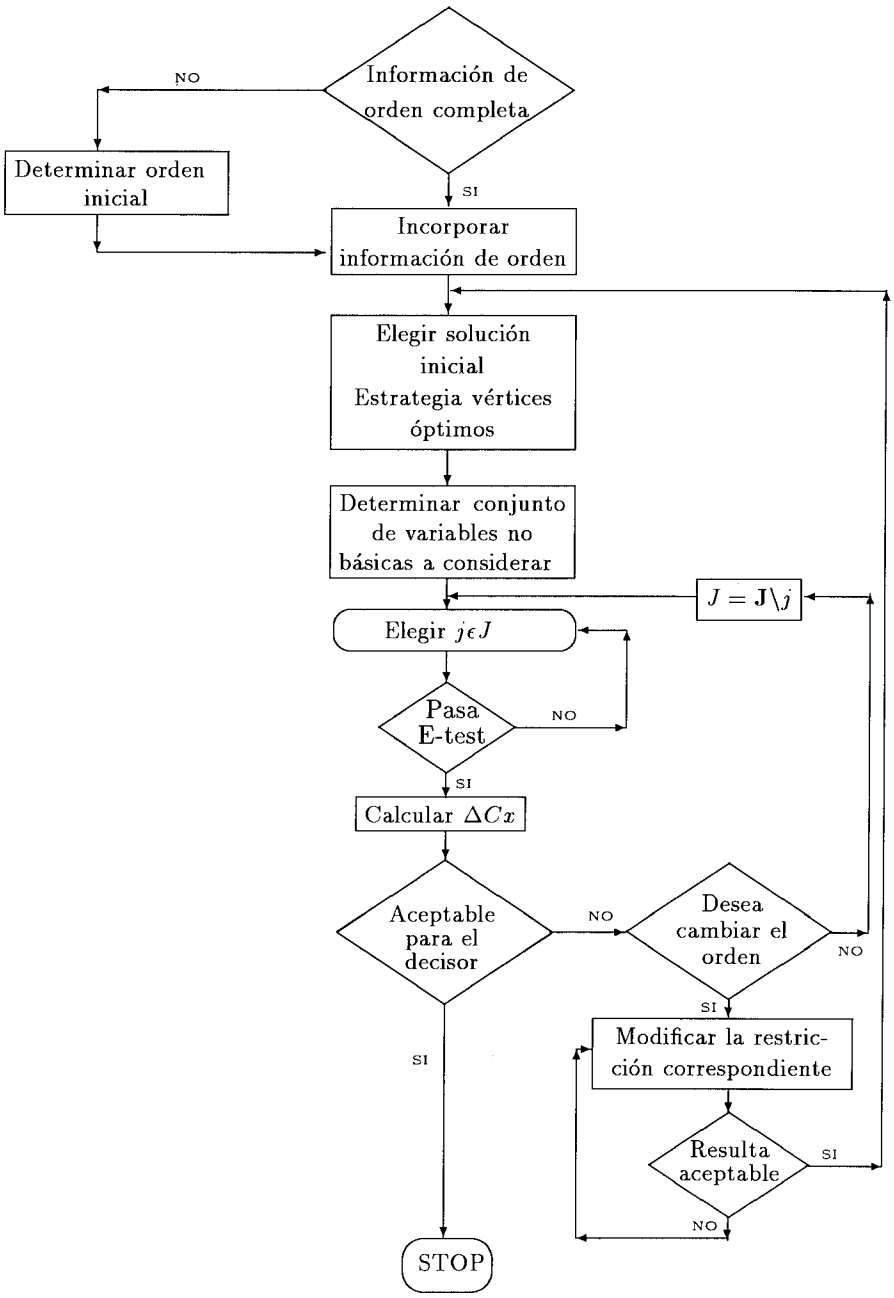
Un diagrama de flujo del esquema interactivo puede verse en el diagrama 1.

5. CONCLUSIONES

En este trabajo se desarrolla un procedimiento interactivo que recoge la información de orden entre objetivos, suministrada por el decisor, y determina una solución eficiente del problema de programación lineal con objetivos múltiples. En el proceso, tanto la fase de completación de la información de orden, como aquella en la que se genera una solución inicial, son flexibles en el sentido de poder ser modificadas y adaptadas a la naturaleza del problema considerado. En ambos casos se pueden utilizar métodos alternativos, existiendo por tanto la posibilidad de crear esquemas algorítmicos concretos en los que interactúen el método propuesto con nuevos desarrollos en el campo de la decisión multiatributo.

El uso de la hipótesis de linealidad de objetivos y restricciones del problema original ha sido utilizada en todo el esquema. La generalización a problemas no lineales no es inmediata puesto que en general no es posible la convexificación de los objetivos sobre regiones convexas del espacio factible, véase Puerto (1990), siendo esta línea de investigación objeto de estudio en este momento.

DIAGRAMA 1. ESQUEMA INTERACTIVO



AGRADECIMIENTOS

Los autores desean agradecer a un experto anónimo los interesantes comentarios y sugerencias realizados durante el período de revisión de este trabajo.

REFERENCIAS

[1] **Barron, H.** & **Schmidt, C.P.** (1988). "Sensitivity analysis of additive multiattribute value models". *ORSA,* vol. **36, 1**, 122–127.

[2] **Claessen, M., Lootsma, F.** & **Vogt, F.** (1991). "An elementary proof of Paelinck's Theorem on the Convex Hull of Ranked Criterion Weights". *EJOR,* vol. **52, 2**, 255–258.

[3] **Dyer, M.E.** (1983). "The complexity of vertex enumeration methods". *Mathematic of Operations Research,* vol. **8, 3**, 381–402.

[4] **Fernández, F.R.** & **Puerto, J.** (1992). "Análisis de Sensibilidad de las soluciones del problema lineal múltiple ordenado". *Trabajos de Investigación Operativa,* vol. **17, 1**, 17–29.

[5] **Gal, T.** (1977). "A general method for determining the set of all efficient solution to a linear vectormaximum problem". *EJOR,* vol. **1**, 307–322.

[6] **Gal, T.** (1979). *Postoptimal Analysis, Parametric Programming and related topics.* MacGraw-Hill.

[7] **Halpern, J.** (1978). "Finding minimal center-median convex combinations of a graph". *Management Science,* vol. **24**, 535–544.

[8] **Hwang, C.L.** & **Yoon, K.** (1981). *Multiple Attribute Decision Making.* Springer Verlag.

[9] **Kmietowicz, Z.W.** & **Pearman, A.D.** (1981). *Decision Theory and Incomplete Knowledge.* Gower.

[10] **Paelinck, J.H.P.** (1975). "Qualitative multiple criteria analysis, environmental protection and multi-regional development". *European Meeting of the Regional Science Association.* Budapest, 26–29 August.

[11] **Puerto, J.** (1990). "Programación Múltiple Ordenada". *Tesis Doctoral.* Universidad de Sevilla.

[12] **Rios, D.** & **French, S.** (1991). "A framework for sensitivity analysis in discrete multi-objective decision.making". *EJOR,* vol. **54, 2**, 176–190.

[13] **Schrijver, A.** (1986). *Theory of Linear and Integer Programming.* Wiley.

[14] **Steuer, R.** (1986). *Multiple criteria optimization.* Wiley.

[15] **Van Huylenbroeck, G.** (1990). "Applications of Multicriteria Analysis in the Environmental Impact Report for the High Speed Train". *JORBEL,* vol. **30, 4**, 32–52.

[16] **Wald, A.** (1950). *Statistical Decision Functions.* Chelsea, P.C.

[17] **White, D.J.** (1982). *Optimality and Efficiency.* Wiley.

[18] **Yu, P.** & **Zeleny, M.** (1976). "Linear Multiparametric Programming by Multicriteria Simplex Method". *Management Science,* vol. **23, 2**, 159–170.

ENGLISH SUMMARY:

INTERACTIVE ANALYSIS OF THE MULTIPLE ORDERED LINEAL PROBLEMS SOLUTIONS

E. Conde Sánchez, F.R. Fernández García and J. Puerto Albandoz

The Multiple Objective Linear Problem consists of finding a vector x in the feasible set $X = \{x \in \mathbb{R}^n : Ax \leq b,\ x \geq 0,\ b \in \mathbb{R}^m\}$ maximazing the linear functions $c^1x, \ldots, c^px$. (See e.g. Yu, Zeleny (1976), Gal (1977, 1979), Steuer (1986).

Since, an ideal solution maximazing simultaneously all these functions does not exist, the set of nondominated vectors is taken as solution-set for this problem. However, obtaining this set is a difficult task, but even assuming it is obtained, choosing a particular efficient solution all of them is not an esay problem (see Steuer (1986), White (1980)).

In order to select a particular efficient solution, the decision-maker (D–M) have to apport some extra-information to the original problem (see Rios and French (1991)). For instance, if the existence of a linear value function is assumed, the D–M must give the weights associated with de different objective functions. On the other hand, only ordering relations might be given, which

lead to the so called "*qualitative approach*". See e.g. Paelinck (1975), Claessen *et al.* (1991).

The additional information constraints the solution-set of this problem, although it cannot assure the uniquiness of solution. Hence, it is necessary to give to the D–M a methodology which simplifies the choice.

The goal of this paper is to ease the process of selection of a solution according with the fact that and extra-information is given. Two different ways are addressed. The first on deals with the generation of ordering among the objective functions, compatible with the extra-information given. The second one develops globalizing functions according to systems of weights.

The generation of compatible orderings covers two different approaches.

1. The D–M has a complete information and he is able to structure this information in the required kind of relations.

2. The D–M has incomplete information or he is not able to structure this information in the required kind of relations.

For the first case, these relations are incorporated to the feasible set, and for the second case, we give a procedure which generates a set of compromise relations. This arrangement is obtained as the optimal solution of assignment problem, where the coefficients are a measure of the adequation of a particular objective to a particular position in the arrangement of the objective functions.

Once the extra-information is adequately structured, we proposed different globalizing functions. The first approach leads us to the so called "*optimal vertex strategy*". This strategy chooses a feasible vertex with quit good properties as a solution. The second approach gives us two families of linear value functions with seights in two different sets. These families include most of the classical value functions in Decision Theory as particular instances.

The paper finishes developing an interactive algorithm which maintains the efficiency of the initial solution, using the E-test (Gal (1977)), and improves the adequation of the final solution with respect to the extra amount of information given by the D–M. This analysis is possible using the expressions for the variation in the objective functions given by

$$\Delta Cx = Cx - Cx' = \theta_{rj}(Z_j - C_j)$$

The algorithm stops in a finite number of steps, and the obtained solution is either admissible for the D–M or has the highest scores in the objectives among the feasible solutions.

Estadística Oficial

QÜESTIIÓ, Vol 18, 3 pp. 419-446, 1994

NOMENCLATURES I CLASSIFICACIONS ESTADÍSTIQUES

JOSEP MARIA MARTÍNEZ LÓPEZ*

Institut d'Estadística de Catalunya

Les classificacions econòmiques i laborals han experimentat en els diferents àmbits espacials un procés d'harmonització derivat de la necessitat de comptar amb un sistema integrat de nomenclatures i classificacions que permeti la comparabilitat de les dades estadístiques. Aquesta harmonització s'inicia, d'una banda, amb les classificacions d'activitats i productes que permeten elaborar estadístiques relacionades amb el funcionament econòmic i, d'altra banda, amb les classificacions d'ocupacions que permeten elaborar estadístiques del mercat de treball.

Les classificacions constitueixen uns llenguatges estadístics de referència, senzills i precisos, que responen a les necessitats de productors i usuaris de la informació estadística i s'avancen a les necessitats derivades de les noves tecnologies i productes.

A l'hora d'establir classificacions, els organismes estadístics, en primer lloc, recorren a la utilització de tècniques d'anàlisi multivariant que permeten conèixer millor les estructures productives i, en segon lloc, elaboren taules de correspondències per aconseguir la comparabilitat de la informació que prové de diferents classificacions.

Statistical nomenclature and classifications.

Key words: Nomenclatura, classificació, harmonització, taula de correspondències.

* Josep Maria Martínez López. Institut d'Estadística de Catalunya Departament d'Economia i Finances. Generalitat de Catalunya. Via Laietana, 58. 08003 Barcelona

-Article rebut l'octubre de 1994.
-Acceptat el gener de 1995.

0. PRESENTACIÓ

Tal i com s'indica a la introducció de la publicació de l'Institut d'Estadística de Catalunya (1993)4, una de les tasques que encomana la Llei d'estadística i la Llei del pla estadístic de Catalunya 1992-1995 a l'Institut d'Estadística de Catalunya consisteix en l'homogeneïtzació de la presentació dels resultats estadístics de Catalunya, amb la finalitat de facilitar la comparabilitat espacial i temporal de la informació estadística que es generi.

Un dels àmbits més característics, dins d'aquestes tasques, és la sistematització de nomenclatures i classificacions estadístiques per a l'harmonització de l'estadística catalana i la seva adaptació als sistemes estadístics estatal, comunitari i internacional. La trascendència d'aquesta activitat instrumental es posa de relleu en aspectes tant essencials per a la coordinació estadística com ara normalitzar l'elaboració, presentació i anàlisi de dades estadístiques, o bé facilitar la implementació de registres i arxius administratius susceptibles d'aprofitament estadístic.

D'aquest plantejament se'n deriva la necessitat de procedir prèviament a l'adopció de les classificacions bàsiques que s'han utilitzat històricament i majoritàriament en l'estadística de Catalunya, les quals constitueixin el marc on integrar posteriorment classificacions específiques i, a la vegada, el punt de referència per a futures actualitzacions de les nomenclatures adoptades. En aquest sentit, cal tenir present que l'establiment d'una nova classificació implica normalment utilitzar informació estadística elaborada segons l'antiga nomenclatura que es pretén actualitzar.

Així doncs, els criteris de comparabilitat espacial i temporal orienten les actuacions inicials de l'Institut d'Estadística de Catalunya cap a les traduccions adaptades de les classificacions vigents en l'àmbit d'activitat econòmica i productes i en el d'ocupacions, per tal de caracteritzar les unitats estadístiques bàsiques com són empreses, establiments i individus. Del seu caràcter bàsic i vertebrador en dóna compte el fet de que la majoria d'elles han estat recentment l'objecte prioritari d'un intens procés d'actualització per part dels principals organismes comunitaris i internacionals, amb la revisió subsegüent de les nomenclatures corresponents establertes a nivell estatal.

Aquesta circumstància justifica addicionalment la urgent traducció adaptada dels seus precedents més immediats, en els que encara s'emmarca la majoria de l'actual producció estadística catalana, per tal d'avaluar l'abast dels canvis que es plantegen i poder disposar de les correspondències que s'elaborin entre versions succesives quan s'escaigui.

Els treballs pel que fa a l'estadística pública catalana se centren en la traducció adaptada de les nomenclatures implicades, les notes explicatives derivades, l'elaboració de les principals abreviatures per a la tabulació estadística de resultats, l'articulació d'índexs temàtics per a la seva consulta i localització i l'establiment de correspondències parcials o completes amb altres classificacions prèviament adaptades. Les nomenclatures i classificacions adaptades i traduïdes esdeven de compliment obligatori, d'acord amb les normes que les fan oficials. L'Institut d'Estadística de Catalunya facilita la seva difusió en suport informàtic, a disposició dels agents i usuaris d'informació estadística.

Aquest treball es justifica per la necessitat d'adoptar i conceptualitzar un sistema integrat de classificació de la informació estadística i de la presentació dels seus resultats qualitatius i quantitatius.

