Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 1999, volum 23, núm. 3
Segona època

Entitats patrocinadores:

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

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 1999, volum 23, núm. 3

SUMARI

EDITORIAL

En aquest darrer número del volum 23 (1999) s'hi publiquen un total de vuit articles, adscrits en tres de les quatre seccions temàtiques, als quals cal afegir-hi una ressenya a l'apartat dedicat a les novetats sobre programari estadístic. Amb l'edició del tercer número, la producció editorial del volum corresponent a l'any 1999 totalitza 23 articles que, juntament amb la informació d'altres apartats de la revista, han suposat la publicació de 645 pàgines. Les xifres de l'exercici confirmen un augment de la producció respecte de l'any 1998 i, a la vegada, permeten millorar de nou la producció mitjana de la 2a època de la revista, que ja se situa en 19.6 articles i 485 pàgines impreses per volum/any.

També val la pena destacar que les millores qualitatives i quantitatives de l'edició impresa s'han vist acompanyades enguany d'una projecció creixent de la difusió electrònica: més de 15.000 consultes al web en els primers onze mesos del 1999, que han crescut ininterrompudament des de les 570 del gener fins a les 3.200 peticions d'accessos el novembre. En aquest sentit, el Consell Editor es planteja l'ampliació dels continguts i les facilitats d'accés al web de *Qüestiió* en el decurs del 2000, com ara la consulta interactiva dels problemes proposats i resolts o l'ampliació de l'avanç dels sumaris dels propers números, paral·lelament a les innovacions que millorin igualment la versió impresa actual de la revista.

Per últim, com és costum incloure en el darrer número de cada volum, *Qüestiió* fa balanç de la presentació i avaluació d'originals que han enregistrat les seves seccions. Atès que l'anterior relació feia referència al període desembre 1997-novembre 1998 (publicat en el número 3 del volum 22), la següent informació fa referència a l'interval desembre 1998-desembre 1999.

- articles sotmesos: 39
- articles en procés d'avaluació: 27
- articles acceptats: 25 (15 publicats i 10 en espera de publicació)
- articles rebutjats: 11

És interessant constatar que el nombre d'articles sotmesos ofereix també una evolució positiva, de manera que el volum d'aquest darrer any representa uns increments del 43%, 75% i 17.5% respecte del 1998, 1997 i 1996, respectivament; de fet, la xifra d'enguany suposa el registre més elevat de tota la 2a època de *Qüestiió*, un indicador significatiu de l'evolució prou satisfactòria de la revista. Les xifres anteriors es complementen amb el manteniment del percentatge de rebuig d'originals a l'entorn del 15% i el 30% dels articles sotmesos, per bé que el temps d'acceptació d'un article en els darrers dos anys s'ha situat en una mitjana de 10.7 mesos, que contrasta amb els 9.1 mesos del període 1992-99 o la mitjana de 7.2 mesos durant el 1987-91.

C.M. Cuadras i E. Ripoll, editors executius

Comentari de les seccions «Estadística» i «Biometria»

En aquesta ocasió, la secció «Estadística» conté tres originals. El primer, *The relative potency of two preparations applied in the supplemented block design with multivariate responses*, de Z. Hanusz, conté algunes contribucions rellevants sobre inferències en relació a la potència relativa de dues preparacions i, en particular, desenvolupa un contrast d'hipòtesi sobre la pendent d'un model lineal general que s'il·lustra amb un exemple. El segon article, *Análisis comparativo de estimadores pretest de heterocedasticidad en modelos econométricos. Un estudio Monte Carlo*, de R. Dios i C. Rodríguez, és un estudi comparatiu de constrastos d'heterocedasticitat en el model lineal amb variàncies dels errors possiblement desiguals, que inclou l'anàlisi de la potència dels diferents tests, l'estudi de les conseqüències d'utilitzar estimadors de mínims quadrats generalitzats i la proposta d'estratègies d'actuació òptimes davant la incertesa del tipus d'heterocestadicitat que presenti el model. El tercer article, *Independencia entre las cuestiones en el análisis factorial de tablas disyuntivas incompletas con preguntas condicionadas*, d'A. Zárraga i B. Goitisolo, és una contribució a l'anàlisi de correspondències múltiples, sobre dades d'enquestes en els casos de dades absents i preguntes condicionades a no resposta; els autors mostren que l'aplicació de l'anàlisi clàssica podria ser poc adequada i proposen l'alternativa d'introduir una marginal modificada.

L'article de la secció «Biometria», *Generalization of the kappa coefficient for categorical data, multiple observers and incomplete designs*, de V. Abraira i A. Pérez de Vargas, proposa una generalització del coeficient kappa que mesura la concordància entre una proporció observada i una proporció esperada, quan hi ha molts observadors i els dissenys són incomplets (és a dir, no tots els subjectes són classificats per tots els observadors). Aquesta generalització contempla ponderacions, així com el càlcul de la proporció d'acord observat i esperat entre múltiples observadors, de manera que el disseny complet quedi com un cas particular; els autors també proposen un interval de confiança i descriuen una il·lustració amb dades clíniques.

Carles M. Cuadras, editor executiu

Comentari de la secció «Estadística Oficial» i altres apartats

La secció «Estadística Oficial» prossegueix amb la publicació selectiva de les ponències presentades a les jornades internacionals «Generació d'informació estadística: qualitat i limitacions», que va organitzar el 1998 la Xarxa Temàtica «Enquestes i Qualitat de la Informació Estadística». Després dels treballs publicats de L. Lebart i V. Martínez, aquest número inclou quatre ponències a l'entorn del disseny i recollida de dades mostrals, amb aportacions provinents de l'àmbit acadèmic (corresponents a sociòlegs de

la Universitat Autònoma de Barcelona) i d'instituts d'estadística (a càrrec d'experts en disseny de mostres de l'INSÉE i l'INE). El primer article, *Utilisation d'informations auxiliaires dans les enquêtes par sondage*, d'Y. Tillé, planteja reduir ambigüitats de la noció de representativitat associada a les mostres estadístiques amb els "plans de mostreig equilibrats", vinculats a l'ús d'informació auxiliar (en forma de variables auxiliars) sobre la població que es mostreja. En el treball *Muestreo y recogida de datos para el análisis de redes sociales*, de J.M. Verd i J. Martí, s'exposen les insuficiències del disseny clàssic de mostres representatives pel que fa a les relacions entre individus d'una col·lectivitat i, com a proposta alternativa, les virtualitats de l'anàlisi de xarxes socials quan s'aplica en la definició de la "població" a investigar i en el procés de recollida de les "dades". En tercer lloc, l'article *La articulación entre lo cuantitativo y lo cualitativo: de las grandes encuestas a la recogida de datos intensiva*, de V. Borràs, P. López i C. Lozares, també insisteix en les mancances de les investigacions estadístiques per mostreig quan es tracta d'integrar aspectes o atributs qualitatius de la població; mitjançant l'Enquesta de la Regió Metropolitana de Barcelona, s'il·lustra la complementarietat amb les tècniques d'anàlisi qualitativa, basada en els grups de discussió, quan s'apliquen amb posterioritat a la generació de l'enquesta, aprofitant els isomorfismes d'espais topològics amb l'anàlisi multivariada. Per últim, en el treball de J. Porras, *Diseño de muestras en encuestas de población y hogares*, es presenta l'esquema de disseny mostral que l'INE utilitza en les enquestes —contínues i estructurals— adreçades a llars o a la població, ressaltant les actualitzacions dels marcs d'enquesta i aventurant l'impacte que suposarà el futur registre de població.