1. INTRODUCCIÓ

En el marc de la Unió Europea, tant els estats membres (administracions públiques, empreses, institucions financeres i agents socials) com les institucions comunitàries, necessiten disposar d'informació estadística fiable, comparable i precisa. Aquesta necessitat requereix d'un sistema per classificar, d'una forma clara i ordenada, el gran nombre de fenòmens individuals en què hauran de basar-se les activitats estadístiques orientades a la producció i difusió de dades estadístiques.

D'acord amb el document de García-Parras, E. (1989)3, el desenvolupament de les classificacions estadístiques ha seguit un curs paral·lel al de l'aplicació creixent de mètodes estadístics per al coneixement de la realitat econòmica i social de cada país i s'ha vist ampliat, posteriorment, per les necessitats de comparabilitat internacional de les dades estadístiques. Així, les primeres classificacions van néixer de l'impuls dels professionals que es relacionaven amb el món de l'estadística i que es veien obligats a treballar amb un considerable nombre d'àmbits temàtics susceptibles de classificació; aquest impuls va ser el primer pas per determinar les classificacions estatals i internacionals que sovint difereixen entre elles, degut als criteris de conceptualització i sistematització utilitzats. Els canvis que s'han succeït en les classificacions han dificultat en molts casos la continuïtat de les sèries estadístiques temporals que s'havien establert. Vista la necessitat de poder comparar les dades estadístiques, els organismes estadístics oficials s'esforcen no només a actualitzar, sinó també a harmonitzar les nomenclatures i classificacions estadístiques.

Donada l'absència de referents normalitzats, l'Institut d'Estadística de Catalunya ha prioritzat el seu treball en matèria de nomenclatures estadístiques, des del primer any de desplegament del seu pla estadístic, a partir de la consideració simultània de tres requisits bàsics: procurar la màxima comparabilitat amb les principals classificacions estadístiques vigents, especialment en l'àmbit estatal i, subsidiàriament, en àmbits comunitaris o internacionals; preservar la continuïtat més àmplia possible de les sèries estadístiques disponibles i afavorir l'elaboració de classificacions pròpies per a l'estadística catalana.

Dins la metodologia científica les classificacions apareixen com un instrument lògic d ordenació sistemàtica i categoritzada dels coneixements adquirits, a partir del qual poden ser endegades noves investigacions. Les classificacions constitueixen un model on les modalitats d un determinat fenomen s ordenen segons una definició o criteri de classificació. En conseqüència, aquest model es transforma en un llenguatge de referència que permet la comparació i l'anàlisi del fenomen estudiat.

Els llenguatges estadístics han de satisfer unes determinades condicions:

a) Respondre a les necessitats dels principals productors i usuaris d'informació estadística, públics i privats, els quals no solament necessiten disposar de dades precises i comparables, sinó també d'un llenguatge estadístic comú que faciliti els seus treballs.

b) Ser senzills i precisos, evitant desenvolupaments injustificats.

c) Avançar-se, pel que fa a les futures necessitats d'ampliació del llenguatge derivades de l'evolució, cap a noves tecnologies i productes, evitant així revisions massa freqüents que invalidin les classificacions anteriors.

Les nomenclatures i classificacions constitueixen un sistema bàsic de llenguatges que asseguren la coordinació entre la recollida, la presentació i l'anàlisi de les dades estadístiques. Els termes nomenclatura i classificació són utilitzats moltes vegades com a sinònims, mentre que, per definició, corresponen a conceptes substantivament diferents.

El terme nomenclatura es defineix com una llista objectiva de terminologia de les característiques o modalitats que presenta un fenomen en estudi. Perquè aquesta llista sigui manejable, les característiques s'agrupen en categories en funció d'elements o propietats comunes, amb la qual cosa entrem en el camp de les classificacions.

El concepte de classificació implica l'organització i ordenació dels ítems d'una nomenclatura seguint uns criteris subjectius de classificació que actuen com a directrius per determinar la seva estructura, dins la qual s'establiran diversos nivells d'agregació.

2. CLASSIFICACIONS ECONÒMIQUES AMB FINALITATS ESTADÍSTIQUES

Les classificacions econòmiques amb finalitats estadístiques s'agrupen en dos grans tipus: les d'activitats econòmiques i les de productes. Les classificacions d'activitats econòmiques es concebeixen per classificar les unitats productives segons el tipus d'activitat a què es dediquin, amb vista a l'elaboració d'estadístiques sobre els aspectes relacionats amb el seu funcionament econòmic: producció, utilització dels factors de producció (treball, capital, materials) o rendibilitat. Les classificacions de productes es concebeixen per classificar els productes per tal d'elaborar estadístiques que permitin observar aquests productes en les seves relacions funcionals amb les diverses operacions econòmiques específiques a què siguin sotmesos (per exemple: producció, comerç exterior o transport). Es poden distingir dos tipus de productes, d'una banda, els béns transportables o mercaderies i, d'altra banda, els béns no transportables o serveis.

2.1. Classificacions d'activitats econòmiques

L'objectiu d'aquest tipus de classificació és establir un conjunt jerarquitzat de categories homogènies d'activitats econòmiques que permeti classificar les unitats que les porten a terme i implantar estadístiques diferenciades sobre aquestes activitats (producció, consum de primeres matèries, rendiments, ocupació, inversions, etc.).

S'entén per activitat econòmica tota acció productora resultant d'una concurrència de mitjans o factors productius (equipament, mà d'obra, procediments de fabricació, productes, etc.) que porta a la creació d'uns determinats béns o a la prestació de serveis concrets. Les activitats poden realitzar-se amb finalitats lucratives o sense.

A la pràctica, les unitats es classifiquen segons la seva activitat principal, definida aquesta com l'acció d'una unitat productiva de la qual s'obté el valor afegit màxim general (producció de béns o prestació de serveis). La resta d'activitats productores es classifiquen com a secundàries i s'ignoren des del punt de vista estadístic. Les seves dades queden assignades a l'activitat principal de l'explotació. La disponibilitat de dades comptables desagregades suficientment per a cadascuna de les activitats que realitza la unitat és un requisit difícil de satisfer per tal de poder efectuar una assignació directa a cadascuna d'elles.

2.2. Classificacions de productes

Per a la conceptualització estadística dels fluxos de béns transportables o mercaderies han estat dissenyades diferents classificacions que responen a les necessitats de les estadístiques de producció, consum, preus, transport, comerç exterior o necessitats duaneres. De fet, existeixen dos prototips d'aquestes classificacions: el que va adreçat a les estadístiques industrials i de producció, i el que va adreçat a les estadístiques de comerç exterior o transports.

Els criteris de classificació varien segons la finalitat; així, l'origen industrial i la destinació dels béns són els criteris que més s'utilitzen, quan la finalitat de la classificació respon a les necessitats de les estadístiques industrials o de producció.

La majoria d'aquestes classificacions de mercaderies són incomparables les unes amb les altres, de forma que per comparar la informació entre elles s'ha de fer mitjançant unes taules de correspondència parcials i complexes que mai asseguren una coherència perfecta de les dades i, per tant, hi pot haver una possible pèrdua d'informació.

Pel que fa als béns no transportables o serveis han estat dissenyades diferents classificacions que tenen com a objectiu recollir el resultat de les activitats de serveis.

3. L'HARMONITZACIÓ DE LES CLASSIFICACIONS D'ACTIVITATS ECONÒMIQUES

Durant els anys 1972 i 1973, a la Comissió d'Estadística de Nacions Unides i a la Conferència d'Estadístics Europeus, respectivament, per primera vegada es va arribar a un acord generalitzat per millorar l'harmonització entre les classificacions econòmiques internacionals, que fins aquell moment eren molt diferents i plantejaven problemes de comparabilitat de dades econòmiques entre països. És així com els principals organismes comunitaris i internacionals del sistema estadístic requereixen, entre d'altres instruments de coordinació, d'un sistema harmonitzat de nomenclatures.

Vistes les mancances existents en el marc de les classificacions econòmiques amb finalitats estadístiques, l'any 1974 la Comissió d'Estadística de Nacions Unides, conjuntament amb la Comunitat Europea, va aprovar un programa en què s'establien les principals directrius per a l'harmonització, entre les que figuraven

l'elaboració d'un sistema integrat de nomenclatures d'activitats i productes, i la creació d'una classificació central de productes no integrada amb les activitats, basada en el Sistema harmonitzat (SH) que servís tant per a les estadístiques de producció com per a les de comerç exterior.

Com a resultat, dins l'àmbit de les classificacions d'activitats econòmiques, al febrer de 1989, les Nacions Unides van aprovar la tercera revisió de la seva Classificació industrial internacional uniforme per a totes les activitats econòmiques (CIIU-Rev.3) i, a l'octubre de 1990, el reglament del Consell número 3037/90 va aprovar l'estructura de la Nomenclatura estadística d'activitats a la Comunitat Europea (NACE-Rev.1), les quals van substituir, respectivament, la Classificació industrial internacional uniforme per a totes les activitats econòmiques en la seva segona revisió (CIIU-Rev.2) de 1968, i la Nomenclatura general d'activitats econòmiques en les Comunitats Europees de 1970 (NACE-70), fins aleshores vigents.

L'estructura de la NACE-Rev.1 aprovada va obligar tots els països de la Comunitat Europea a la seva adopció, entre els quals figurava Espanya. Aquesta estructura va quedar establerta fins a un nivell de quatre dígits i la informació estadística de cada país membre havia d'integrar-se en aquesta estructura. Partint d'aquesta, l'Oficina d'Estadística de les Comunitats Europees (EUROSTAT) va deixar en llibertat, els països membres, per establir desglossaments a cinc dígits.

Al desembre de 1993, el Reial Decret 1560/1992 del Ministeri d'Economía i Hisenda va aprovar una nova Classificació nacional d'activitats econòmiques, vigent a partir de l'1 de gener de 1993 (CNAE-93), que es correspon a 4 dígits amb la NACE-Rev.1 i que incorpora un nivell de desglòs a 5 dígits. Aquesta classificació va substituir l'anterior Classificació nacional d'activitats econòmiques de l'any 1974 (CNAE-74)7, fins aleshores vigent.

En el context regional això suposa, d'una banda, assumir una estructura i criteris de classificació d'àmbit internacional, comunitari i estatal, que no sempre recull de forma òptima la realitat i, d'altra banda, afrontar la desagregació de la informació, subjecta al compliment inexcusable de preservació del secret estadístic i al condicionant d'una disponibilitat d'informació menor.

4. ASPECTES METODOLÒGICS DE L'ELABORACIÓ DE LA CNAE-93

A continuació s'exposen alguns dels aspectes metodològics que van servir de directriu en l'elaboració de la Classificació nacional d'activitats econòmiques de

l'any 1993 (CNAE-93), amb l'aplicació de tècniques d'anàlisi multivariant a les dades recollides en l'enquesta industrial. Tanmateix, aquest exercici, amb algunes variants, es pot realitzar amb les dades recollides de l'economia catalana, el qual ens permetrà obtenir una classificació d'activitats econòmiques que s'adeqüi de forma òptima a la realitat econòmica catalana.

Partint de la Nomenclatura estadística d'activitats a la Comunitat Europea (NACE-Rev.1) a 4 dígits, es planteja l'estudi d'un desglòs a 5 dígits. Per això és necessari tenir presents una sèrie d'elements: el primer és la partició que l'anterior Classificació nacional d'activitats econòmiques de l'any 1974 (CNAE-74) indueix en la classificació NACE-Rev.1, és a dir, les subdivisions que afageix la correspondència amb la CNAE-74 i que podrien ser elements d'estudi per a un possible desglossament; el segon que cal tenir en compte és la importància o entitat, present o futura, en termes de valoració monetària, nombre d'unitats o personal ocupat en l'activitat a desglossar; el tercer és l'especialització, és a dir, perd sentit dividir un agregat a 4 dígits en el qual les unitats de producció (empreses i establiments) classificades en ell, estiguin indiferenciades, per tant, que tots o quasi tots els productes són produïts de la mateixa forma per a cadascuna de les unitats de producció sense que prevalguin criteris majoritaris; el quart a tenir en compte és el fet de que sigui possible mesurar les variables d'interès en les unitats classificades en el mateix, i en cinquè i últim lloc, conservar sempre que sigui possible les sèries històriques.

El punt de partida d'aquest treball és el referit a l'any 1987, empleat per l'Institut Estadístic Suec (SCB) per a la revisió de la seva classificació d'activitats econòmiques i utilitzada per l'Instituto Nacional de Estadística per estudiar els apartats segon i tercer anteriors. Les descripcions han estat preses de Carlstrom, C (1988)2 i de Ballano, C., García-Parra, E. i Prieto, G. (1992)1.

L'objetiu és estudiar la informació que ens ofereix l'enquesta industrial en l'àmbit territorial de Catalunya per a l'any 1987 i analitzar las seccions C, D, E i F de la NACE-Rev.1 (divisions 10 a 45) per agregats a 4 dígits. En aquest marc és possible obtenir la següent informació:

a) Mesures de tamany. La més important d'elles és el valor de producció segons l'activitat principal dels establiments i la seva desagregació en valor de producció dels diferents productes.

b) Coeficients de cobertura i d'especialització. El coeficient de cobertura d'una determinada activitat és el tant per u de la producció de productes d'una classe que és produïda per una determinada indústria. El coeficient d'especialització d'una determinada activitat és el tant per u de la producció d'una determinada indústria que recau en productes d'una classe. Les mesures de tamany i els coeficients de cobertura i d'especialització

s'obtenen només per aquelles agrupacions NACE-Rev.1 a 4 dígits, que comparen activitats industrials estudiades en l'enquesta industrial i no delegades a altres organismes.

L'objetiu de l'estudi és agrupar els productes, segons la Classificació nacional de béns i serveis de l'any 1978 (CNBS-78)8 a 7 dígits, amb similar comportament per als establiments que tenen la mateixa activitat principal, segons la NACE-Rev.1 a 4 dígits. L'observació de les característiques comunes d'aquests grups de productes poden servir per determinar el criteri de divisió a partir del quart dígit.

El registre del fitxer de productes utilizat conté l'activitat principal de l'establiment, segons la CNAE-74 a 4 dígits, el codi de producte segons la CNBS-78 a 7 dígits, el valor de producció en aquest establiment d'aquest producte i el factor d'elevació de l'establiment.

La idea del procediment a utilitzar consisteix en calcular, per als establiments l'activitat principal dels quals és la categoria NACE-Rev.1, objecte d'estudi, les matrius dels coeficients d'especialització i cobertura a nivell de la CNBS-78 a 7 dígits. Aquestes matrius ja suggereixen els productes que convé que estiguin en un mateix grup. Malgrat això, el problema que es planteja en aquestes matrius és la quantitat de productes, que fa impossible realitzar el treball d'agrupació per la simple inspecció visual.

Per solucionar aquest problema, s'utilitza un procediment d'anàlisi cluster que automàticament porta a terme una agrupació, basant-se en els coeficients de cobertura i d'especialització. Es defineix, per això, el concepte de divergència entre dos productes com la diferència entre la unitat i el percentatge màxim de cobertura o d'especialització registrat.

Definida la divergència entre els productes, per fer els clusters es valoren diferents mètodes i diferents distàncies. El procés comença amb cada producte considerat com un grup per separat. En el primer pas, les divergències entre els grups corresponen amb les divergències entre els productes. El procés finalitza amb tots els productes en un únic grup.

Un cop finalitzat el procés es representen els grups que s'han anat formant mitjançant un dendograma, considerat aquest com una forma de fer patents les similituds entre els productes.

La utilització de mètodes d'anàlisi multivariant que permetin reconèixer suficientement les estructures productives i la construcció de taules de correspondències entre diferents classificacions de productes i nomenclatures de diferent naturalesa, que proporcionin informació coherent i comparable, són algunes

de les aproximacions metodològiques més freqüents que s'exposen i es valoren en aquest article.

5. LA CLASSIFICACIÓ CATALANA D'ACTIVITATS ECONÒMIQUES

La Classificació catalana d'activitats econòmiques (CCAE-93) és el resultat de l'adaptació a la llengua catalana de la Classificació nacional d'activitats econòmiques de l'any 1993 (CNAE-93), elaborada per l'Instituto Nacional de Estadística (INE), i que fou aprovada pel Reial Decret 1560/1992, de 18 de desembre de 1992. En aquest sentit, els comentaris sobre l'estructura i criteris de la classificació estadística esmentada que s'efectuen són plenament vigents per aquesta traducció adaptada.

Tal com indica la publicació de l'Instituto Nacional de Estadística (1993)10, l'estructura d'una classificació es troba determinada per diferents nivells d'agregació, els quals són el resultat lògic d'ordenar i organitzar els fenòmens a classificar. Cada nivell constitueix una nomenclatura, entenent com a tal la llista de designacions, on a cada fenomen a classificar li correspon un sol element de la nomenclatura.

La CNAE-93 està estructurada en cinc nivells més un nivell intermedi d'una forma jeràrquica piramidal.

Nom	Nivell	Nombre de rúbriques	Identificació
Secció	Primer	17	Codi alfabètic d'1 dígit
Subsecció	Intermedi	31	Codi alfabètic de 2 dígits
Divisió	Segon	60	Codi numèric de 2 dígits
Grup	Tercer	222	Codi numèric de 3 dígits
Classe	Quart	503	Codi numèric de 4 dígits
Subclasse	Cinquè	762	Codi numèric de 5 dígits

Característiques específiques:

1. El codi del primer nivell (seccions) és representat per un dígit alfabètic que no està integrat en el codi numèric de divisió. Per tant, el primer dígit numèric no té cap significat propi, contràriament al que passava a la

CNAE-74. Els codis numèrics tenen significat a partir de les agrupacions a dos dígits.

2. El nivell intermedi (subseccions), entre el primer i segon nivell, recull la desagregació de dues de les 17 seccions; la secció C que correspon a les *Indústries extractives* amb 2 subseccions i la secció D que correspon a les *Indústries manufactureres* amb 14 subseccions.

3. Partint del segon nivell, existeix una integració numèrica de les rúbriques que formen l'estructura piramidal de la classificació.

4. De forma similar a la CNAE-74, quan una rúbrica no presenta subdivisions en un nivell immediat inferior, en aquest últim hi figurarà un 0.

Diferències estructurals entre la CNAE-74 i la CNAE-93

Nivell	CNAE-74	CNAE-93
Primer	10	17
Intermedi	—	31
Segon	63	60
Tercer	283	222
Quart	507	503
Cinquè	—	762

1. S'ha suavitzat el nombre de rúbriques entre els diferents nivells d'agregació de la classificació.

2. La CNAE-93 no utilitza el sistema decimal en el primer nivell de la classificació, constituït aquest per 17 seccions, que ha estat substituït per un codi alfabètic d'un sol dígit no integrat en la resta del codi.

3. Un cinquè nivell que possibilita una major desagregació de la classificació.

Diferències en continguts entre la CNAE-74 i la CNAE-93

El contingut de la divisió 1 que correspon a *Energia i aigua*, de la CNAE-74, es troba repartit entre la secció C que correspon a *Indústries extractives*, la secció D que correspon a *Indústries manufactureres*, i la secció E que correspon a *Producció i distribució d'energia, gas i aigua*, de la CNAE-93.

En el contingut de la divisió 2 que correspon a *Extracció i transformació de minerals no enèrgetics i productes derivats. Indústries químiques* de la CNAE-74 es classifica en dues seccions diferents de la CNAE-93, les seccions C i D.

El contingut de la divisió 6 que correspon a *Comerç, restaurants i hoteleria.* Reparacions de la CNAE-74 es troba en dues seccions diferents de la CNAE-93, d'una banda, el *Comerç* (secció G), i de l'altra, l'*Hosteleria* (secció H).

El contingut de la divisió 8 que correspon a *Institucions financeres, assegurances, serveis prestats a les empreses i lloguers* de la CNAE-74, es classifica en dues seccions diferents de la CNAE-93, d'una banda, les *Activitats d'institucions financeres i assegurances* (secció J), i de l'altra, les *Activitats dels serveis prestats a les empreses* (secció K).

El contingut de la divisió 9 que correspon a *Altres serveis* de la CNAE-74, ha estat molt dividit, distribuint-se en diferents seccions de la CNAE-93.

6. ASPECTES METODOLÒGICS DE L'ELABORACIÓ DE TAULES DE CORRESPÒN-DENCIA ENTRE DIFERENTS CLASSIFICACIONS DE PRODUCTES

Per a les estadístiques de fluxos de béns transportables o mercaderies han estat dissenyades diferents classificacions que responen a las necessitats de les estadístiques de producció, consum, preus, transport, comerç exterior o necessitats duaneres.

La majoria d'aquestes classificacions de mercaderies són incomparables entre elles. Per a la comparabilitat de la informació entre distintes classificacions de mercaderies i classificacions de diferent naturalesa, es requereix de la construcció de taules de correspondència. Amb aquestes es pot comparar la informació que prové de les estadístiques de comerç exterior codificades mitjançant la Nomenclatura combinada (NC) o l'Aranzel integrat comunitari (TARIC) i la que prové de les estadístiques de producció codificada mitjançant la Classificació nacional d'activitats econòmiques de l'any 1993 (CNAE-93). Aquesta comparabilitat es redueix als béns transportables o mercaderies (seccions A, B, C, D i E de la CNAE-93).

Entre els objectius de la CNAE-93 destaca el fet de formar part d'un sistema central de classificacions que permeti la comparabilitat de la informació pel que fa als diferents tipus de classificacions econòmiques, principalment d'activitats i de productes. A partir d'aquest moment és la Nomenclatura combinada (NC)

la que s'utilitza per a la codificació de mercaderies que són objecte de comerç entre els països de la Unió Europea. A nivell estatal, integrat dins de la NC, Espanya utilitza per a la codificació de mercaderies que són objecte de comerç amb tercers països, és a dir, aquells que no formen part de la Unió Europea, el TARIC, que coincideix en les seves vuit primeres xifres amb la NC.

D'altra banda, el 29 d'octubre de 1993, la Comunitat Europea va establir la Classificació de productes ordenats per activitat (CPA), que és una classificació de productes en què els seus elements estan estructurats d'acord amb l'origen industrial definit en la NACE-Rev.1, i que té idèntics els seus quatre primers dígits. Aquesta classificació es va elaborar per ser una classificació de productes central en el seu àmbit, així, per exemple, la classificació de productes industrials comunitaris (PRODCOM) és l'utilitzada en l'enquesta industrial europea i està integrada en la CPA, a més incorpora una taula de correspondències amb la NC. Mitjançant la classificació PRODCOM, la qual es correspon en els seus quatre primers dígits amb la CPA, el treball ha consistit en reelaborar la taula de correspondències establerta per EUROSTAT en sentit invers, per transformar la informació codificada amb una classificació de mercaderies, en una informació codificada amb una classificació d'activitats econòmiques. La relació que existeix entre aquestes dues classificacions, la NC i la CPA, pel que fa als béns transportables, ve establerta per la pròpia elaboració de la CPA a partir de rúbriques elementals de la Classificació central de productes (CPC), ja que aquestes al mateix temps es van formar per la unió de rúbriques elementals del Sistema harmonitzat (SH) i, per tant, de la NC. La correspondència entre la CPA o la PRODCOM i la NC s'obté mitjançant una matriu de pas, i entre la CPA o la PRODCOM i la NACE-Rev.1 existeix una relació d'ordre donat, que ambdues tenen una estructura comú, el que permetrà la comparabilitat de dades del comerç intra i extra comunitari amb les de producció a nivell de Unió Europea.

En aquest mateix àmbit, l'Institut d'Estadística de Catalunya ha elaborat una taula de correspondències entre la Nomenclatura combinada (NC) i la Nomenclatura comú de mercaderies (NCM), utilitzada per RENFE, per a la classificació de mercaderies fins a 1993, any que va ser substituïda per la Nomenclatura harmonitzada de mercaderies (NHM), una classificació comparable amb la NC. Aquesta taula permet la comparabilitat de la informació dels béns transportables per ferrocarril amb la informació que ja es disposa dels fluxos de duanes. Aquest apartat es complementa amb l'elaboració d'una taula de correspondències amb la Nomenclatura estadística de transports (NST), utilitzada per les estadístiques de transport per carretera, amb diverses classificacions que permetran la comparabilitat d'aquesta informació amb les dades duaneres. Aquestes taules ens poden aproximar a l'estimació de les dades de comerç de Catalunya amb la resta d'Espanya, que juntament amb les dades de comerç amb l'estranger conformarien les dades de comerç exterior de Catalunya.

7. CLASSIFICACIONS D'OCUPACIONS AMB FINALITATS ESTADÍSTIQUES

El primer objectiu que una classificació d'ocupacions pretén satisfer és el de disposar d'un instrument estadístic de classificació laboral que permeti la comparabilitat, en l'espai i en el temps, de dades estatals i internacionals.

També pretén oferir una llista uniforme de grups d'ocupacions que permeti als organitzadors de programes de migració o de programes internacionals de formació professional identificar en els diferents països categories professionals comparables.

Dins l'àmbit exclusivament estatal és evident la necessitat d'una classificació estadística d'aquest tipus i resulta útil per a estudis d'ocupació, atur, migració o classificacions censals i mostrals.

En aquestes classificacions les ocupacions s'agrupen en funció de la naturalesa del treball o de la tasca realitzada. Els grans grups constitueixen amplis camps professionals i, per tant, la seva delimitació no presenta cap tipus de problema. Però aquests problemes es plantegen a l'establir subgrups i grups primaris. Els subgrups s'estableixen quan existeix un nombre suficientment important de treballadors que ho justifiqui. A més hi ha una coincidència amb els grups convencionals utilitzats en estudis estadístics, especialment, en els censos de població.

Cada grup primari comprèn un nombre d'ocupacions que fa que existeixi una afinitat i homogeneïtat dins de cada grup; aquesta afinitat i homogeneïtat es perd quan ens trobem en presència d'un grup primari residual.

L'ocupació és l'últim nivell de treball, el més restringit, limitat i concret d'una classificació d'ocupacions. Aquest nivell engloba, en la seva denominació, ocupacions o càrrecs dels treballadors que realitzin qualsevol de les diferents combinacions que es puguin fer de les tasques.

A l'establir les denominacions dels grups i ocupacions s'utilitza un llenguatge d'ús corrent que reflecteix el contingut del grup o ocupació.

8. L'HARMONITZACIÓ DE LES CLASSIFICACIONS D'OCUPACIONS

Segons la publicació de l'Instituto Nacional de Estadística (1980)5/6, en l'àmbit internacional els primers resultats visibles en l'harmonització de classifi-

cacions d'ocupacions o professions se situen en el decurs de la setena Conferència Internacional d'Estadígrafs del Treball, a l'any 1949, on s'acordà l'establiment d'una classificació d'ocupacions, la primera edició de la qual va ser publicada per l'Organització Internacional del Treball (OIT), a l'any 1958.

Posteriorment, l'onzena Conferència, a l'any 1966, va representar el començament de l'estudi per a la revisió d'aquesta classificació, el qual es va perllongar fins al 1968, any en què es va publicar la Classificació internacional uniforme d'ocupacions de l'any 1968 (CIUO-68). Aquesta classificació suposava el marc de referència dels treballs que, a partir d'aquest moment, van portar a terme els instituts i organismes oficials de l'estadística estatal, quant a l'adaptació harmonitzada de classificacions en aquest àmbit.

La finalitat de la CIUO-68 era, doncs, oferir un sistema de classificació de les ocupacions que pogués servir de base per a l'establiment de comparacions internacionals adients en l'estadística laboral i en el mercat de treball, i per a les tasques de revisió de les classificacions nacionals d'ocupacions, de forma que es garantís la convertibilitat temporal d'aquestes amb la internacional.

Quant a l'estadística oficial d'àmbit estatal, l'Instituto Nacional de Estadística, amb la col·laboració de diversos organismes de l'Administració pública i privada, va elaborar una primera classificació nacional d'ocupacions (CNO-61), adaptada a l'àmbit internacional del 1958 i aprovada, oficialment, per la presidència del govern, Ordre de 20 de febrer de 1961, la qual va esdevenir la primera classificació d'utilització obligada.

Aquesta primera Classificació nacional d'ocupacions de l'any 1961 (CNO-61) constava d'una estructura més simple que la Classificació nacional d'ocupacions de l'any 1979 (CNO-79), basada en la desagregació dels anomenats grans grups, subgrups i grups unitaris identificats, respectivament, per un, dos i tres dígits. Addicionalment, la classificació incorporava les denominacions i les definicions, com és habitual. La classificació numèrica continuava subdividint-se fins a arribar a la categoria ocupacional concreta, codificada amb set dígits, encara que aquests darrers ítems s'establiren sense la corresponent definició o descripció de les tasques adscrites a l'ocupació o professió identificada. Aquesta classificació va representar una aportació molt important en el camp estadístic, atès que fou la primera vegada que es comptava amb un instrument de classificació laboral tan interessant.

La vigència de la Classificació internacional uniforme d'ocupacions de l'any 1968 (CIUO-68) es mantingué fins que els desenvolupaments científics i tecnològics de la dècada dels 80 generaren, progressivament, l'aparició de noves ocupacions i, en alguns casos, la quasi desaparició de determinades professions. En conseqüència, també l'Instituto Nacional de Estadística inicià el procés

d'elaboració d'una nova classificació nacional d'ocupacions, actualitzada per als àmbits internacional i estatal, basada en la revisió sistemàtica de la CNO-61 i amb la col·laboració d'experts del Ministeri de Treball.

Els principals trets que caracteritzen la Classificació nacional d'ocupacions de l'any 1979 (CNO-79) han estat, en una primera etapa, el criteri basat en l'ocupació (i no la professió o la categoria professional), el manteniment dels mateixos grans grups de la CIUO-68 (encara que s'introduïren algunes variacions en la seva titulació) i, en una segona etapa, l'establiment del quart nivell, de cinc dígits, el qual ha permès considerar la descripció o definició de les tasques individualitzades de les 1.568 ocupacions resultants.

A l'estudiar els subgrups i grups primaris de la CIUO-68, com a regla general només es van establir subgrups quan existia un nombre suficientment elevat per justificar la seva agrupació en un apartat diferenciat. En canvi, alguns subgrups de la CIUO-68 es van dividir en dos, perquè tenien ambdós entitat suficient en la realitat espanyola per constituir nuclis separats. En relació amb els grups primaris es van fer igualment canvis amb el mateix criteri anteriorment exposat.

Establerts els grans grups, subgrups i grups primaris, es va fer l'assignació de les claus numèriques fins al nivell de tres dígits i es van establir les denominacions i definicions fins aquest nivell.

En aquesta tasca van intervenir experts i es va requerir d'un estudi minuciós de les reglamentacions professionals i de treball, les ordenances laborals i els convenis col·lectius, amb aquestes dades es van identificar més de 8.000 possibles ocupacions.