A continuació, la «Secció docent i problemes» inclou la presentació successiva de nous enunciats i la resolució dels problemes publicats en el número anterior d'aquest volum. Seguidament, la secció «Novetats de software» acull una detallada presentació de l'*Statistical Application Building Environment*, sistema estadístic per al desenvolupament d'aplicatius basat en la programació visual que culmina un projecte del programa ESPRIT IV desenvolupat, entre d'altres, per l'equip d'A. Prat de la UPC, amb l'assessorament de l'Idescat com un dels usuaris finals del prototipus STABLE. La recensió de la seva arquitectura i prestacions, a càrrec dels seus dissenyadors principals, evidencia la potència del nou paradigma visual que integra el sistema i inclou una aplicació pràctica de la capacitat d'ajustar modelitzacions estadístiques a mida.

El darrer apartat, dedicat a «Ressenyes d'activitats institucionals», inclou, com ja és costum, una revisió actualitzada d'activitats de la *Sociedad Española de Biometría*, amb l'anunci dels cursos monogràfics organitzats en col·laboració amb altres entitats. En segon lloc, s'ofereix una recensió del «Training for European Statisticians Institute» que *Qüestiió* redifon des del 1997, amb la relació dels cursos del *Programme 1999-2000* que s'impartiran de gener a juny del 2000, adreçats principalment als membres dels instituts d'estadística oficial en l'àmbit comunitari. Seguidament, es reprodueix l'anunci del «Tercer Congrés Europeu de Matemàtiques» —3ECM— (Barcelona, 10-14 juliol 2000) que organitza la Societat Catalana de Matemàtiques-IEC, sota els auspicis

de la European Mathematical Society i la col·laboració de la UB, la UPC i l'Idescat, entre d'altres. Finalment, s'anuncia l'edició de l'«International Workshop on Statistical Modelling» –15th IWSM– (Bilbao, 17-21 juliol 2000) que organitza la Universitat del País Basc.

Enric Ripoll, editor executiu

Estadística

QÜESTIIÓ, vol. 23, 3, p. 425-436, 1999

THE RELATIVE POTENCY OF TWO PREPARATIONS APPLIED IN THE SUPPLEMENTED BLOCK DESIGN WITH MULTIVARIATE RESPONSES

Z. HANUSZ
Agricultural University*

The method of point and interval estimation of the relative potency of two preparations: Standard and Test in the multivariate case is presented. General formulae for testing the hypotheses about the parallelism of regression lines and the relative potency have been adopted to experiments in which doses of preparations are applied in the supplemented block designs. The designs of this kind, with two groups of treatments, the first group comprising the doses of the Standard preparation and the second group- the doses of the Test preparation have been used for bioassay. The details of the experimental plans as well as the test functions are presented. The theoretical considerations are illustrated with an example involving a generated data set.

Keywords: Relative potency, parallel-line design, supplemented block design, multivariate response

AMS Classification: primary 62H15; secondary 62K10

* Institute of Applied Mathematics. Agricultural University. Akademicka 13, PL-20-934 Lublin. E-mail: Hanusz@ursus.ar.lublin.pl.

- Received June 1998.
- Acepted April 1999.

1. INTRODUCTION

One of the methods of comparing two preparations, where one preparation is known (Standard) and the other is new (Test), is estimation of their relative potency. In the case of parallel-line designs, the relative potency, ρ, is defined as the ratio of a dose of the Test preparation to such a dose of the Standard preparation that produces the same average response. The relative potency allows us to indicate which dose of the Test preparation produces the same response as one dose of the Standard preparation. This problem concerning univariate and multivariate observations was considered by many authors: Finney (1978), Meisner et al. (1986), Laska et al. (1985), Williams (1988), Vølund (1980, 1982), Carter and Hubert (1985), Rao (1954), Hanusz (1995) and many others. In the multivariate case, most of the authors considered the problem of point and interval estimation of the relative potency of preparations administered on homogenous experimental units. A similar problem arises when we apply doses of the preparations to units which are not homogenous. Especially, with agricultural experiments involving herbicides, for example, the most suitable designs for experiments are blocks. However, in the case where doses of two preparations are administered in blocks, then the supplemented block design should be recommended. In literature, supplemented block designs, also referred to as augmented or reinforced block designs were considered in papers: Nigam et al. (1988), Ceranka, Krzyszkowska, (1992, 1994), Caliński, Ceranka (1974). Blocks of the supplemented block designs contain basic and additional treatments. In particular, these designs can be adopted to bioassays if the doses of the Standard preparation constitute the basic treatments and the doses of the Test-additional treatments. In the paper we consider the multivariate setting where for each dose of the preparations a multivariate response is measured. On the responses we make basic assumptions: normality, the same covariance matrix for all responses, mutual uncorrelation between the responses, and the linear relation between the responses and the logarithm to base 10 of the doses. The formulae for testing hypotheses connected with parallelism and relative potency according to the experimental plan are presented. Finally, theoretical considerations are illustrated with an example involving a simulated data set.

2. NOTATIONS AND LINEAR MODEL

To describe a model of responses to the doses of the preparations administered in the supplemented block design let us introduce some notations. Let us consider a design with b blocks which are divided into two subblocks where the doses of the Standard preparation are applied on the first subblock and the doses of the Test preparation on the second subblock of each block. Let $\mathbf{k}_S$, $\mathbf{k}_T$ be the $(b \times 1)$ vectors of numbers of plots in the subblocks in each block. Suppose that the ith preparation is applied on ν_i doses denoted by u_{ij} $(i = S, T;\ j = 1, \ldots, \nu_i)$. The doses of the preparations are

replicated in the experiment, so let $\mathbf{r}_i$ be the $(\nu_i \times 1)$ vector of dose replications of the ith preparation. For example, let us consider the experimental plan with the doses of the Standard and the Test preparations administered in four blocks in the following way:

(2.1)

B_1	B_2	B_3	B_4
u_{S1}	u_{S1}	u_{S1}	u_{S2}
u_{S3}	u_{S3}	u_{S2}	u_{S1}
u_{S2}	u_{T1}	u_{S3}	u_{T3}
u_{T2}	u_{T2}	u_{T1}	u_{T2}
u_{T1}	u_{T3}	u_{T3}	u_{T1}
u_{T3}		u_{T2}	