Els criteris generals seguits en la presentació d'aquesta Classificació nacional d'ocupacions poden sintetitzar-se en els següents:

a) Respecte a la naturalesa d'una classificació d'ocupacions, aquesta requereix una descripció de tasques i no una enumeració de categories, facultats o atribucions.

b) Manteniment de la comparabilitat internacional.

c) Utilització de termes usuals en les denominacions i definicions, en la pràctica o en la legislació, els quals facilitin la identificació i reconeixement de l'ocupació. Això obliga, en alguns casos, a utilitzar denominacions de cossos o professions per a algunes ocupacions.

d) Descripció de tasques fonamentals, evitant les definicions extenses o reiteratives que puguin induir a confusió amb les d'altres ocupacions que tinguin tasques accessòries similars o idèntiques.

8.1. La classificació internacional uniforme d'ocupacions de 1988

Recentment, l'Oficina Internacional del Treball (OIT) ha atès la necessitat ineludible d'actualitzar la seva Classificació internacional uniforme d'ocupacions de l'any 1968 (CIUO-68), per motius similars als que va obligar a la revisió de la Nomenclatura d'ocupacions de l'any 1958. En aquest context, s'afegeix un nou element determinant de la vigència tècnica de l'antiga CIUO-68: el procés harmonitzador propiciat per l'Oficina Estadística de les Comunitats Europees (EUROSTAT) en la revisió del sistema actual de nomenclatures sobre activitat econòmica (NACE Rev.1) i l'adopció progressiva del Sistema europeu de comptes econòmics integrats (SEC), per part de la majoria d'instituts d'estadística oficials als estats membres de la Comunitat Europea.

Segons la publicació de l'Oficina Internacional del Treball (1990)9, els estudis iniciats en el si de les Conferències Internacionals d'Estadígrafs del Treball, en especial la tretzena i la catorzena que van tenir lloc a Ginebra el 1982 i el 1987, va originar, el 6 de novembre de 1987, en el marc de la catorzena conferència, l'anomenada Classificació internacional uniforme d'ocupacions de l'any 1988 (CIUO-88), publicada finalment el 1990 per l'OIT i que substitueix la de l'any 1968.

La Classificació internacional uniforme d'ocupacions revisada (CIUO-88) presenta un sistema de classificació i agregació de dades d'informació sobre les ocupacions obtingudes mitjançant els censos de població, altres estudis estadístics i els registres de les administracions públiques.

L'elaboració d'aquesta classificació va començar fa alguns decenis i ha estat sempre associada al treball desenvolupat en les Conferències Internacionals d'Estadígrafs del Treball, que tingueren lloc sota els auspicis de l'Oficina Internacional del Treball (OIT). La necessitat d'establir una classificació internacional uniforme d'ocupacions es va debatre per primera vegada el 1921, però el primer pas concret per a la seva realització fou l'adopció, per a la setena Conferència Internacional d'Estadígrafs del Treball el 1949, d'una classificació provisional en nou grans grups. Pocs anys més tard, a l'any 1952, l'OIT va publicar la Classificació internacional d'ocupacions per a les migracions i la col·locació, en la qual es descriuen detalladament 1.727 ocupacions, a partir de vuit classificacions nacionals de països industrialitzats. La primera edició de la CIUO va ser publicada l'any 1958 i revisada deu anys més tard, l'any 1968.

El nou marc de referència ha estat concebut per a la comparació homogènia d'estadístiques internacionals de caràcter laboral i suposa, de nou, un impuls paral·lel d'actualització de la majoria de classificacions d'ocupacions d'àmbit

estatal per part dels organismes competents, per a les quals la CIUO-88 pot servir de model.

Grans objectius

La CIUO-88 té tres grans objectius. El primer és facilitar, a escala internacional, la comunicació en matèria d'ocupacions oferint, als estadístics dels diferents països, un instrument que permeti situar les dades estatals sobre ocupacions en una perspectiva internacional.

El segon és el de permetre la presentació de les dades internacionals sobre ocupacions sota una forma que s'adeqüi, tant a la investigació com a l'adopció de decisions i a les iniciatives d'acció concreta en certs casos determinats, per exemple, pel que fa a les migracions internacionals i a la col·locació.

El darrer és el de servir de model als països que desitgin desenvolupar o revisar la seva pròpia classificació d'ocupacions. La CIUO-88 no està destinada a substituir cap de les classificacions nacionals d'ocupacions, les quals haurien normalment de reflectir, el més fidelment possible, l'estructura dels seus mercats d'ocupació. Malgrat això, els països, les classificacions dels quals ja s'inspiren en la CIUO-88, tant en la seva concepció com en la seva estructura, podran adoptar amb major facilitat els procediments que els permetin fer comparables les seves estadístiques de treball en el pla internacional.

Cal assenyalar, a més, que en molts casos els països poden desitjar que la classificació tingui una estructura més desagregada i definicions més elaborades que la classificació internacional. És possible que desitgin incorporar informacions codificades sobre elements del contingut de les tasques i la descripció detallada de les ocupacions que presentin un interès especial des del punt de vista de la solució dels conflictes sobre salaris, orientació i formació professionals, els serveis de col·locació, l'anàlisi de la morbiditat i la mortalitat en relació amb l'ocupació.

Marc conceptual

El marc necessari per a la concepció de la CIUO-88 es basa en dos grans conceptes: el tipus de treball realitzat, és a dir, l'ocupació, i la competència.

L'ocupació —definida com un conjunt de tasques complides o que se suposa que seran complides per una mateixa persona— constitueix la unitat estadística

de la CIUO-88. Les persones es classifiquen per ocupacions en funció de la seva relació amb una feina passada, present o futura.

La competència —definida com la capacitat de dur a terme les tasques inherents a un treball determinat— es caracteritza, dins la CIUO-88, pels següent trets:

a) El nivell de competències, el qual és funció de la complexitat i de la diversitat de les tasques.

b) L'especialització de les competències que estan relacionades amb l'amplitud dels coneixements exigits, les eines i màquines utilitzades, el material que es treballa i la naturalesa dels béns i serveis produïts.

Partint del concepte de competència, així definit, es van delimitar i agregar novament els grups ocupacionals de la CIUO-88.

Atès el caràcter internacional de la Classificació, es van utilitzar únicament quatre nivells de competències i, per a la seva definició, es van fer servir les categories i nivells que apareixen en la Classificació internacional normalitzada de l'educació (CINE).

Pla i estructura

L'enfocament conceptual adoptat per realitzar la CIUO-88 va donar com a resultat una estructura jeràrquica piramidal formada per deu grans grups al nivell més elevat d'agregació, subdividits, successivament, en vint-i-vuit subgrups principals, cent setze subgrups i tres-cents noranta grups primaris.

Vuit dels deu grans grups han estat correlacionats amb els quatre nivells de competències de la CIUO-88, els quals van ser definits fent referència a les categories i graus d'ensenyament de la Classificació internacional normalitzada de l'educació (CINE). El concepte de nivell de competències no s'ha aplicat ni al gran grup 1, membres del poder executiu i dels cossos legislatius i personal directiu de l'Administració pública i d'empreses, ni al gran grup 0, forces armades. Això és motivat pel fet que, segons les informacions recollides en les fonts nacionals, les competències necessàries per dur a terme les tasques inherents a les ocupacions de cadacun d'aquests dos grans grups difereixen, fins al punt de fer impossible, la seva vinculació amb qualsevol dels quatre nivells de competències de la CIUO-88.

Els grups ocupacionals han estat subdividits i se'ls ha diferenciat de manera més detallada, segons l'especialització de les competències que es defineix, fent

referència al camp de coneixement exigit, a les màquines i les eines utilitzades, als materials que es treballa i, també, al tipus de béns i serveis produïts.

Els vint-i-vuit subgrups, en el segon nivell d'agregació de la CIUO-88, constitueixen una novetat, ja que totes les classificacions internacionals d'ocupacions anteriors presentaven un buit important quant al nombre de grups en els nivells primer i segon d'agregació. Per exemple, la CIUO-68 contenia vuit grups al primer nivell d'agregació i vuitanta tres en el segon. Com a conseqüència, existia, d'una banda, un desequilibri entre el nombre de grups necessaris per presentar l'estructura general de les ocupacions amb una classificació paral·lela en funció de variables com branca d'activitat o grups d'edat detallats i, de l'altra, per presentar l'estructura general de les ocupacions sense una classificació paral·lela de cap tipus o en funció de variables com el sexe o grups amplis d'edat.

Els tres-cents noranta grups primaris que, en la CIUO-88, se situen a nivell de la major diferenciació comprenen, en general, més d'una ocupació. A cada país, el nombre d'ocupacions compreses i la seva diferenciació dependran, en gran mesura, de la importància de l'economia i del grau de desenvolupament; del nivell i de l'orientació de la tecnologia; de l'organització del treball, i de les tradicions. Per aquesta raó, no s'ha fet una descripció detallada de les ocupacions esmentades en cadacun dels tres-cents noranta grups primaris de CIUO-88.

9. LA CLASSIFICACIÓ CATALANA D'OCUPACIONS

La Classificació catalana d'ocupacions (CCO-94) és el resultat de l'adaptació a la llengua catalana de la Classificació nacional d'ocupacions de l'any 1994 (CNO-94), elaborada per l'Instituto Nacional de Estadística i que fou aprovada pel Reial Decret 917/1994, de 6 de maig de 1994. En aquest sentit, els comentaris sobre l'estructura i criteris de la classificació estadística esmentada que s'efectuen són plenament vigents per aquesta traducció adaptada.

Tal i com indica la publicació de l'Instituto Nacional de Estadística (1994)11, la CNO-94 està estructurada en quatre nivells d'agregació i per comprendre-la, cal veure com es genera el primer nivell (grans grups) partint de la definició de qualificació.

La CIUO-88 utilitza quatre nivells de qualificació que es relacionen amb les categories de la Classificació internacional normalitzada d'educació (CINE).

Gran grup	Descripció CCO	Nivell de qualificació de la CIUO-88
1	Personal directiu de les empreses i de les administracions públiques	—
2	Tècnics i professionals científics i intel·lectuals	4t.
3	Tècnics i professionals de suport	3r.
4	Empleats administratius	2n.
5	Treballadors de serveis d'hoteleria, personals, protecció i venedors de comerços	2n.
6	Treballadors qualificats en activitats agràries i pesqueres	2n.
7	Artesans i treballadors qualificats de les indústries manufactures, la construcció i la mineria, llevat dels operadors d'instal·lacions i maquinària	2n.
8	Operadors d'instal·lacions i maquinària, i muntadors	2n.
9	Treballadors no qualificats	1r.
0	Forces armades	—

Els nivells de qualificació de la CIUO-88, definits per l'Oficina Internacional del Treball (OIT), són els següents:

Nivell de qualificació	Categories de la CINE
Primer	Categoria 1 de la CINE, que correspon a l'ensenyament primari que comença, generalment, a l'edat de 6 o 7 anys i sol durar uns 5 anys.
Segon	Categories 2 i 3 de la CINE, que correspon als cicles 1r. i 2n. de l'ensenyament secundari. El 1r. cicle comença a l'edat d'11 o 12 anys i sol durar 3 anys, i el 2n. cicle comença a l'edat de 14 o 15 anys i sol durar, aproximadament, uns 3 anys. Podrà requerir-se un període de formació i d'experiència en l'ocupació. Aquest període pot completar una formació de tipus formal, reemplaçar-la en part, i en alguns casos, en la seva totalitat.
Tercer	Categoria 5 de la CINE (la categoria 4 de la CINE s'ha deixat deliberadament sense contingut), que comprèn l'educació que comença a l'edat de 17 o 18 anys i sol durar uns 4 anys i condueix a un diploma que no és equivalent a un 1r. grau universitari.
Quart	Categories 6 i 7 de la CINE que comprèn l'educació que comença a l'edat de 17 o 18 anys i sol durar uns 3, 4 o més anys, i dóna accés a un grau universitari, de postgrau universitari o a un títol equivalent.

En aquestes taules s'observa que tant el gran grup 1, *Personal directiu de les empreses i de les administracions públiques*, com el gran grup 0, *Forces armades*, no estan relacionats amb el nivell de qualificació de la CIUO-88, perquè són heterogenis respecte a la qualificació de les ocupacions contingudes en ell. Aquests grans grups són homogenis des del punt de vista de la finalitat d'aquestes ocupacions.

Per poder diferenciar els treballadors determinats pel nivell de qualificació 2n. de la CIUO-88, s'utilitza el concepte d'especialització de la qualificació. Així, d'una banda, es diferencien els *Empleats d'oficina* (gran grup 4) i els *Treballadors dels serveis de restauració, personals, protecció, etc.* (gran grup 5). Als altres treballadors, associats al 2n. nivell de qualificació, se'ls diferencia, d'una banda, pels oficis tradicionals (gran grup 6, *Treballadors qualificats en activitats agràries i pesqueres*, i gran grup 7, *Treballadors artesans i d'altres oficis*), i d'altra banda, per les ocupacions que essencialment estan orientades a la utilització de maquinària i instal·lacions industrials (gran grup 8, *Operadors d'instal·lacions i maquinària, i muntadors*).

Partint d'aquest primer nivell d'agregació, la CNO-94 presenta una estructura jeràrquica i piramidal de la següent forma:

Nom	Nivell	Nombre rúbriques	Identificació
Gran grup	Primer	10	Codi numèric d'1 dígit
Grup principal	Intermedi	19	Codi alfabètic d'1 dígit
Subgrup principal	Segon	65	Codi numèric de 2 dígits
Subgrup	Tercer	206	Codi numèric de 3 dígits
Grup primari	Quart	493	Codi numèric de 4 dígits

Diferències estructurals entre la CNO-79 i la CNO-94

Nivell	CNO-79	CNO-94
Primer	8	10
Intermedi	—	19
Segon	85	65
Tercer	309	206
Quart	—	493
Cinquè	1568	—

L'estructura de la CNO-94 presenta diferències substancials enfront la CNO-79. Així, en general, s'ha suavitzat l'estructura, com s'aprecia a la taula anterior. En particular, en la CNO-94 tots els grans grups estan identificats per un únic codi numèric, mentre que a la CNO-79 hi havia, per exemple, un gran grup identificat per tres dígits numèrics, el gran grup 7/8/9.

La principal diferència de l'estructura de la CNO-94 respecte a la de la CIUO-88 (o CIUO-88.COM) és l'existència d'un nivell intermedi denominat grup principal entre el primer i segon nivell d'agregació i que s'ha d'entendre com a possible alternativa als grans grups. El grup principal queda caracteritzat perquè pertany a un únic gran grup de la CIUO-88, i està format per agregacions exactes de rúbriques de 2 dígits de la CIUO-88.

En els codis de les rúbriques de la CNO-94 existeixen dos números amb un significat especial: el 0 i el 9. El 0 quan està situat a partir de la tercera posició significa que no hi ha desagregacions del nivell immediat anterior. El 9, a la quarta posició, significa una rúbrica del tipus Altres.

REFERÈNCIES

[1] **Ballano, C., García-Parra, E.** i **Prieto, G.** (1992). "Aspectos metodológicos en la elaboración de la Clasificación Nacional de Actividades Económicas (CNAE-93) a partir de NACE.Rev.1". *Revista Estadística Española.* **Vol. 34, 131**, 491–504.

[2] **Carlstrom, C.** (1988). "Revision of the swedisch SIC of all Economic Activities". *Proceedings of the Fourth Annual Research of the Bureau of the Census.*

[3] **Garcia-Parra, E.** (1989). *La nueva nomenclatura de actividades económicas de la Comunidad Económica Europea. (NACE Revisión 1).* Madrid: Instituto Nacional de Estadística.

[4] **Institut d'Estadística de Catalunya** (1993). *Classificació d'Activitats Econòmiques. Adaptació normalitzada de la CNAE-74.* Barcelona.

[5] **Institut d'Estadística de Catalunya** (1994). *Classificació d'Ocupacions. Adaptació normalitzada de la CNO-79.* Document de treball. Barcelona.

[6] **Instituto Nacional de Estadística** (1980). *Clasificación Nacional de Ocupaciones 1979 (CNO-79).* Madrid.

[7] **Instituto Nacional de Estadística** (1984). *Clasificación Nacional de Actividades Económicas 1974 (CNAE-74).* Madrid.

[8] **Instituto Nacional de Estadística** (1987). *Clasificación Nacional de Bienes y Servicios.* Madrid.
[9] **Oficina Internacional del Trabajo** (1990). *Clasificación Internacional Uniforme de Ocupaciones (CIUO-88).* Ginebra.
[10] **Instituto Nacional de Estadística** (1993). *Clasificación Nacional de Actividades Económicas 1993 (CNAE-93).* Madrid.
[11] **Instituto Nacional de Estadística** (1994). *Estructura de la Clasificación Nacional de Ocupaciones 1994 (CNO-94).* Madrid.

ENGLISH SUMMARY:

STATISTICAL NOMENCLATURE AND CLASSIFICATIONS

Josep Maria Martínez López

0. PRESENTATION

One of the tasks entrusted to the Catalan Institute of Statistics, in response to the Law on statistics and the Law on the statistical plan for Catalonia for 1992-1995, consists in standardising the presentation of Catalonia's statistical results, in order to facilitate the spatial and temporal comparability of the statistical information which is produced. One of the most important facets of this work is the systematisation of statistical nomenclature and classifications, in order to coordinate Catalan statistics and adapt these to Spanish, Community and international statistical systems.

1. INTRODUCTION

Given the absence of standardised references, the Catalan Institute of Statistics has given priority in its work in the field of statistical nomenclature and classifications to the equal consideration of three basic requirements: to obtain

the maximum degree of comparability with the main statistical classifications in force, to maintain the greatest possible continuity of the statistical series available and to favour the development of classifications which are suitable for Catalan statistics.

2. ECONOMIC CLASSIFICATIONS FOR STATISTICAL PURPOSES

There are two important types of economic classifications for statistical purposes: classifications of economic activities and those of products. The classifications of economic activities are designed to classify production units, whereas the classifications of products are designed to elaborate statistics which will enable these products to be observed in their functional relationships with the various economic operations. Two types of products can be distinguished: on the one hand, transportable goods or merchandise, and on the other hand, non-transportable goods or services.

3. COORDINATION OF THE CLASSIFICATIONS OF ECONOMIC ACTIVITIES

During 1972 and 1973, an overall agreement was reached by the United Nations Statistical Commission and at the Conference of European Statisticians to improve the degree of concordance between international economic classifications, by virtue of which it was planned to elaborate an integrated nomenclature system for activities and products. As a result, in February 1989, the United Nations approved the third revision of its International Standard Industrial Classification of all Economic Activities (ISIC Rev.3), and in 1990, the structure of the Statistical Classification of Economic Activities in the European Community (NACE Rev.1) was approved, whereupon these substituted the classifications which had been in force up to that time.

Adoption of the structure of the approved NACE Rev.1 became obligatory for all the countries of the European Community. In December 1993, the Spanish Treasury department approved a new National Classification of Economic Activities, to take effect from 1 January 1993 (CNAE-93), which corresponds to NACE Rev.1 with 4 digits and incorporates a level of breakdown of 5 digits.

4. METHODOLOGICAL ASPECTS OF THE ELABORATION OF CNAE-93

On the basis of the Statistical Classification of Economic Activities in the European Community (NACE Rev.1) with 4 digits, the study of a breakdown of 5 digits is planned. The starting point for this work is the 1987 study, employed by the Swedish Institute of Statistics (SCB) to revise its classification of economic activities and subsequently used by the National Institute of Statistics (INE) in the elaboration of CNAE-93.

5. THE CATALAN CLASSIFICATION OF ECONOMIC ACTIVITIES

The Catalan Classification of Economic Activities (CCAE-93) is the result of adapting the 1993 National Classification of Economic Activities (CNAE-93).

The structure of CNAE-93 is that of a pyramid, consisting of five hierarchical levels, plus an intermediate level. The main structural differences between CNAE-74 and CNAE-93 are: firstly, the number of sections between the different aggregation levels of classification has been reduced; secondly, CNAE-93 does not use the decimal system in the first level of classification, but uses an alphabetical code of one single digit which is not integrated into the rest of the code, and thirdly, CNAE-93 makes use of a fifth level which facilitates a greater disaggregation of the classification.

6. METHODOLOGICAL ASPECTS OF THE ELABORATION OF TABLES OF CORRESPONDENCE BETWEEN DIFFERENT PRODUCT CLASSIFICATIONS

The majority of the classifications of merchandise are not intercomparable. In order for the information presented by different classifications of merchandise and classifications of different kinds to be compared, the construction of tables of correspondence is required. These make it possible to compare the information presented by external trade statistics, coded by means of the Combined Nomenclature or the Integrated Community Tariff, and the information presented by production statistics, coded by means of CNAE-93. This comparability is limited to transportable goods or merchandise.

7. CLASSIFICATIONS OF OCCUPATIONS FOR STATISTICAL PURPOSES

In these kinds of classifications the occupations are grouped according to the nature of the work or task performed. These classifications are structured in different levels of aggregation: the large groups represent broad professional fields; the sub-groups are established when there is a sufficiently large number of workers to justify this; the groups are those used in statistical surveys, especially in population censuses; the primary groups consist of a certain number of occupations, making homogeneity possible within each group; finally, the occupation is the last level of work and the most restricted, limited and specific level of an occupational classification.

8. COORDINATION OF THE OCCUPATIONAL CLASSIFICATIONS

The 1968 International Standard Classification of Occupations (ISCO-68) served as a frame of reference for the work involved in the coordinated adaptation of classifications in this field. With respect to official statistics covering Spain, the INE produced a first National Classification of Occupations (CNO-61), adapted to the work on an international scale of 1958, which became the first classification of obligatory use. Subsequently, the INE produced the 1979 National Classification of Occupations (CNO-79), updated for international and national purposes and based on the systematic revision of CNO-61.

The International Labour Office (ILO) updated its 1968 International Standard Classification of Occupations (ISCO-68) for reasons similar to those which made the revision of the Nomenclature of Occupations necessary in 1958. These studies resulted in the 1988 Standard International Classification of Occupations of (ISCO-88), which was finally published by the ILO in 1990 and superseded the 1968 classification.

ISCO-88 provides a system for classifying and aggregating occupational information obtained by means of population censuses, other statistical surveys and administrative records. The framework necessary for designing ISCO-88 has been based on two main concepts: the occupation and the skill.

The occupation is defined as a set of tasks executed by one person and constitutes the statistical unit classified by ISCO-88, and the skill is defined as the ability to carry out the tasks of a given job. On the basis of the skill concept

thus defined, the occupational groups of ISCO-88 were delineated and further aggregated. The structure of CNO-94 is that of a pyramid, consisting of four hierarchical levels, based on the four levels of qualification related to the categories which appear in the International Standard Classification of Education (ISCED). On the other hand, differences can be observed between the structure of CNO-94 and that of CNO-79, in that the structure of the former has been simplified; in particular, all the large groups are identified by one single numerical code, whereas in CNO-79 there was a large group identified by three numerical digits.

The Catalan Classification of Occupations (CCO-94) is the result of adapting the 1994 National Classification of Occupations (CNO-94) produced by the INE.

QÜESTIIÓ, Vol 18, 3 pp. 447-459, 1994

LA TERMINOLOGIA ESTADÍSTICA I ALGUNS ASPECTES TERMINOLÒGICS DE LES CLASSIFICACIONS ESTADÍSTIQUES

GRISELDA MARTÍ*

Institut d'Estadística de Catalunya

Després d'una breu reflexió sobre els problemes del desenvolupament de la terminologia científica catalana i de la terminologia estadística, s'analitzen els aspectes lingüístics de les classificacions estadístiques. Es fa una breu referència a la metodologia que s'ha seguit, a alguns dels problemes que ha calgut solucionar, a les diferències i semblances que hi ha entre les diferents classificacions des del punt de vista terminològic, i al cabal lèxic que aquestes classificacions comporten.

Statistical terminology and some terminological aspects of statistical classifications.

1. LA TERMINOLOGIA CIENTÍFICA EN CATALÀ

Totes les branques del coneixement en general i en aquest cas concret les ciències, necessiten disposar de mots especialitzats que puguin designar les seves nocions. Aquests mots o unitats lexicals, usats amb significat unívoc en una ciència o una disciplina determinada per tal d'evitar una correspondència

* Griselda Martí. Institut d'Estadística de Catalunya Departament d'Economia i Finances. Via Laietana, 58. 08003 Barcelona

–Article rebut l'octubre de 1994.
–Acceptat el gener de 1995.

equívoca entre els conceptes i llur expressió, reben el nom de "termes" i com es desprèn de la seva definició són elements imprescindibles a l'hora de poder expressar o descriure qualsevol concepte de forma inequívoca.

En el món actual, en què els avenços tecnològics progressen a un ritme vertiginós, i per tant, a un ritme també vertiginós cal descriure i expressar nous conceptes, la terminologia "com a disciplina que s'ocupa de l'estudi i de la recopilació dels termes especialitzats" està esdevenint una eina cada vegada més indispensable per a la investigació científica. En un principi estava només en mans dels científics i no serà fins fa pocs anys que s'hi han incorporat els lingüistes.

Pel que fa a la terminologia científica en català cal no oblidar que ha hagut de superar algunes dificultats addicionals respecte a la terminologia científica d'altres llengües.

Com és lògic, els problemes que han afectat el normal desenvolupament del català com a llengua de comunicació en general, han incidit també en el camp de la terminologia, encara que potser amb més intensitat. No es tracta de lamentar res sinó de descriure una situació real que ha tingut unes conseqüències força greus: la diglòssia i tot allò de negatiu que aquest concepte comporta per a la llengua menys afavorida, ha estat (i encara és) una realitat a Catalunya. La ciència i la tecnologia durant molt anys, precisament aquells en què hi ha hagut un major desenvolupament científic, van utilitzar a Catalunya com a vehicle obligatori de comunicació el castellà i, per tant, una gran part de la terminologia s'ha elaborat en aquesta llengua.

Durant molts anys el català no es va utilitzar en l'ensenyament, ni en els mitjans de comunicació (escrits o orals), ni en l'Administració pública. Reduït a l'àmbit estrictament col·loquial els termes científics i tècnics eren necessàriament castellans, malgrat que algunes vegades existia el mot corresponent en català.

La situació ha començat a canviar i el català és cada vegada més una llengua d'ús normal en la ciència i la tecnologia. Això ha comportat unes necessitats terminològiques que han fet imprescindible treballar a marxes forçades per poder cobrir el desfasament que s'havia produït entre la llengua i les necessitats lingüístiques dels parlants. Bàsicament, en terminologia els problemes han estat els següents: nocions noves sense denominacions concretes, denominacions poc precises o estranyes als parlants, impossibilitat de disposar del temps necessari per tal que els termes catalans s'incorporin al cabal lèxic, desorientació sobre la fiabilitat de les fonts, interposició del castellà entre el català i la resta de llengües, etc.

Aquests mateixos problemes, amb més o menys intensitat, els podem trobar en totes aquelles comunitats lingüístiques on hi ha dues o més llengües en con-

tacte, és a dir, que utilitzen dues o més llengües de comunicació: una d'elles, la de més prestigi (social, cultural, econòmic, etc.) s'usarà per a les funcions formals o escrites i l'altra, per a les informals o orals. La diglòssia és un fenomen inseparable del bilingüisme perquè difícilment, en un mateix àmbit d'ús podrem utilitzar indiferentment totes dues llengües. El problema de la terminologia catalana és exactament el mateix que el de totes aquelles llengües minoritàries supeditades a altres llengües majoritàries i per tant caldrà mirar més enllà del nostre entorn immediat. Probablement, si observem com se n'han sortit altres comunitats lingüístiques amb situacions molt similars a la nostra, veurem que cal començar a treballar sense prejudicis de cap mena ni escatimar esforços.

L'anglès ha estat una altra llengua que ha influït notablement en la fixació de la terminologia no tan sols catalana sino de qualsevol altra llengua. La investigació científica duta a terme sobretot en països de llengua anglesa ha comportat la introducció massiva de terminologia anglòfona en totes les llengües, sobretot en camps tant estesos com la informàtica, les telecomunicacions o les ciències en general. Llengües culturalment prou fortes com el francès han arribat a promulgar normatives per defensar-se de la "invasió" lèxica anglesa.

Per al català l'anglès no és una amenaça de substitució lingüística immediata com el castellà i això té força influència a l'hora d'introduir certes denominacions d'origen anglès. No és que s'hagin acceptat sense restriccions però sí que s'han estudiat d'una forma menys políticament apassionada i més lingüísticament informada (mètode infinitament més productiu que l'invers).

Per altra banda, paral·lelament al lèxic s'ha hagut d'anar fixant una varietat estàndard que serveixi, entre altres usos, per a la comunicació científica i tècnica. Això que en principi pot semblar secundari per la poca repercussió que té en l'ortografia pot arribar a dificultar o fins i tot impossibilitar la comprensió d'un text. Per exemple, imaginem que en una anàlisi sobre uns resultats estadístics l'autor omple el text de figures retòriques, evidentment, l'expressió en un registre literari en dificultarà moltíssim la comprensió.

Com ja s'ha comentat, fins fa pocs anys la terminologia va córrer a càrrec dels especialistes (científics o tècnics) en cada camp del saber sense el concurs dels lingüistes, més ocupats en els principis que regeixen les llengües (reals o hipotètiques) que no pas a estudiar el llenguatge com a eina de comunicació. Poc a poc aquesta col·laboració ha anat esdevenint cada vegada més necessària i més en el cas del català on la investigació s'havia fet utilitzant pràcticament sempre el castellà i la poca que s'havia fet en català era pràcticament desconeguda. Per tant, els mateixos especialistes s'adonen que necessiten la col·laboració dels lingüistes i paral·lelament aquests no podran elaborar cap treball terminològic coherent sense l'assessorament conceptual de científics i tècnics.

2. LA TERMINOLOGIA ESTADÍSTICA

L'estadística com qualsevol altra ciència necessita expressar-se per mitjà d'un llenguatge precís i adequat a les seves necessitats. Tots el que s'ha tractat en els apartats anteriors és vàlid per a la terminologia estadística en català.

Cal assenyalar, abans d'entrar a comentar amb detall els aspectes terminològics d'aquesta ciència, que ens referirem sempre al vessant aplicat de l'estadística, és a dir, si l'estadística és "la ciència, mètode, tècniques, operació d'anàlisi matemàtica, que permeten estudiar numèricament amb el màxim de precisió els fenòmens col·lectius incompletament coneguts" no tan sols farem referència a la terminologia que utilitza com a ciència teòrica sinó també la que usa a l'hora d'estudiar numèricament els fenòmens i fins i tot la denominació d'aquests mateixos fenòmens (que de fet són mots del lèxic comú amb un significat restringit).

Per tant, veiem que les necessitats terminològiques de l'estadística aplicada són tan àmplies com ampli és l'univers que es vol estudiar. Si el lèxic comú o especialitzat ja s'ha anat fixant al llarg dels anys mitjançant l'ús social que en fan els parlants, lògicament quan l'estadística aplicada necessiti descriure els fenòmens que es denominen mitjançant aquest lèxic no haurà de fer més que utilitzar-lo. Però si aquest procés no s'ha donat pels motius que hem esmentat anteriorment, ens trobem amb el problema de la falta d'una terminologia específica que ens permeti anomenar els fenòmens que volem descriure. A més a més, hi haurà també mancances terminològiques per anomenar mètodes i operacions, i fins i tot en l'estudi més teòric de l'estadística.