The plan (2.1) is described by: $b = 4$, $\nu_S = \nu_T = 3$, $\mathbf{k}_S = [3,2,3,2]'$, $\mathbf{k}_T = [3,3,3,3]'$, $\mathbf{r}_S = [4,3,3]'$, $\mathbf{r}_T = [4,4,4]'$. The above vectors fulfill the following relations: $\mathbf{r}'_S\mathbf{1}_{\nu_S} = \mathbf{k}'_S\mathbf{1}_b = n_S = 10$, $\mathbf{r}'_T\mathbf{1}_{\nu_T} = \mathbf{k}'_T\mathbf{1}_b = n_T = 12$, where $\mathbf{1}_i$ denotes the vector of i ones and n_S, n_T are the total numbers of plots where the doses of S and T are applied. This experimental plan is also uniquely characterized by the incidence matrix $\mathbf{N}$, defined as:

$$\mathbf{N} = \begin{bmatrix} \mathbf{N}_S \\ \mathbf{N}_T \end{bmatrix} = \begin{matrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \end{matrix} \begin{matrix} u_{S1} \\ u_{S2} \\ u_{S3} \\ u_{T1} \\ u_{T2} \\ u_{T3} \end{matrix}$$
$$B_1 \; B_2 \; B_3 \; B_4$$

The matrix $\mathbf{N}$ shows that in the plan (2.1) the doses of the preparations appeared once in each block, only the second and the third dose of the Standard did not appear in the second and the fourth block. Moreover, the submatrices $\mathbf{N}_S$ and $\mathbf{N}_T$ fulfil the equalities: $\mathbf{N}_S\mathbf{1}_b = \mathbf{r}_S$, $\mathbf{N}_T\mathbf{1}_b = \mathbf{r}_T$, $\mathbf{N}'_S\mathbf{1}_{\nu_S} = \mathbf{k}_S$, $\mathbf{N}'_T\mathbf{1}_{\nu_T} = \mathbf{k}_T$.

Let us assume that for each dose of the preparations S and T a p-variate response vector is observed. Let us denote this response by $\mathbf{y}_{ijkl}$, where $i = S,T$; $j = 1,\ldots,\nu_i$; $k = 1,\ldots,r_{ij}$; $l = 1,\ldots,b$ and r_{ij} is the jth component of the vector $\mathbf{r}_i$. In most assays, the responses are linearly related to the logarithm of the doses (Finney, 1978). Therefore, the response can be written as:

$$\mathbf{y}_{ijkl} = \alpha_i + \beta_i x_{ij} + \tau_l + \mathbf{e}_{ijkl} \tag{2.2}$$

where τ_l denotes the $(p \times 1)$ vector of the effects of the lth block in which the dose u_{ij} was applied, α_i, β_i are the $(p \times 1)$ vectors of intercepts and regression slopes,

respectively, $x_{ij} = \log(u_{ij})$ denotes the logarithm to base 10 of the dose u_{ij}, $\mathbf{e}_{ijkl}$ is the vector of errors corresponding to $\mathbf{y}_{ijkl}$. As the whole experiment involves the total number of experimental units $n = n_S + n_T$ so the matrix $(n \times p)$ of all observations, $\mathbf{Y}$, whose rows are $\mathbf{y}'_{ijkl}$, can be written in the following form:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2.3}$$

and $\mathbf{B} = \begin{bmatrix} \tau' \\ \alpha' \\ \beta' \end{bmatrix}$ is the $((b+4) \times p)$ matrix of parameters, where $\tau = [\tau_1, \tau_2, \ldots, \tau_b]$ is the $(p \times b)$ matrix of blocks effects, $\alpha = [\alpha_S, \alpha_T]$, $\beta = [\beta_S, \beta_T]$ are the $(p \times 2)$ matrices of intercepts and slopes, $\mathbf{X} = [\mathbf{D}_1, \mathbf{D}_2, \Delta]$ is the $(n \times (b+4))$ matrix connected with the matrix of parameters $\mathbf{B}$, where $\mathbf{D}_1$ is an $(n \times b)$ matrix connected with τ, having the entries equal to 1 if the considered dose appeared in the block or 0 otherwise, $\mathbf{D}_2 = \begin{bmatrix} \mathbf{1}_{n_S} & \mathbf{0}_{n_S} \\ \mathbf{0}_{n_T} & \mathbf{1}_{n_T} \end{bmatrix}$, $\Delta = \begin{bmatrix} \mathbf{x}_S & \mathbf{0}_{n_S} \\ \mathbf{0}_{n_T} & \mathbf{x}_T \end{bmatrix}$, and $\mathbf{x}_i$ is the $(n_i \times 1)$ vector of logarithms of all doses of the ith preparation applied in (2.1) and located in the same order as responses $\mathbf{y}'_{ijkl}$ in the matrix $\mathbf{Y}$, $\mathbf{E}$ is an $(n \times p)$ matrix composed of all $\mathbf{e}'_{ijkl}$. About the model (2.3) we make assumptions that the rows of $\mathbf{Y}$ are independent and have the p-variate normal distribution with the same $(p \times p)$ unknown, positively defined covariance matrix, Σ.

3. TESTING HYPOTHESIS ABOUT THE SAME SLOPE

Two preparations can be compared by the relative potency if they similarly influence the responses. This similarity exists when the vectors of slopes for the Standard and the Test preparations in model (2.3) are equal. It means that for each measured feature of the observations, the regression coefficients (slopes) are equal, so the regression lines of each feature of responses versus the doses of Standard and the Test preparations are parallel. Such models are called a parallel- line model. If the model (2.3) has this characteristic, then the following hypothesis should be true:

$$H^0_\beta : \mathbf{L}'\mathbf{B} = \mathbf{0}' \text{ versus } H^1_\beta : \mathbf{L}'\mathbf{B} \neq \mathbf{0}' \tag{3.1}$$

where $\mathbf{L}' = [\mathbf{0}'_b, \mathbf{0}'_2, \mathbf{m}']$, $\mathbf{m}' = [1, -1]$, and $\mathbf{0}'_i$ is the $(1 \times i)$ vector of nulls. To test the hypothesis H^0_β in (3.1) we can use *Wilks' lambda* or *Lawley- Hotelling trace* statistic and because $rank(\mathbf{L}') = 1$ then both statistics are equivalent (see Appendix B). Let

us take the *Wilks' lambda* statistic which is defined as the ratio of two determinants:

$$\Lambda = \frac{|\mathbf{S}_E|}{|\mathbf{S}_E + \mathbf{S}_H|} \tag{3.2}$$

where

$$\begin{aligned}
\mathbf{S}_H &= \left(\mathbf{L}'\widetilde{\mathbf{B}}\right)' \left(\mathbf{L}' \left(\mathbf{X}'\mathbf{X}\right)^- \mathbf{L}\right)^{-1} \left(\mathbf{L}'\widetilde{\mathbf{B}}\right), \\
\widetilde{\mathbf{B}} &= \left(\mathbf{X}'\mathbf{X}\right)^- \mathbf{X}'\mathbf{Y}, \\
\mathbf{S}_E &= \left(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}\right)' \left(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}\right).
\end{aligned}$$

Using the transformation given in Meisner et al. (1986) we can write Λ in the following form:

$$\Lambda = \frac{1}{1+V} \tag{3.3}$$

where $V = \frac{(\mathbf{L}'\widetilde{\mathbf{B}})(\mathbf{S}_E)^{-1}(\mathbf{L}'\widetilde{\mathbf{B}})'}{\mathbf{L}'(\mathbf{X}'\mathbf{X})^-\mathbf{L}}$. In the formula for V the general inverse to the matrix $\mathbf{X}'\mathbf{X}$ appears, so $\mathbf{S}_H$, $\widetilde{\mathbf{B}}$ and $\mathbf{S}_E$ depend on the general inverse $(\mathbf{X}'\mathbf{X})^-$. As this inverse we propose the matrix given in Appendix A1. Moreover, in the vector $\mathbf{L}'$, only the subvector $\mathbf{m}'$ has not null elements, so V can be calculated using the formula:

$$V = \frac{\left(\mathbf{m}'\widetilde{\mathbf{B}}_1\right) \mathbf{S}_E^{-1} \left(\mathbf{m}'\widetilde{\mathbf{B}}_1\right)'}{\mathbf{m}'\mathbf{H}^{-1}\mathbf{m}} \tag{3.4}$$

where $\widetilde{\mathbf{B}}_1 = \mathbf{H}^{-1}\Delta'\Phi\mathbf{Y}$, and formulae for $\mathbf{H}$, Δ, Φ are given in A1 of the Appendix.