Gràcies a la recent publicació del Diccionari d'estadística elaborat pel TERMCAT, la terminologia de l'estadística teòrica i aplicada queda pràcticament resolta, però queda pendent el problema dels termes que serveixen per denominar els fenòmens que descriu l'estadística i que poden formar part del lèxic comú o ser termes de camps de la ciència o la tècnica diferents de l'estadística. Seran aquests mots els que analitzarem a partir d'aquest moment.

Esmentarem l'exemple concret de l'enquesta industrial anual de productes, la versió bilingüe de la qual ha elaborat l'Institut d'Estadística de Catalunya en col·laboració amb l'Institut Nacional d'Estadística (INE). Aquesta operació de camp es basa en setanta-sis models diferents de qüestionaris dirigits a una mostra d'empreses, on es descriuen setanta-sis tipologies de productes distints; és a dir, cada qüestionari detalla els productes que a grans termes estan definits en els setanta-sis epígrafs del sector industrial de la Classificació d'activitats econòmiques.

La dificultat de traducció de tots els productes que s'esmenten al llarg d'aquests qüestionaris és òbvia. Primer per les grans diferències conceptuals que hi ha entre grups de productes i segon pel nivell de detall amb què aquests són descrits dintre de cada qüestionari. N'hi ha de referits a les empreses de productes químics (orgànics o inorgànics), altres a les empreses manufactureres de la pell, productes de vidre, adobs, pintures o productes alimentaris, però també cal tenir en compte que, per exemple, dintre del qüestionari de productes químics utilitzats en la indústria podem arribar a descriure fins a gairebé tres-cents compostos químics diferents.

Un altre exemple seria el del cens agrari on es fa servir un ventall amplíssim de terminologia del món agrícola i ramader: des dels diferents tipus d'automòbils utilitzats en les explotacions fins a tota la varietat de bestiar, passant per la seva destinació segons el tipus de consum que se'n faci, etc.

3. ASPECTES TERMINOLÒGICS DE LES CLASSIFICACIONS ESTADÍSTIQUES

Fins aquí s'ha pogut posar de relleu la importància de la terminologia en la ciència en general i, per tant, també en l'estadística, però en el camp on realment s'ha fet imprescindible és en el de les nomenclatures i classificacions que s'utilitzen per presentar les dades estadístiques de forma coherent, ordenada i comparable.

Les classificacions són, segons la seva definició general, la distribució d'un conjunt d'objectes en un cert nombre de classes, coordinades o subordinades, segons un criteri determinat. Les classificacions que s'utilitzen en estadística estableixen un conjunt jerarquitzat de categories homogènies que permeten classificar els elements que les componen. Aquests elements poden ser activitats econòmiques, ocupacions, béns o serveis. D'aquesta manera les dades estadístiques poden ser recollides i difoses de manera ordenada, amb més o menys nivells de desagregació o detall, o per a uns àmbits determinats.

La importància que té l'aplicació rigorosa d'una metodologia terminològica es pot intuir fàcilment. En primer lloc, la denominació de les categories es farà mitjançant termes, la qual cosa requerirà una precisió molt gran per tal que tant l'abast dels conceptes continents com el dels continguts quedi perfectament delimitat i fixat. En segon lloc és imprescindible evitar la sinonímia tant entre termes d'una mateixa categoria jeràrquica com de categories jeràrquiques diferents. I en tercer lloc, caldrà desfer també els casos d'homonímia mitjançant la

utilització de grups sintagmàtics (substantius amb complement de nom o adjectiu) o bé de sinònims per a un dels significats.

De moment, les classificacions a les quals ens referirem són traduccions de les que elabora o tradueix (de classificacions europees) l'INE per a la resta de l'Estat i que, per tant, ens arriben en castellà. Com que l'estadística ha de ser comparada i comparable en tot el territori espanyol, a Europa i arreu del món, no té sentit crear classificacions noves vàlides només per a Catalunya que no puguin comparar-se amb les d'altres àmbits geogràfics.

En un altre nivell ens trobem amb les dificultats pròpies de la manca d'experiència de terminologia elaborada en català, paral·lela al poc desenvolupament de la ciència i la tècnica en aquesta llengua. Això té una incidència molt gran a l'hora de localitzar bibliografia de consulta fiable (diccionaris, vocabularis, etc.) o de poder resoldre alguns problemes conceptuals amb els especialistes perquè també ells ignoren el terme en català.

En el cas de les classificacions, on els nivells de desagregació són extraordinàriament detallats, ens trobem amb el problema de la inexistència d'alguns conceptes en la nostra llengua o a la inversa, quan caldria distingir dos conceptes en català per motius de claredat o per ajustar els significats i el castellà només disposa d'un únic terme. Aquest problema es concreta quan s'han de traduir productes molt locals o ocupacions molt lligades a costums o tradicions d'una àrea concreta. Aquest és el cas d'algunes espècies marines comercials o les ocupacions relacionades amb el món de la tauromàquia.

Hi ha, per altra banda, un gran nombre d'oficis o productes tradicionals amb noms ben fixats en català i que tothom coneixia prou bé fa uns anys, però el fet que el català no hagi estat un vehicle d'expressió escrita en els mitjans de comunicació públics o en l'ensenyament, ha comportat un progressiu oblit d'aquests termes fins al punt que alguns d'ells s'han perdut o són molt difícils de recuperar. Altres ocupacions o productes nous s'han introduït amb el terme castellà directament i a vegades és difícil poder-los traduir o adaptar.

En algun moment de la traducció de les classificacions ha estat necessari recórrer a fonts d'informació no normatives però que ens han proporcionat una primera solució al problema terminològic, que ha quedat pendent d'un posterior seguiment més acurat. Algunes vegades s'ha hagut d'optar per una variant dialectal o per un estrangerisme.

El procés de validació de les traduccions es desenvolupa de la manera següent. A l'Institut d'Estadística es fa la primera traducció i es comprova i corregeix tantes vegades com sigui necessari, es consulten les fonts adequades i els especialistes corresponents. Amb aquest primer esborrany es fa una revisió juntament amb la

Secció de Normalització Lingüística del Departament d'Economia i Finances. Un cop introduïdes les modificacions i aïllats els problemes pendents de solució, es tramet el text al TERMCAT per tal que el validi definitivament, juntament amb la traducció se li fa arribar un resum de la metodologia terminològica que s'ha seguit, una llistá completa de la bibliografia consultada, una còpia en castellà de l'original de l'INE i tots aquells materials que puguin ajudar a comprendre el text. Posteriorment el TERMCAT ens tramet la solució dels dubtes que havien quedat pendents i les esmenes que ha considerat oportunes. La difusió final d'aquestes traduccions pot ser en forma de publicacions, de decrets al Diari oficial de la Generalitat de Catalunya (DOGC), epígrafs de taules estadístiques en publicacions editades per l'Institut, descripcions de codis en els fulls padronals i censals, etc.

Actualment podem disposar de la traducció de la Classificació nacional d'activitats econòmiques de l'any 1974, la Classificació nacional d'ocupacions de 1979, la Classificació d'activitats econòmiques de 1993, la Classificació d'ocupacions de 1994 i la Classificació de béns i serveis de 1978. Algunes d'elles estan ja disponibles en forma de publicació, altres seran editades a més a més com a decrets en el DOGC i altres es troben en fase d'una última revisió per donar-les ja per definitives.

Paral·lelament a la traducció dels epígrafs de les classificacions, es traueixen també les anomenades "Notes explicatives" de cada una d'elles. Aquestes notes especifiquen quins són els conceptes concrets que s'inclouen dintre de la descripció genèrica de cada epígraf. Per tant, malgrat que aquesta part no tingui la rellevància del text de la classificació, el treball terminològic que s'hi ha de fer és equivalent o fins i tot més extens.

Les dificultats específiques que han condicionat la traducció de cada classificació les veurem per separat a continuació.

3.1. Classificacions d'activitats econòmiques

Descriuen de forma jeràrquica les diferents activitats econòmiques, és a dir, "tenen per objecte l'agrupació i la classificació de les unitats productores, segons l'activitat que exerceixin, per a l'elaboració d'estadístiques relatives als fenòmens lligats al seu funcionament econòmic: producció, utilització de factors productius, rendiment, etc.".

Ja s'han traduït les dues editades fins ara per l'INE: la de 1974 i la de 1993. La primera s'ha difòs en forma de publicació i la segona es farà oficial amb la publicació d'un decret en el DOGC a principi del 1995.

L'estructura de la classificació del 1974 consta de quatre nivells jeràrquics numerats amb un, dos, tres o quatre dígits a mesura que van detallant les activitats corresponents. El nivell d'un dígit s'anomena "divisió"; el segon nivell, amb dos dígits, és l'"agrupació"; amb tres dígits tenim el "grup", que identifica el tercer nivell; i finalment, amb quatre dígits i amb el màxim nivell de desagregació, hi ha el "subgrup".

L'estructura de la del 1993 és una mica més complicada però és essencialment la mateixa. El primer nivell amb un codi alfabètic d'un dígit s'anomena "secció"; a continuació ve un nivell intermedi anomenat "subsecció" amb codi alfabètic de dos dígits; el segon nivell, amb codi numèric de dos dígits, és la "divisió"; el tercer nivell, amb tres dígits, s'anomena "grup"; la "classe" és el quart nivell amb quatre dígits; i finalment, tenim el cinquè nivell o "subclasse" amb un codi numèric de cinc dígits.

El treball terminològic que s'ha dut a terme en les dues classificacions ha estat molt diferent malgrat que es tractés d'una temàtica molt similar. En la de 1974, la traducció de la qual va començar l'any 1990, es va partir de zero. L'únic material de què disposàvem eren els diccionaris de l'Enciclopèdia i alguns vocabularis especialitzats que van ajudar força a resoldre la traducció dels tres primers nivells, però el quart nivell, el de subgrup, pel seu grau de detall va ser el més complicat de resoldre perquè s'hi havia de diferenciar la producció o fabricació de diverses varietats d'un mateix producte.

També va ser necessari unificar certes expressions que es repetien al llarg de la classificació castellana de maneres diferents tot i tenir el mateix significat. Es va intentar donar uniformitat a les construccions amb preposicions i articles, i estructurar de forma coherent els signes de puntuació (comes, punts i comes, punts, etc.).

Les fases de revisió i correcció d'aquest text van ser moltes i molt laborioses perquè era la primera vegada que ens enfrontàvem amb un document d'un abast terminològic tan gran. Aquesta classificació conté 863 epígrafs diferents i a cada un hi pot aparèixer una mitjana de dos termes diferents, per tant, cal comptar uns 1.600 termes referits a deu àmbits diferents d'acitivitats econòmiques.

En la classificació del 93 la situació de sortida era molt diferent perquè comptàvem amb una experiència anterior. De fet, més que per resoldre problemes terminològics el que més es va poder aprofitar eren els criteris d'unificació d'estructures sintàctiques, signes puntuació, construccions fixes, etc. Aquesta classificació conté 1.595 epígrafs i restant-ne 503 de repetits donen 1.092 epígrafs diferents. Per tant, tindrem uns 2.000 termes.

3.2. Classificacions d'ocupacions

En aquest cas descriuen de forma jeràrquica les ocupacions que es poden desenvolupar en el mercat laboral.

Estan ja traduïdes les dues que fins ara ha editat l'INE: la de 1979 i la de 1994. La Classificació d'ocupacions de l'any 79 s'editarà pròximament en forma de publicació i la del 94 a més a més es publicarà en el DOGC en forma de decret a principi del 1995.

Les estructures són les següents. La de l'any 79 està dividida en quatre nivells jeràrquics. El nivell més general, anomenat "gran grup," identificat per un dígit numèric; el segon nivell, o "subgrup", identificat per dos dígits; el tercer nivell, o "grup primari", amb tres dígits; i finalment, l'ocupació en sentit estricte amb quatre dígits.

La del 1994 consta de cinc nivells. El "gran grup" amb un dígit numèric; el "grup principal", o nivell intermedi, amb un dígit alfabètic; el "subgrup principal", com a segon nivell, amb dos dígits numèrics; el tercer nivell anomenat "subgrup" amb tres dígits numèrics; i finalment, el "grup primari" amb un codi numèric de quatre dígits.

La feina de traducció de les dues classificacions d'ocupacions editades per l'INE, la del 79 i la del 94, ha estat molt diferent de les altres dues. D'entrada hem pogut disposar d'un material de consulta específic molt important: el Diccionari de professions editat per la Direcció General d'Ocupació del Departament de Treball de la Generalitat de Catalunya amb la col·laboració i l'assessorament del TERMCAT.

Les fases de treball per a la classificacions d'ocupacions van ser les mateixes que s'han descrit per a les anteriors d'activitats econòmiques però s'hi va afegir una fase més: el treball de confrontació entre la nostra traducció i la de la Classificació internacional uniforme d'ocupacions de l'any 1988, elaborada pel Departament de Treball de la Generalitat de Catalunya.

Malgrat l'existència d'uns materials de consulta tan vàlids com els que hem esmentat, les dificultats terminològiques van ser molt importants. Per una banda hi havia la impossibilitat de traduir algunes professions inexistents en el domini de parla catalana, per l'altra, algunes que hi són usuals no consten a la classificació i altres no s'han anomenat mai amb termes catalans, i finalment, la gran varietat de sufixos que comporten el sentit de "persona dedicada a..." o "que desenvolupa una determinada professió".

En aquestes dues classificacions hi ha la dificultat afegida (inexistent a les classificacions d'activitats econòmiques i de béns i serveis que veurem a continuació) de la possibilitat de doble gènere en el terme que descriu l'ocupació. Aquest punt es va estudiar molt a fons i posteriorment va ser analitzat i valorat amb el TERMCAT i el Departament de Treball. De moment aquestes dues classificacions no tenen la flexió de gènere femení en les ocupacions que s'hi descriuen. El TERMCAT està treballant en un projecte d'aquest tipus.

En la Classificació d'ocupacions del 79 hi ha 1.970 epígrafs i uns 4.000 termes diferents. En la del 94 hi ha 793 epígrafs i, per tant, uns 1.600 termes.

3.3. Classificacions de béns i serveis

De moment només s'ha traduït la que va ser editada per l'INE el 1978. S'hi descriuen tota mena de productes, és a dir, béns, susceptibles de ser fabricats, transformats, cultivats, etc., en un espai econòmic, i tots aquells serveis que poden ser prestats a una col·lectivitat. A causa de la gran diversitat de sectors productius a què fa referència i al gran nivell de detall a què s'arriba, aquesta classificació esdevé un gran vocabulari extensíssim.

Actualment s'està acabant la revisió de la primera traducció i es preveu tenir-la llesta a final del 95.

Malgrat el gran volum de terminologia que conté l'estructura és força senzilla ja que consta només de quatre nivells jeràrquics. El primer amb quatre dígits numèrics, el segon amb cinc, el tercer amb sis i l'últim amb set.

S'hi ha hagut de fer una gran labor d'unificació de criteris perquè la gran extensió d'aquesta classificació provoca canvis injustificats entre la primera part i l'última. Per altra banda, el nivell de detall és tan gran que molt sovint els productes molt locals que s'hi descriuen (per exemple alguns embotits) no tenen equivalent en català.

La Classificació de béns i serveis del 78 consta de 12.500 epígrafs, la qual cosa representa (amb una mitjana de dos termes per epígraf) uns 25.000 termes.