Under the null hypothesis H_β^0, $\frac{n-b-p-2}{p} \cdot V$ has *Snedecor's* F distribution with (p, $n-b-p-2$) degrees of freedom.

4. ESTIMATION OF THE RELATIVE POTENCY

Assuming the hypothesis (3.1) to be true, the model (2.3) can be reparametrized by replacing β consisting of two vectors β_S and β_T with one vector called also β. A new model takes a form:

$$\mathbf{Y} = \mathbb{X}\mathbb{B} + \mathbb{E} \tag{4.1}$$

where $\mathbb{B} = \begin{bmatrix} \tau' \\ \alpha' \\ \beta' \end{bmatrix}$, and τ, α remain the same as in (2.3) but β is the $(p \times 1)$ vector of slopes and, consequently, in the matrix $\mathbb{X} = [\mathbf{D}_1, \mathbf{D}_2, \vec{\Delta}]$, $\mathbf{D}_1$, $\mathbf{D}_2$ remain the same but $\vec{\Delta} = \Delta \cdot \mathbf{1}_2 = \begin{bmatrix} \mathbf{x}_S \\ \mathbf{x}_T \end{bmatrix}$ becomes a column vector.

In parallel- line designs with the linear relation between the responses and the logarithm of the doses we get the logarithm of the relative potency, $\mu = \log(\rho)$, which is the distance between the logarithms of doses of both preparations giving the same average responses. Let α_{Sj}, α_{Tj}, β_j denote the jth components of the vectors α_S, α_T, β correspond to the jth feature. Then the dependence of μ on intercepts and slope can be illustrated on Figure 1.

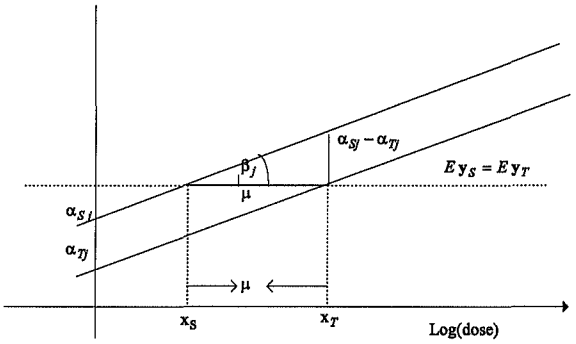


Figure 1. Logarithm of the relative potency in parallel-line design.

This figure shows that for the jth feature $(j = 1, \ldots, p)$, $\mu = \frac{\alpha_{Sj} - \alpha_{Tj}}{\beta_j}$ and if for each feature the same μ satisfies the above equality then in multivariate case the equality $\alpha_S - \alpha_T - \mu\beta = \mathbf{0}$ should be true. In the matrix notation the equality takes a form:

$$H_\mu^0 : \mathbf{L}'_\mu \mathbb{B} = \mathbf{0}' \text{ versus } H_\mu^1 : \mathbf{L}'_\mu \mathbb{B} \neq \mathbf{0}' \tag{4.2}$$

where $\mathbf{L}'_\mu = [\mathbf{0}'_b, \mathbf{m}', -\mu]$. To test (4.2) *Wilks' lambda* statistic in the form (3.3) is used:

$$\Lambda(\mu) = \frac{1}{1 + V(\mu)}$$

with $V(\mu)$ defined in the same manner as V in (3.3), taking $\mathbf{L}'_\mu$ instead of $\mathbf{L}'$ and $\mathbb{X}$ instead of $\mathbf{X}$. Using the formula for the general inverse to the matrix $\mathbb{X}'\mathbb{X}$ given in

Appendix A2, we get the formula for $V(\mu)$ as the ratio of two quadratics on μ:

$$V(\mu) = \frac{A\mu^2 - 2B \cdot \mu + C}{a\mu^2 - 2b \cdot \mu + c} \tag{4.3}$$

where $A = \widetilde{\beta}'\mathbb{S}_E^{-1}\widetilde{\beta}$, $\widetilde{\beta} = \vec{\Delta}'\Phi\mathbf{Y}$, $B = \mathbf{m}'\widetilde{\alpha}'\mathbb{S}_E^{-1}\widetilde{\beta}$, $\widetilde{\alpha} = \mathbf{F}_2\left(\mathbf{Y} - \vec{\Delta}'\widetilde{\beta}\right)$, $C = \mathbf{m}'\widetilde{\alpha}'\mathbb{S}_E^{-1}\widetilde{\alpha}\mathbf{m}$, $a = 1/h$, $b = \mathbf{m}'\mathbf{F}_2$, $c = \mathbf{m}'(\mathbf{C}^- + h\mathbf{b}_2\mathbf{b}_2^*)\mathbf{m}$, and h, $\mathbf{F}_2$, $\mathbf{b}_2$, $\mathbf{b}_2^*$, Φ and $\mathbf{C}$ are given in A1 and A2 of the Appendix.

The problem of testing the hypothesis H_μ^0 in (4.2) was discussed by Williams (1988), Carter and Hubert (1985), Meisner et al. (1986) or Hanusz (1995). The test derived by Carter and Hubert, improved by using Bartlett correction factor, compares $n^* \cdot \ln(1 + \min V(\mu))$ with the χ^2 distribution with $(p-1)$ degrees of freedom, where $n^* = n - r(\mathbb{X}) - \frac{p-1}{2} - \frac{1}{\max V(\mu)}$, ln is natural logarithm, min and max denote the minimum and maximum. If the hypothesis (4.2) is true then we take $\widehat{\mu}$ as the estimator of the logarithm of the relative potency, for which the test function $\Lambda(\mu)$ achieves its maximum. The $(1-\alpha)$ confidence interval for the logarithm of the relative potency is a set of μ satisfying the following inequality (see, Williams (1988), Meisner et al. (1986)):

$$P\left\{\Lambda(\mu) \geqslant \Lambda(\widehat{\mu}) \exp\left(-\chi^2_{p-1}(\alpha)/n^*\right)\right\} = 1 - \alpha \tag{4.4}$$

or :

$$P\left\{V(\mu) \leq (1 + V(\widehat{\mu})) \exp\left(\chi^2_{p-1}(\alpha)/n^*\right) - 1\right\} = 1 - \alpha$$

where $\exp(\cdot)$ denotes the exponential function.