4. CONCLUSIONS

Malgrat que el to general de l'article s'hagi pogut centrar en l'exposició dels problemes amb què es troba la terminologia catalana i els que hi ha ha-

gut a l'hora de fer la traducció de les classificacions, és important ressenyar que l'objectiu principal de plena normalització lingüística en l'àmbit de la terminologia científica i tècnica s'està aconseguint de forma gradual també en l'estadística.

La traducció de les classificacions que s'utilitzen per presentar les dades estadístiques és una petita contribució a aquest objectiu malgrat el volum de material que representa. Aquesta contribució tant pot servir per ampliar la terminologia estadística com el lèxic comú gràcies a les característiques d'aquestes classificacions.

Com ja s'ha dit, posar al dia una llengua que durant tants anys no s'ha utilitzat com a instrument de comunicació formal és un treball titànic i contrarellotge i, per tant, amb moltes possibilitats d'error, de dubte i de rectificació. El més important, i això no s'ha de perdre mai de vista, és que el català arribi a ser una llengua útil per expressar tot allò que de vell o de nou hi ha al món.

BIBLIOGRAFIA

[1] **Cabré, M.T.** (1992). *La terminologia. La teoria, els mètodes, les aplicacions.* Barcelona: Empúries.

[2] **Enciclopèdia Catalana** (1993). *Diccionari de la llengua catalana.* Barcelona.

[3] **Il·lustre Col·legi d'Advocats de Barcelona**. (1989). *Diccionari de les professions. Barcelona: Generalitat de Catalunya.* Direcció General d'Ocupació.

[4] **Institut d'Estadística de Catalunya** (1993). *Classificació d'activitats econòmiques. Adaptació normalitzada de la CNAE-74.* Barcelona, Institut d'Estadística de Catalunya, 1993.

[5] **Riera, C.** (1994). *El llenguatge científic català.* Barcelona: Barcanova.

ENGLISH SUMMARY:

STATISTICAL TERMINOLOGY AND SOME TERMINOLOGICAL ASPECTS OF STATISTICAL CLASSIFICATIONS

Griselda Martí

1. SCIENTIFIC TERMINOLOGY IN CATALAN

The Catalan language has not been able to develop in a normal fashion in the field of scientific and technical terminology, because in recent years the language of communication in science has been Castilian. Considerable efforts are currently being made to bring Catalan up to date in the field of terminology. With respect to the aim of achieving the complete standardisation of the language, account should also be taken of the development of other minority languages which are in contact with majority languages.

2. STATISTICAL TERMINOLOGY

Problems concerning terminology arise not only in the field of statistical theory, but also in the field of applied statistics and with relation to statistical phenomena which it is wished to study. It is important to emphasise the great lexical diversity of these phenomena.

3. TERMINOLOGICAL ASPECTS OF STATISTICAL CLASSIFICATIONS

An analysis is made of the process of translating the classifications which assist in the collection, ordering and dissemination of statistical data. Reference is made to common aspects which affect the three most important types of classifications: economic activities, occupations and goods and services.

3.1. Classifications of economic activities

Both classifications which the INE has published to date, those of 1974 and 1993, have been translated. Special problems arise in each of these with respect to terminology.

3.2. Occupational classifications

Both classifications which the INE has published to date, those of 1979 and 1994, have been translated. Among other considerations, mention should be made of the difficulties involved in the translation of some occupations which are linked with a very specific culture. Gender inflections are not present in either of the two classifications.

3.3. Classifications of goods and services

The 1978 classification has been translated. It is the most extensive with respect to the volume of terms and therefore the most complicated from a terminological point of view.

4. CONCLUSIONS

The importance of the development of terminology to the process of standardising the Catalan language is highlighted, as well as the small contribution made to this process by the lexical volume of statistical classifications.

SECCIÓ DOCENT I PROBLEMES

La introducció de la nova "SECCIÓ DOCENT I PROBLEMES" a la revista QÜESTIIÓ es fa amb l'objectiu d'incloure una secció on es publiquen articles de caire docent, difícilment publicables en revistes de recerca. Alhora es continua amb l'antiga secció de problemes. A cada número de QÜESTIIÓ s'inclourà d'un a tres problemes i les solucions es donaran en el número següent.

Els lectors poden, si ho volen, proposar problemes amb les solucions pertinents i enviar-los a QÜESTIIÓ, que farà una selecció i en publicarà els més adequats, fent la corresponent referència a l'autor.

També seran ben rebudes solucions alternatives a les propostes fetes per l'autor dels problemes; l'editorial es reservarà, però, el dret a publicar-les.

PROBLEMES PROPOSATS

PROBLEMA N° 55

Sea un juego bipersonal de suma nula con matriz de ganancias:

		Jugador B	
Jugador A		0	a
		$-a$	0

, $\quad a > 0$

Obtener las estrategias óptimas de ambos jugadores y el valor del juego, aplicando el método gráfico.

Ramón Alonso Sanz
E.T.S.I. Agrónomos
Madrid

PROBLEMA N° 56

Sea un juego bipersonal de suma nula con matriz de ganancias:

		Jugador B	
Jugador A		a	b
		α	α

Obtener las estrategias óptimas de ambos jugadores y el valor del juego.

Ramón Alonso Sanz
E.T.S.I. Agrónomos
Madrid

SOLUCIONS ALS PROBLEMES PROPOSATS AL VOLUM 18. N° 2

PROBLEMA N° 51

Admitamos que una población finita de tamaño N, estratificada en L clases o estratos, es a su vez una muestra aleatoria simple, s, de una población infinita madre con varianza $\overline{\sigma}^2 < \infty$; entonces si $P_h = N_h/N$ es el peso relativo del estrato h en la muestra s, tenemos que (Deming, 1960)

$$\sigma^2 = \sum_{h=1}^{L} P_h \sigma_h^2 + \sum_{h=1}^{L} P_h \left(\mu_h - \mu\right)^2 ,$$

siendo μ y σ^2 la media y la varianza de la muestra s, y μ_h y σ_h^2 la media y la varianza del estrato h en la muestra s.

Para s fijada,

$$\begin{aligned} E(\sigma^2) &= \frac{N-1}{N}\overline{\sigma}^2, \\ E\left(\sum_{h=1}^{L} P_h \sigma_h^2 | s\right) &= \sum_{h=1}^{L} \frac{N_h - 1}{N} \overline{\sigma}_h^2 \end{aligned}$$

y

$$\begin{aligned} E\left[\sum_{h=1}^{L} P_h \left(\mu_h - \mu\right)^2 | s\right] &= \sum_{h=1}^{L} P_h E\left[\left(\mu_h - \overline{\mu}_h\right)^2 | s\right] + \sum_{h=1}^{L} P_h \left(\overline{\mu}_h - \overline{\mu}\right)^2 - \\ &- E\left[\left(\mu - \overline{\mu}\right)^2 | s\right] = \sum_{h=1}^{L} \left[\frac{\overline{\sigma}_h^2}{N} + P_h \left(\overline{\mu}_h - \overline{\mu}\right)^2\right] - \frac{\overline{\sigma}^2}{N}. \end{aligned}$$

Así

$$E(\sigma^2) = E\left\{E\left[\sum_{h=1}^{L} P_h \sigma_h^2 + \sum_{h=1}^{L} P_h \left(\mu_h - \mu\right)^2 | s\right]\right\},$$

o sustituyendo

$$\overline{\sigma}^2 = E\left[\sum_{h=1}^{L} P_h \overline{\sigma}_h^2 + \sum_{h=1}^{L} P_h \left(\overline{\mu}_h - \overline{\mu}\right)^2\right] = \sum_{h=1}^{L} \overline{P}_h \overline{\sigma}_h^2 + \sum_{h=1}^{L} \overline{P}_h \left(\overline{\mu}_h - \overline{\mu}\right)^2 ,$$

siendo $N_h = NP_h$ una variable aleatoria binomial de parámetros N y $\overline{P}_h$, donde $\overline{P}_h, \overline{\sigma}_h^2$, $\overline{\mu}_h, \overline{\mu}$ y $\overline{\sigma}^2$ son los parámetros asociados a la población madre estratificada.

REFERENCIA

[1] **Deming, W.E.** (1960). *Sample Design in Business Research.* Wiley. Nueva York.

M. Ruiz Espejo
Universidad Complutense de Madrid

PROBLEMA N° 52

Siguiendo la exposición más reciente de Sánchez-Crespo (1980), la varianza del estimador de la media poblacional de Hansen y Hurwitz es:

$$V(\widehat{\overline{X}}) = (1-f)\frac{S^2}{n} + \frac{1}{fN^2}\left(\frac{1}{f_{21}} - 1\right) N_2 S_2^2$$

donde f, n, N, f_{21} y N_2 son conocidos. Bastará estimar insesgadamente S^2 y S_2^2.

En efecto, la cuasivarianza poblacional S^2, puede ser estimada por la cuasivarianza muestral s^2 de tamaño $n_{21} = f_{21}n_2$ compuesta por las n_{21} primeras unidades de la muestra ordenada. También, la cuasivarianza muestral del estrato de no respuesta S_2^2, puede ser estimada insesgadamente por la cuasivarianza muestral s_2^2 de tamaño n_{21} compuesta por las n_{21} primeras unidades de la muestra que sean de no respuesta en un primer intento. De aquí es inmediato proponer un estimador insesgado de $V(\widehat{\overline{X}})$, frente al problema de la no respuesta en el método de Hansen y Hurwitz (1946).

En líneas similares puede estimarse la varianza del estimador propuesto por Ruiz (1988) para la estimación insesgada con observaciones erradas y no respuesta. Observar que $V(\widehat{\overline{X}}')$ de Ruiz (1988, p. 79) es exactamente $V(\widehat{\overline{X}}'|n)$, y que el llamado valor esperado de la varianza de pág. 80, es propiamente la varianza

$$V(\widehat{\overline{X}}') = E\left[V(\widehat{\overline{X}}'|n)\right],$$

puesto que $V\left[E(\widehat{\overline{X}}'|n)\right] = 0$ para todo $n \geq 1$. Entonces, $S^2_{x'_N}$ puede ser estimada insesgadamente por la cuasivarianza muestral $s^2_{x'_{n_{11}+n_{21}}}$ de las $n_{11}+n_{21}$ primeras unidades de la muestra.

También $S^2_{e_{N_1}}$ puede ser estimada insesgadamente por la cuasivarianza de errores $s^2_{e_{n_{11}}}$ de las n_{11} primeras unidades de respuesta en el primer intento.

Por último, $S^2_{x'_{N_2}}$ puede ser estimada insesgadamente por la cuasivarianza muestral $s^2_{x'_{n_{21}}}$ de las n_{21} primeras unidades de la muestra en la segunda fase de la variable x' entre las n_2 que no respondieron en la primera fase.

De aquí, es inmediato construir un estimador insesgado de $V(\widehat{\overline{X}}')$ frente a los problemas de errores de medida u observaciones erradas y no respuesta.

Nota. Aunque N_2 no sea conocido, es posible estimarlo insesgadamente por Nn_2/n. De modo similar N_1N_2 se puede estimar insesgadamente por el estadístico $N^2n_1n_2/[n(n-1)]$.

REFERENCIA

[1] **Hansen, M.H.** y **Hurwitz, W.N.** (1946). "The problem of non response in sample surveys". *J. Amer. Statist. Assoc.*, **41**, 517–529.

[2] **Ruiz, M.** (1988). "Estimación insesgada con observaciones erradas y no respuesta". *Trabajos Estadíst.*, **3 (1)**, 71–80.

[3] **Sánchez-Crespo, J.L.** (1980). *Curso Intensivo de Muestreo en Poblaciones Finitas.* (2ª edición). INE. Madrid.

M. Ruiz Espejo

Universidad Complutense de Madrid

PROBLEMA N° 53

I) estrategias óptimas y valor del juego[1]

Jugador A:

max v

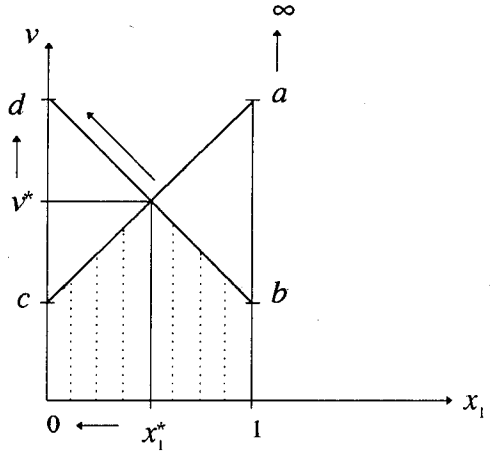
$$\begin{aligned} ax_1 + cx_2 &= ax_1 + c(1-x_1) = (a-c)x_1 + c \geq v \\ bx_1 + dx_2 &= bx_1 + d(1-x_1) = (b-d)x_1 + d \geq v \\ 0 \leq x_1 \leq 1, &\qquad 0 \leq x_2 \leq 1 \end{aligned}$$

$$(a-c)x_1 + c = (b-d)x_1 + d \Rightarrow (a-c-(b-d))x_1 = d-c \Rightarrow$$

$$\Rightarrow \boxed{\begin{aligned} x_1^* &= \frac{d-c}{(a-c)+(d-b)} \\ x_2^* &= \frac{a-b}{(a-c)+(d-b)} \end{aligned}} \Rightarrow$$

$$v^* = (a-c)x_1^* + c = (a-c)\frac{d-c}{(a-c)+(d-b)} + c \Rightarrow$$

$$\Rightarrow \boxed{v^* = \frac{ad-bc}{(a-c)+(d-b)}}$$



[1] Adoptando el convenio habitual de que A juega *a ganar* y B *a no perder*.

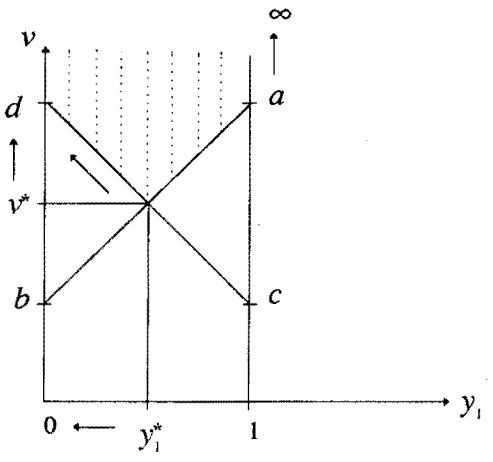
Jugador B:

$\min v$

$$\begin{array}{rcl} ay_1 + by_2 & = & ay_1 + b(1-y_1) = (a-b)y_1 + b \geq v \\ cy_1 + dy_2 & = & cy_1 + d(1-y_1) = (c-d)y_1 + d \geq v \\ 0 \leq y_1 \leq 1, & & 0 \leq y_2 \leq 1 \end{array}$$

$$(a-b)y_1 + b = (c-d)y_1 + d \Rightarrow (a-b-(c-d))y_1 = d-b \Rightarrow$$

$$\Rightarrow \boxed{\begin{array}{l} y_1^* = \dfrac{d-b}{(a-c)+(d-b)} \\ y_2^* = \dfrac{a-c}{(a-c)+(d-b)} \end{array}}$$



Las condiciones supuestas en este problema impiden la anulación del denominador (común) de las expresiones de las componentes de ambas estrategias óptimas y del valor del juego. En efecto, por ser los valores de la matriz de ganancias distintos entre sí, las diferencias $a-c$ y $d-b$ no se anulan. Por otra parte, al no presentarse dominio entre líneas, dichas diferencias no pueden tomar valores opuestos, la otra posible causa de anulación del denominador. Así, si $a > c \Rightarrow a - c > 0$ (como en la figura) debe ser $d > b \Rightarrow d - b > 0$.