5. NUMERICAL EXAMPLE

To illustrate the theoretical consideration in Sections 2, 3 and 4 we consider a generated data set corresponding to the experimental plan (2.1). Let us take: the number of blocks, $b = 4$, the number of features in each observation, $p = 3$, the vectors of intercepts: $\alpha_S = [11, 21, 31]'$, $\alpha_T = [10, 20, 30]'$, the vectors of slopes: $\beta_S = \beta_T = [1, 1, 1]'$, the matrix of block effects: $\tau = 0.01 \begin{bmatrix} 1 & -2 & 5 & -4 \\ -1 & 2 & -3 & 2 \\ 2 & 3 & -1 & -4 \end{bmatrix}$, the same number of doses: $\nu_S = \nu_T = 3$ and the same doses for the Standard and the Test preparations: 1, 10, 100 applied according to plan (2.1). Moreover, we assume that the covariance matrices

are the same for the Standard and the Test, and are equal to $\Sigma = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 4 \end{bmatrix}$. Using the MapleV packet we generate the data set from the normal distribution, having a null expectation and unit variance. The data is allocated in a (22×3) matrix, and after some mathematical transformations, using Cholesky decomposition of Σ (see, Krzanowski, 1988, p.478) we obtain the matrix of the observation $\mathbf{Y}$ satisfying (2.3). To calculate the test function of the hypothesis about the same slopes in (3.1), the test function V is calculated using formula given in (3.4). We obtained : $V = 0.227$, Snedecor F statistic, under the truthfulness H_β^0 is equal to $F^0 = \frac{13}{3} \cdot V = 0.98$ and the probability that F is smaller then F^0 is equal to 0.57, so the hypothesis in (3.1) is not rejected, therefore the model (2.3) describes a parallel- line design.

When we consider the model (4.1) and the hypothesis H_μ^0 in (4.2), we obtained:

$$V(\mu) = \frac{0.095\mu^2 - 0.060\mu + 0.049}{\frac{15}{221}\mu^2 + \frac{10}{663}\mu + \frac{370}{1989}}. \tag{5.1}$$

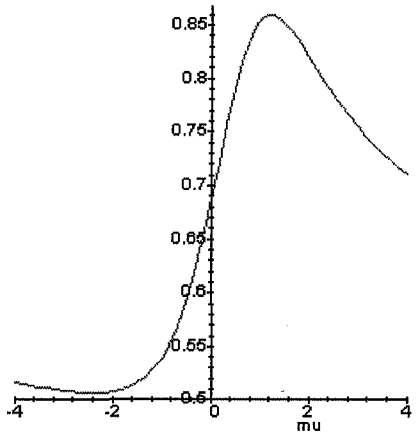


Figure 2. Shape of *Wilks' lambda* statistic for testing the hypothesis about the logarithm of the relative potency.

and *Wilks' lambda* statistic $\Lambda(\mu) = \frac{1}{1+V(\mu)}$, for $V(\mu)$ described in (5.1) has a plot given on Figure 2:

Calculating the extrema of $\Lambda(\mu)$ we obtain the minimum at point $\mu = -2.36$ and the maximum at $\mu = 1.21$. Moreover, $\chi_0^2 = n^* \cdot \ln(1 + \min V(\mu)) = 2.43$ and the probability that χ^2 is smaller then χ_0^2 is equal to 0.70, so the hypothesis in (4.2) is also not rejected. The point in which $\Lambda(\mu)$ achieves its maximum is taken as the estimator of

the logarithm of the relative potency, so $\widehat{\mu} = 1.21$. Using the inequality given in (4.4), the 95 per cent confidence interval for μ is given by the interval (0.99, 1.46). Putting to the equality: $\alpha_S - \alpha_T - \mu\beta = \mathbf{0}$, the values of α_S, α_T and β, let us notice that the true value of μ is equal to 1. Having the point and interval estimators for the logarithm of the relative potency we obtain the suitable estimators for the relative potency, namely, $\widehat{\rho} = 16.34$ and $\rho \in (9.92, 28.82)$.

6. CONCLUSION

In the paper we present the method of estimation of the relative potency in parallel-line assays. In literature, in the multivariate setting, the test functions using to test the multivariate hypotheses about parallelism and relative potency are presented in the case where the doses of preparations are administered to homogenous experimental units. In the experiments with homogenous experimental units, in the test functions given by (3.2) and (3.4) the inverse to $\mathbf{X}'\mathbf{X}$ and $\mathbb{X}'\mathbb{X}$ appeared and have unique forms. The problem of testing the same hypotheses in the case where the data is obtained on nonhomogenous experimental units is more complicated. The matrices $\mathbf{X}$ and $\mathbb{X}$ in (2.3) and (4.1) are not of full rank, so the general inverse matrices to $\mathbf{X}'\mathbf{X}$, and $\mathbb{X}'\mathbb{X}$ have to be known to get the test functions. This problem is solved in the case where doses of both preparations are administered to units with one directional changeability of units formed the supplemented block designs. The formulae given in (3.4) and (4.3) gives us possibility to calculate the values of the test functions and as the result of estimating the relative potency, which was the objective of the paper.

7. REFERENCES

Ahrens, H. and Läuter, L. (1974). *Mehrdimensionale Varianzanalyse.* Berlin, Akademie-Verlag.

Caliński, T. and Ceranka, B. (1974). «Supplemented block designs». *Biom. J.*, 16, 299-305.

Carter, E.M. and Hubert, J.J. (1985). «Analysis of Parallel-Line Assays with Multivariate Responses». *Biometrics*, 41, 703-710.

Ceranka, B. and Krzyszkowska, J. (1992). «Block designs with two groups of treatments». *Biometrical Letters*, 29(2), 33-44.

Ceranka, B. and Krzyszkowska, J. (1994). «Reinforced block designs with two groups of treatments». *Biometrical Letters*, 29(2). 17-25.

Finney, D.J. (1978). *Statistical method in Biological Assays*, (3rd edition). London: Charles Griffin and Co., Ltd.

Hanusz, Z. (1995). «Relative Potency of Two Preparations in Two-way Elimination of heterogeneity Designs with Multivariate Responses». *Biometrics*, 51, 1133-1139.

Krzanowski, W.J. (1988).*Principles of Multivariate Analysis.* Clarendon Press-Oxford.

Laska, E.M., Kushner,H.B. and Meisner, M. (1985). «Multivariate bioassay». *Biometrics*, 41, 547-554.

Meisner, M., Kushner, H. B. and Laska, E.M. (1986). «Combining multivariate bioassay». *Biometrics*, 42, 421-427.

Nigam, A.K., Puri, P.D. and Gupta, V.K. (1988). *Characterization and Analysis of Block Designs.* John Wiley & Sons.

Rao, C.R. (1954). «Estimation of the relative potency from multiple response data». *Biometrics*, 10, 208-220.

Rao, C.R. and Mitra, S.K. (1971). *Generalized Inverse of Matrices and Its Applications.* John Wiley, New York.

Vølund, Aa. (1980). «Combination of Multivariate Bioassay Results». *Biometrics* 38, 181-190.

Vølund, Aa. (1982). «Multivariate Bioassay». *Biometrics*, 36, 225-236.