II) ¿Qué fila (columna) deberá elegirse con mayor probabilidad?

Jugador A

$$x_1^* \left\{ \begin{array}{c} \geq \\ \leq \end{array} \right\} x_2^* \Leftrightarrow a - b \left\{ \begin{array}{c} \geq \\ \leq \end{array} \right\} d - c,$$

es decir, la fila más elegida será la más *equilibrada:* aquella en la que la diferencia de sus elementos es menor.

La preferencia por la fila con valores más próximos indica una tendencia a la *seguridad*, en el sentido de evitar la alternativa más descompensada.

$$x_1^* = x_2^* \Leftrightarrow a - b = d - c$$

Jugador B

$$y_1^* \left\{ \begin{array}{c} \geq \\ \leq \end{array} \right\} y_2^* \Leftrightarrow a - c \left\{ \begin{array}{c} \geq \\ \leq \end{array} \right\} d - b,$$

es decir, de forma equivalente a lo que ocurre en la elección del jugador A, la columna más elegida será la más equilibrada.

$$y_1^* = y_2^* \Leftrightarrow d - b = a - c$$

III)

Jugador A

$$\lim_{a \to \infty} \boldsymbol{X}^* = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \lim_{a \to \infty} v^* = d$$

Este resultado contradice la *intuición* que llevaría a pensar que el jugador A tenderá a elegir la fila donde se encuentra el valor que crece hasta infinito, la primera en este ejemplo. Hay que tener en cuenta que, en el límite, B nunca elegirá la primera columna (ver Jugador B), por lo que en él han de compararse sólo los valores de la segunda columna. Pero si se mantiene la condición de ausencia de dominio, ha de mantenerse la condición $d > b$, por lo que la primera fila queda descartada.

Jugador B

$$\lim_{a \to \infty} \boldsymbol{Y}^* = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

El Jugador B nunca elegirá la primera columna, en la que se encuentra el valor que crece indefinidamente.

IV)

IV.1 Haciendo $a = n, b = n+2, c = n+3, d = n+1$ en las fórmulas generales:

$$x_1^* = \frac{1}{2}, \quad y_1^* = \frac{1}{4}, \quad v^* = n + \frac{3}{2}$$

IV.2 Haciendo $a = n, b = 3n, c = 4n, d = 2n$ en las fórmulas generales:

$$x_1^* = \frac{1}{2}, \quad y_1^* = \frac{1}{4}, \quad v^* = \frac{5}{2}n$$

Los juegos IV.1 y IV.2 presentan por tanto las mismas estrategias óptimas independientes de n:

$$\boxed{x_1^* = x_2^* = \frac{1}{2}}, \quad \boxed{y_1^* = \frac{1}{4}, \quad y_2^* = \frac{3}{4}}.$$

Los juegos difieren en su valor, pero en ambos casos éste es la media aritmética de los elementos de la matriz de ganancias.

El Problema 1 puede calificarse como el *normal*, en la medida en que los elementos de la matriz de ganancias no presentan ninguna característica especial. En el siguiente problema se mantiene la propiedad de no dominio, pero se postula la igualdad de los elementos de la diagonal principal. Un problema semejante al que se describe seguidamente se plantearía suponiendo la igualdad de los elementos de la diagonal no principal.

Ramón Alonso Sanz
E.T.S.I. Agrónomos
Madrid

PROBLEMA N° 54

I) estrategias óptimas y valor del juego

Jugador A:

$\max v$

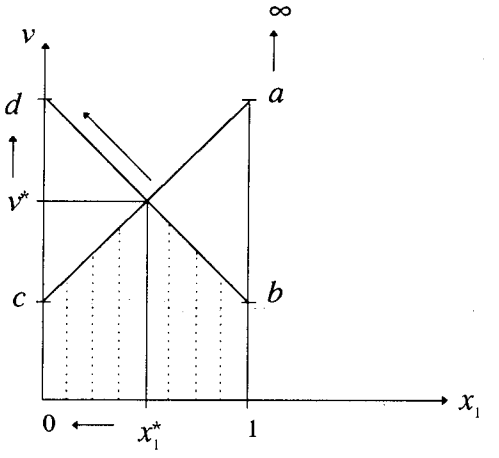
$$\begin{aligned} ax_1 + cx_2 &= ax_1 + c(1-x_1) = (a-c)x_1 + c \geq v \\ bx_1 + ax_2 &= bx_1 + a(1-x_1) = (b-a)x_1 + a \geq v \\ 0 \leq x_1 \leq 1, &\qquad 0 \leq x_2 \leq 1 \end{aligned}$$

$$(a-c)x_1 + c = (b-a)x_1 + a \Rightarrow (a-c-(b-a))x_1 = a-c \Rightarrow$$

$$\Rightarrow \boxed{\begin{aligned} x_1^* &= \frac{a-c}{2a-c-b} \\ x_2^* &= \frac{a-b}{2a-c-b} \end{aligned}} \Rightarrow$$

$$v^* = (a-c)x_1^* + c = (a-c)\frac{a-c}{2a-c-b} + c \Rightarrow$$

$$\Rightarrow \boxed{v^* = \frac{a^2-bc}{2a-b-c}}$$



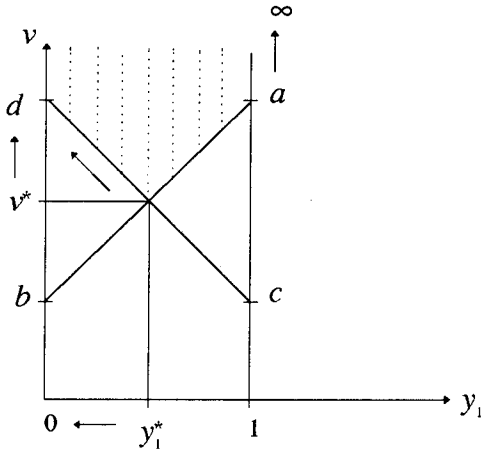
Jugador B:

$\min v$

$$\begin{array}{rcl} ay_1 + by_2 &=& ay_1 + b(1-y_1) = (a-b)y_1 + b \geq v \\ cy_1 + ay_2 &=& cy_1 + a(1-y_1) = (c-a)y_1 + a \geq v \\ 0 \leq y_1 \leq 1, & & 0 \leq y_2 \leq 1 \end{array}$$

$$(a-b)y_1 + b = (c-a)y_1 + a \Rightarrow (a-b-(c-a))y_1 = a-b \Rightarrow$$

$$\Rightarrow \boxed{\begin{array}{l} y_1^* = \dfrac{a-b}{2a-c-b} \\ y_2^* = \dfrac{a-c}{2a-c-b} \end{array}}$$



Las condiciones supuestas en este problema impiden la anulación del denominador de las expresiones de las componentes de X^*, Y^* y de v^*. Además, dichas expresiones se pueden obtener de las del Problema 1, haciendo $d = a$.

En el caso particular $b = c$:

$$\boxed{x_1^* = x_2^* = y_1^* = y_2^* = \frac{1}{2}} \qquad \boxed{v^* = \frac{b+a}{2}}$$

Es decir, en este supuesto se presenta una perfecta simetría en las estrategias y el valor del juego es la media aritmética de los elementos de la matriz de ganancias.

II)

Jugador A

$$x_1^* \left\{ \begin{matrix} \geq \\ \leq \end{matrix} \right\} x_2^* \Leftrightarrow b \left\{ \begin{matrix} \geq \\ \leq \end{matrix} \right\} c,$$

es decir, en contra de la *intuición*, pero manteniendo en criterio de preferencia por la fila más equilibrada (*más segura*), el jugador A deberá elegir más frecuentemente la fila donde se encuentre el menor de los valores b y c.

Jugador B

$$y_1^* \left\{ \begin{matrix} \geq \\ \leq \end{matrix} \right\} y_2^* \Leftrightarrow c \left\{ \begin{matrix} \geq \\ \leq \end{matrix} \right\} b.$$

El jugador B elige más frecuentemente la columna donde se encuentre el menor de los valores c y b, manteniéndose el principio de búsqueda de *seguridad*.

III)

Si b crece x_1^* decrece $\Rightarrow$ x_2^* crece

(en contra de la *intuición*) $\lim_{b\to\infty} X = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

Si b crece y_1^* crece $\Rightarrow$ y_2^* decrece (*esperable*). $\lim_{b\to\infty} Y = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$$\lim_{b\to\infty} v^* = c$$

IV) $\lim_{a\to\infty} X = \lim_{a\to\infty} Y = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$ $\lim_{a\to\infty} v^* = \infty$

Ramón Alonso Sanz
E.T.S.I. Agrónomos
Madrid

NOVETATS DE SOFTWARE

La introducció de la nova secció de "NOVETATS DE SOFTWARE" a la revista QÜESTIIÓ es fa amb la finalitat de promoure l'intercanvi d'informació relacionada amb programes d'ordinador disponibles, destinats a la implementació de metodologia estadística, d'informàtica o d'investigació operativa.

A causa de l'important creixement que ha experimentat darrerament la utilització dels ordinadors a totes les àrees científiques i tècniques i, a les esmentades més amunt, en particular, hi ha un bon nombre d'investigadors que han desenvolupat un software propi, l'existència del qual és desconeguda, de vegades, per a molts lectors que el podrien aprofitar. Per això, creiem que és convenient i útil fer-lo conèixer mitjançant aquesta revista, amb el benentès que només actuaria com a mitjà de difusió.

Per tal d'uniformitzar la descripció del software, adjuntem una butlleta que ha de ser omplenada i tramesa a l'editorial de QÜESTIIÓ.

Amb tota certesa, la vostra col·laboració serà d'utilitat per a molts lectors als qui facilitarà el treball i que, alhora, podran ajudar els autors del programes suggerint-los possibles millores.

Nom del programa: CALIBRA

Area/àrees d'aplicació (Estadística, Sistemes, etc.): Donde se requiera algún método de calibración multivariada, lo que es frecuente en la química, analítica, en la óptica, en la biología, etc.

Descripció del software:

- Llenguatge: Pascal v 6.0
- Ordinador/s: 80×86
- Sistema operatiu: MS DOS

Està disponible en els suports següents:

Floppy disk/diskette: Mida: 5 1/4 o 3 1/2

Distribuït per:

Departamento de Matemática Aplicada
Facultad Matemática — Computación
Universidad de La Habana
Ciudad de la Habana. Cuba

Configuració mínima de hardware requerida:

Microprocesador 80×86 con 1 mega de memoria RAM.

Descripció del que fa l'esmentat software: (200 paraules aproximadament).

El sistema dispone facilidades para la calibración multivariada mediante los métodos de Regresión con Componentes Principales, Mínimo Cuadrática y Mínimo Cuadrática Parcial; siempre que el modelo describa el fenómeno linealmente o de forma potencial; pero que se pueda linealizar mediante transformaciones logarítmicas.

Se adiciona la opción de considerar interacciones binarias entre las variables independientes aunque no pueden ser más de 3.

Brinda posibilidades de aplicar logaritmos a los datos y si se desea, seleccionar transformaciones de los datos a través de un centrado o centrado y escalado con la desviación típica.

Efecturando el cálculo para obtener los coeficientes de los estimadores, según el método seleccionado, se puede realizar el análisis de la precisión del modelo, llevándose a efecto esto imprimiendo diversos criterios de errores, si fue explícitamente solicitado.

Se puede realizar predicciones utilizando los resultados que se obtuvieron durante el proceso de calibración; en la misma sesión de trabajo o en alguna otra efectuada en el sistema.

El sistema puede realizar la captura de los datos de forma interactiva, importándolos de un fichero ASCII o sobre un fichero creado dentro del sistema. Además permite seleccionar la impresión de los cálculos estadísticos efectuados y de resultados intermedios que les sean necesarios al usuario. En el proceso de trabajo se requiere de 2 Kbytes libres en disco, para la creación de ficheros temporales y el modelo puede constar de a lo más 100 observaciones y 10 variables independientes y dependientes.

Possibles usuaris: Los que requieran modelos de calibración multivariada.

Camps d'interès: Aplicaciones que requieran la utilización de métodos de calibración multivariada para la solución de problemas.

Nom de l'autor/s:

Dra. Gladys Linares Fleites
M.C. María Victoria Mederos Bru
Lic. Carmen T. Fernández Montoto

Institució:

Universidad de La Habana
San Lázaro y L
Vedado
Ciudad de La Habana
Cuba
Correo electrónico: glinares%comuh@ceniai.cu:

Corrigendum to Nikulin M.S., Voinov V.G.

"Unbiased estimators of multivariate discrete distributions and chi-square goodness-of-fit test"

(QÜESTIIÓ, Vol. 17,3 pp. 301–326, 1993)

Formula (5) of the Table of MVUE's for functions $u(p)$ of parameter p of the trinomial distribution should be excluded and formula (6) should be read as follows

N	$u(p)$	MVUE	Refs.
5	$p_1^i p_2^j (1-p_1-p_2)^r$ $i,j,r \in \mathbb{N}$	$\dfrac{\binom{mn-r-i-j}{Z_1-i,\ Z_2-j}}{\binom{mn}{Z_1,\ Z_2}}$, $mn-r-i-j>0$	

Moreover, formula (7) of the Table of MVUE's for functions $u(\boldsymbol{\theta})$ of parameter $\boldsymbol{\theta}$ of the negative multinomial distribution should be read as

N	$u(\boldsymbol{\theta})$	MVUE	Refs.
7	$\mathcal{P}\{X_1=x_1,\dots,$ $X_\ell=x_\ell;\theta\}$ $\ell \le n$	$\dfrac{\Gamma(mn)\prod_{i=1}^{k} Z_i!\Gamma' m(n-\ell)+\sum_{i=1}^{k}(Z_i-x_{1i}-\cdots-x_{\ell i})'}{[\Gamma(n)]^\ell \Gamma[m(n-\ell)]\prod_{i=1}^{k}(Z_i-x_{1i}-\cdots-x_{\ell i})!}\times$ $\dfrac{\prod_{j=1}^{\ell}\Gamma\left(m+\sum_{i=1}^{k}x_{ji}\right)}{\Gamma\left(mn+\sum_{i=1}^{k}Z_i\right)\prod_{j=1}^{\ell}\prod_{i=1}^{k}x_{ji}!}$	

Butlleta de subscripció a la revista **Qüestiió**

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ ciutat _____

Tel. _____ Fax _____

Data d'expedició _____

Signatura: DNI o NIF _____

Desitjo subscriure'm a **Qüestiió** per a l'any 1994.
El preu de la subscripció és de 2.700 PTA.

Forma de pagament

☐ Transferència al compte de la Caixa de Catalunya número 6985/77, Agència 100, Comte d'Urgell 162, 08036 Barcelona

☐ Domiciliació bancària

☐ Taló nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retornar aquesta butlleta (o una fotocòpia) a:

Qüestiió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de la revista **Qüestiió**

El sotasignat _____

autoritza el Banc/Caixa _____

Agència núm. _____ adreça _____

Codi postal _____ ciutat _____

a abonar a la revista **Qüestiió** amb càrrec al meu compte número _____

_____, les subscripcions a **Qüestiió.**

_____ , a _____ d _____ de 19 _____

(Signatura)

El sotasignat _____

autoritza la revista **Qüestiió** a carregar al meu compte número _____

_____ , al Banc/Caixa _____

_____

Agència núm. _____ adreça _____

Codi postal _____ ciutat _____

l'import de les tarifes vigents de les subscripcions a la revista **Qüestiió.**

_____ , a _____ d _____ de 19 _____

(Signatura)