Williams, D.A. (1988). «An exact confidence interval for the relative potency estimated from a multivariate bioassay». *Biometrics*, 44, 861-867.

APPENDIX

A1. Formula for the general inverse of the matrix $\mathbf{X}'\mathbf{X}$ used in Section 2

The matrix $\mathbf{X}'\mathbf{X}$ of the model (2.3) has the following form:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{D}_1'\mathbf{D}_1 & \mathbf{D}_1'\mathbf{D}_2 & \mathbf{D}_1'\Delta \\ \mathbf{D}_2'\mathbf{D}_1 & \mathbf{D}_2'\mathbf{D}_2 & \mathbf{D}_2'\Delta \\ \Delta'\mathbf{D}_1 & \Delta'\mathbf{D}_2 & \Delta'\Delta \end{bmatrix} = \begin{bmatrix} \mathbf{k}^{\delta} & \mathbf{K} & \mathbf{D}_1'\Delta \\ \mathbf{K}' & \mathbf{n}^{\delta} & \mathbf{D}_2'\Delta \\ \Delta'\mathbf{D}_1 & \Delta'\mathbf{D}_2 & \Delta'\Delta \end{bmatrix}$$

where

$$\begin{aligned} \mathbf{D}_1'\mathbf{D}_1 &= \mathbf{k}^{\delta} = diag(\mathbf{k}) \text{ and } \mathbf{k} = \mathbf{k}_S + \mathbf{k}_T, \\ \mathbf{D}_1'\mathbf{D}_2 &= \mathbf{K} = [\mathbf{k}_S, \mathbf{k}_T], \\ \mathbf{D}_2'\mathbf{D}_2 &= \mathbf{n}^{\delta} = diag(n_S, n_T). \end{aligned}$$

Using the formula given in Rao and Mitra (1971, p.41), the general inverse to $\mathbf{X}'\mathbf{X}$ has the following form:

$$(\mathbf{X}'\mathbf{X})^{-} = \begin{bmatrix} \mathbf{k}^{-\delta}(\mathbf{I}_b + \mathbf{K}\mathbf{C}^{-}\mathbf{K}'\mathbf{k}^{-\delta}) + \mathbf{b}_1\mathbf{H}\mathbf{b}_1^*, & -\mathbf{k}^{-\delta}\mathbf{K}\mathbf{C}^{-} + \mathbf{b}_1\mathbf{H}\mathbf{b}_2^*, & \mathbf{b}_1 \\ -\mathbf{C}^{-}\mathbf{K}'\mathbf{k}^{-\delta} + \mathbf{b}_2\mathbf{H}\mathbf{b}_1^*, & \mathbf{C}^{-} + \mathbf{b}_2\mathbf{H}\mathbf{b}_2^*, & \mathbf{b}_2 \\ \mathbf{b}_1^*, & \mathbf{b}_2^*, & \mathbf{H}^{-1} \end{bmatrix}$$

where

$$\begin{aligned} &\mathbf{H} = \Delta'\Phi\Delta, && \Phi = \mathbf{I}_n - \mathbf{D}_1\mathbf{k}^{-\delta}\mathbf{D}_1' - \mathbf{F}_2^*\mathbf{C}\mathbf{F}_2, \\ &\mathbf{b}_1 = -\mathbf{F}_1\Delta\mathbf{H}^{-1}, && \mathbf{b}_1^* = -\mathbf{H}^{-1}\Delta'\mathbf{F}_1^*, \\ &\mathbf{b}_2 = -\mathbf{F}_2\Delta\mathbf{H}^{-1}, && \mathbf{b}_2^* = -\mathbf{H}^{-1}\Delta'\mathbf{F}_2^* \\ &\mathbf{F}_2 = \mathbf{C}^{-}(\mathbf{D}_2' - \mathbf{K}'\mathbf{k}^{-\delta}\mathbf{D}_1'), && \mathbf{F}_2^* = (\mathbf{D}_2 - \mathbf{D}_1\mathbf{k}^{-\delta}\mathbf{K})\mathbf{C}^{-}, \\ &\mathbf{F}_1 = \mathbf{k}^{-\delta}(\mathbf{D}_1' - \mathbf{K}\mathbf{F}_2), && \mathbf{F}_1^* = (\mathbf{D}_1 - \mathbf{F}_2^*\mathbf{K}')\mathbf{k}^{-\delta}, \\ &\mathbf{C} = \mathbf{n}^{\delta} - \mathbf{K}'\mathbf{k}^{-\delta}\mathbf{K} \end{aligned}$$

and $\mathbf{C}^{-}$ denotes the general inverse to (2×2) matrix $\mathbf{C}$. Let us notice that only in the case when the general inverse to $\mathbf{C}$ is symmetric then $\mathbf{b}_1^*$, $\mathbf{b}_2^*$, $\mathbf{F}_1^*$, and $\mathbf{F}_2^*$ are the transposition of $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{F}_1$ and $\mathbf{F}_2$, respectively.

A2. Formula for the general inverse of the matrix $\mathbb{X}'\mathbb{X}$ used in Section 4

The general inverse to the matrix $\mathbb{X}'\mathbb{X}$ from the model (4.1) has the following form:

$$(\mathbb{X}'\mathbb{X})^{-} = \begin{bmatrix} \mathbf{k}^{-\delta}(\mathbf{I}_b + \mathbf{K}\mathbf{C}^{-}\mathbf{K}'\mathbf{k}^{-\delta}) + h\mathbf{b}_1\mathbf{b}_1^*, & -\mathbf{k}^{-\delta}\mathbf{K}\mathbf{C}^{-} + h\mathbf{b}_1\mathbf{b}_2^*, & \mathbf{b}_1 \\ -\mathbf{C}^{-}\mathbf{K}'\mathbf{k}^{-\delta} + h\mathbf{b}_2\mathbf{b}_1^*, & \mathbf{C}^{-} + h\mathbf{b}_2\mathbf{b}_2^*, & \mathbf{b}_2 \\ \mathbf{b}_1^*, & \mathbf{b}_2^*, & \frac{1}{h} \end{bmatrix}$$

where $h = \vec{\Delta}'\Phi\vec{\Delta}$, is a constant now, and the others formulae remain the same as in A1 taking the constant h instead of (2×2) matrix $\mathbf{H}$.

B. Equality of *Wilks' lambda* and *Lawley – Hotelling trace* statistics

When we test the hypothesis $\mathbf{L}'\mathbf{B} = \mathbf{0}'$, where the rank of the matrix $\mathbf{L}$ is equal to 1 then *Wilks' lambda* statistic and *Lawley – Hotelling trace* statistic are equivalent. To show this, let us consider *Wilks' lambda* statistic:

$$\Lambda = \frac{|\mathbf{S}_E|}{|\mathbf{S}_E + \mathbf{S}_H|} \tag{5.2}$$

where $|\mathbf{A}|$ denotes the determinant of a matrix $\mathbf{A}$, $\mathbf{S}_H = \left(\mathbf{L}'\widetilde{\mathbf{B}}\right)' \left(\mathbf{L}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{L}\right)^{-1} \left(\mathbf{L}'\widetilde{\mathbf{B}}\right)$, $\widetilde{\mathbf{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{Y}$, $\mathbf{S}_E = \left(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}\right)' \left(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}\right)$. Meisner et al. (1986) showed that Λ can be transform to the following form:

$$\Lambda = \frac{1}{1 + \left(\mathbf{L}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{L}\right)^{-1} \left(\mathbf{L}'\widetilde{\mathbf{B}}\right) \mathbf{S}_E^{-1} \left(\mathbf{L}'\widetilde{\mathbf{B}}\right)'}.$$

On the other hand, *Lawley – Hotelling trace* statistic is equal to:

$$T^2 = trace(\mathbf{S}_E^{-1}\mathbf{S}_H).$$

Using the trace property we can write:

$$\begin{aligned} T^2 &= trace\left(\mathbf{S}_E^{-1} \left(\mathbf{L}'\widetilde{\mathbf{B}}\right)' \left(\mathbf{L}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{L}\right)^{-1} \left(\mathbf{L}'\widetilde{\mathbf{B}}\right)\right) \\ &= trace\left(\left(\mathbf{L}'\widetilde{\mathbf{B}}\right) \mathbf{S}_E^{-1} \left(\mathbf{L}'\widetilde{\mathbf{B}}\right)' \left(\mathbf{L}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{L}\right)^{-1}\right) \\ &= \left(\mathbf{L}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{L}\right)^{-1} \left(\mathbf{L}'\widetilde{\mathbf{B}}\right) \mathbf{S}_E^{-1} \left(\mathbf{L}'\widetilde{\mathbf{B}}\right)'. \end{aligned}$$

Moreover in the case where $rank(\mathbf{L}) = 1$, and the hypothesis $\mathbf{L}'\mathbf{B} = \mathbf{0}'$ is true then $\frac{n-r(\mathbf{X})-p+1}{p} \cdot \frac{1-\Lambda}{\Lambda}$ and $\frac{n-r(\mathbf{X})-p+1}{p} \cdot T^2$ have F *Snedecor* distribution with $(p, n - r(\mathbf{X}) - p + 1)$ degrees of freedom (see, Ahrens, Läuter, 1974). One can see that $\frac{1-\Lambda}{\Lambda} = T^2$.

QÜESTIIÓ, vol. 23, 3, p. 437-464, 1999

ANÁLISIS COMPARATIVO DE ESTIMADORES PRETEST DE HETEROCEDASTICIDAD EN MODELOS ECONOMÉTRICOS. UN ESTUDIO MONTE CARLO

R. DIOS PALOMARES
C. RODRÍGUEZ FONSECA
Universidad de Córdoba*

En el presente artículo se recogen los resultados de una investigación llevada a cabo sobre el comportamiento de pretest de heterocedasticidad. Con este fin se ha diseñado un experimento Monte Carlo, introduciendo como proceso generador de datos un modelo con tres supuestos sobre la estructura de la varianza del error y con distintos niveles de heterocedasticidad para cada uno de ellos. Asimismo, se analiza la potencia de los diferentes contrastes de heterocedasticidad bajo los distintos supuestos de estructura heterocedástica. Además, se estudia la consecuencia de utilizar un estimador por Mínimos Cuadrados Generalizados Factibles cuya matriz $\hat{\Omega}$ no se corresponde con la matriz Ω del proceso generador de datos, mediante una función de riesgo de los estimadores pretest. Se concluye que es más importante el procedimiento de estimación empleado que el contraste aplicado para detectar la heterocedasticidad.

A comparative analysis of heterocedasticity pretest estimators in econometrics models. A Monte Carlo study

Palabras clave: Test de heterocedasticidad, pretest, Monte Carlo, curva de potencia

Clasificación AMS: 62J05, 62J20

* Universidad de Córdoba. Departamento de Estadística, Econometría, Investigación Operativa y Organización de Empresas. Avda. Menéndez Pidal, s/n. 14004 Córdoba.

– Recibido en marzo de 1998.
– Aceptado en junio de 1999.

1. INTRODUCCIÓN

En el modelo lineal general $y = X\beta + e$, el vector de perturbaciones e sigue una distribución normal con vector de medias cero y matriz de varianzas-covarianzas $\Omega = E\lfloor(e-E(e))(e-E(e))'\rfloor$. Si dicha matriz es igual a $\sigma^2 * \mathbf{I}_T$, siendo $\mathbf{I}_T$ la matriz identidad de rango igual al número de observaciones T, estaríamos en presencia de un modelo homocedástico, pero en la práctica hay casos en los que no se cumple la hipótesis de homocedasticidad siendo la matriz de varianzas-covarianzas de e la siguiente:

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdot & \cdot & 0 \\ 0 & \sigma_2^2 & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & \sigma_T^2 \end{pmatrix}$$

es decir, la varianza del error es distinta para cada observación y, por tanto, nos encontramos en el caso de un modelo heterocedástico.

Ante la posible existencia de heterocedasticidad en el modelo, se impone la necesidad de realizar un tratamiento de la misma enfocado en dos aspectos: contraste de homocedasticidad y estimación óptima.

Para el contraste de la hipótesis de homocedasticidad se han desarrollado una serie de métodos cuya potencia es muy variable y depende, entre otras cosas, de la estructura real de la heterocedasticidad.

Los contrastes de heterocedasticidad aplicados con más frecuencia en la bibliografía se clasifican en dos grupos. El primero de ellos está formado por los tests específicos para cada uno de los siguientes supuestos sobre la estructura de la varianza del error (Fomby, 1984): $\sigma_i^2 = z_i'\alpha$, $\sigma_i^2 = (z_i'\alpha)^2$, $\sigma_i^2 = \exp(z_i'\alpha)$. El segundo grupo está compuesto por los tests generales o no específicos. Entre ellos destacan el test de Breush y Pagan (1979), Goldfeld y Quandt (1965), Harvey (1976), test de Picos (Goldfeld y Quandt, 1965), test de Correlación por Rangos (Spearman,1904) y White (1980). La potencia de alguno de ellos ya ha sido cuestionada ampliamente en la bibliografía: Carrol y Ruppert (1981), Glejser (1969), Godfrey (1978), Harrison (1979), Harvey y Phillips (1974) y Hausman (1978).

El segundo aspecto es estimar de forma óptima los parámetros del modelo. La estimación de los parámetros se puede realizar de dos formas distintas:

a) mediante los estimadores por Mínimos Cuadrados Ordinarios (M.C.O.), $\hat{\beta} = = (X'X)^{-1}X'y$, que sólo son óptimos cuando el modelo es homocedástico y

b) a través de los estimadores por Mínimos Cuadrados Generalizados Factibles (M.C.G.F.), $\tilde{\tilde{\beta}} = (X' \hat{\Omega}^{-1} X)^{-1} X' \hat{\Omega}^{-1} y$, que sólo son óptimos cuando la matriz está bien estimada.

Los procedimientos de estimación por M.C.G.F., también se clasifican en dos grupos. En el primer grupo se encuentran los métodos de estimación específicos para cada uno de los supuestos sobre la estructura de la varianza del error (Fomby, 1984) y en el segundo grupo los que denominaremos procedimientos de estimación generales. Estos estimadores generales se van a diferenciar en la forma de estimar la matriz Ω. Las propiedades de los anteriores procedimientos no han sido analizadas en pequeñas muestras.

Las consecuencias que se derivan de un fallo en el procedimiento de estimación son muy importantes, de modo que lo ideal sería que en caso de homocedasticidad se estimara por M.C.O. y, en caso de heterocedasticidad, se aplicaran los M.C.G.F. con una matriz Ω correctamente estimada.

Toda esta problemática se puede enmarcar en el contexto de la estimación pretest. Definimos para ello un estimador pretest de heterocedasticidad como aquel que engloba a los estimadores por M.C.O. y por M.C.G.F., de modo que se decide estimar según uno u otro procedimiento, dependiendo del resultado de un test de hipótesis preliminar. El estimador pretest $\hat{\hat{\beta}}$ que planteamos coincide con el estimador por M.C.O. si se acepta la hipótesis nula de homocedasticidad, y en caso contrario el estimador utilizado es el de M.C.G.F.

Es decir,

$$\hat{\hat{\beta}} = \begin{cases} \hat{\beta} & \text{si se acepta } H_0 \\ \tilde{\tilde{\beta}} & \text{si se rechaza } H_0 \end{cases}$$

Así, la combinación de un test y un procedimiento de estimación de Ω definen un pretest de heterocedasticidad. El estimador pretest es insesgado y su varianza está comprendida entre la de los estimadores por M.C.O. y la de los M.C.G.F.

Las propiedades muestrales del estimador pretest se pueden estudiar en el contexto de la matriz de riesgo o Error Cuadrático Medio (E.C.M.) que será

$$R(\beta, \hat{\hat{\beta}}) = \text{cov}(\hat{\hat{\beta}}) + (\text{sesgo de } \hat{\hat{\beta}})(\text{sesgo de } \hat{\hat{\beta}})' \tag{1}$$

Igualmente, se puede analizar dicho modelo de riesgo en función de la importancia que tenga la heterocedasticidad en el Proceso Generador de Datos (P.G.D.). Si se resume dicha importancia en un parámetro λ y nos interesa conocer el comportamiento del estimador pretest en el espacio paramétrico de λ, tenemos la función

$$R(\lambda,\beta,\hat{\hat{\beta}}) = E\left[(\hat{\hat{\beta}}_{(\lambda)}-\beta)(\hat{\hat{\beta}}_{(\lambda)}-\beta)'\right] + \left[E(\hat{\hat{\beta}}_{(\lambda)})-\beta\right]\left[E(\hat{\hat{\beta}}_{(\lambda)})-\beta\right]'$$

que se podrá calcular sólo en el caso de que λ y β sean conocidos, como ocurre en un experimento Monte Carlo.

Son muchos los planteamientos teóricos de nuevas metodologías que han sido confirmados utilizando el método Monte Carlo. Mikhail (1972) lleva a cabo una simulación, para pequeñas muestras, de las propiedades de estimadores econométricos. Sowey (1973) realiza una revisión bibliografíca en la que clasifica los estudios Monte Carlo aplicados a la Econometría. Hendry y Harrison (1974) emplearon el método Monte Carlo para observar el comportamiento de pequeñas muestras en el procedimiento de mínimos cuadrados. De nuevo Mikhail, esta vez en el año 1975, estudia las propiedades de distintos estimadores a través de un estudio Monte Carlo. Kloek y Dijk (1978) analizaron los estimadores bayesianos mediante el método Monte Carlo. Mizon y Hendry (1980) utilizan un experimento Monte Carlo para los contrastes de especificación dinámica. Surekha y Griffiths (1984) también realizaron un experimento Monte Carlo con el objetivo de estudiar el comportamiento de los estimadores bayesianos bajo dos supuestos distintos para la estructura de la heterocedasticidad.

Teniendo en cuenta todo lo anterior, queda claro que en el análisis del problema de heterocedasticidad en el modelo econométrico, la bondad de un estimador pretest dependerá, por un lado, del acierto que tenga el test en aceptar o rechazar la homocedasticidad y, por otro lado, de lo próxima que esté la estimación de Ω a su verdadera matriz en el P.G.D.

El presente trabajo resume los resultados más relevantes de una investigación llevada a cabo con el objetivo principal de investigar si existe algún estimador pretest que se comporte mejor que los demás, en circunstancias de incertidumbre sobre la estructura de la varianza del error, que es el caso más frecuente que nos encontramos en la práctica.

Con este fin se ha diseñado un experimento Monte Carlo, introduciendo como P.G.D. un modelo con tres supuestos sobre la estructura de la varianza del error y con distintos niveles de heterocedasticidad para cada uno de ellos.

Los objetivos secundarios que se plantean son los siguientes:

1. Analizar la potencia de los diferentes contrastes de heterocedasticidad bajo los distintos supuestos de estructura heterocedástica.
2. Estudiar la consecuencia de utilizar un estimador por M.C.G.F. cuya matriz $\hat{\Omega}$ no se corresponda con la matriz Ω del proceso generador de datos. Por ejemplo, cuando existe homocedasticidad.

3. Estudiar el comportamiento de los distintos pretests con el fin de plantear una estrategia de actuación óptima ante la incertidumbre sobre la posible naturaleza de la heterocedasticidad.

2. METODOLOGÍA

El experimento está diseñado con el fin de analizar el comportamiento de los contrastes de heterocedasticidad y de los estimadores por M.C.O. y por M.C.G.F., cuando el P.G.D. corresponde a modelos con estructuras de varianza del error homocedásticas y heterocedásticas.

Así, se realiza un experimento Monte Carlo tomando como P.G.D. un modelo base $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$, generando muestras con distintas hipótesis sobre la matriz, incluida $\Omega = I$ (caso de homocedasticidad).

En el experimento se han utilizado los datos del gasto total en productos agroalimentarios (**Y**) y de renta familiar disponible (**X**) correspondientes a cuarenta y siete provincias españolas en el año 1990 cuyos datos se presentan en la tabla 1 del anexo. La especificación del modelo es $\boldsymbol{Y}_i = \beta_0 + \beta_1 \cdot \boldsymbol{X}_i + e_i$. Tras realizar una primera estimación, el P.G.D. que tomaremos como base será $\boldsymbol{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \boldsymbol{X}_i + e_i$, ya que asignaremos a los parámetros estructurales β_0 y β_1, los valores estimados por M.C.G.F en el modelo real que son $\hat{\beta}_0 = 2633.135$ y $\hat{\beta}_1 = 0.156$. Todos los contrastes de heterocedasticidad utilizados rechazan la hipótesis de homocedasticidad en el modelo, siendo la varianza residual estimada igual a 550.000.

2.1. Fuentes de variación

Se desarrolla el experimento con dos fuentes de variación: la forma de la función $\sigma_i^2 = h(z_i'\boldsymbol{\alpha})$ y la gravedad de la heterocedasticidad λ.

2.1.1. La forma de la función

Se establece un P.G.D. bajo los siguientes supuestos

a) $\sigma_i^2 = a + b \cdot \boldsymbol{X}_i \quad \Rightarrow \quad$ Supuesto A.

b) $\sigma_i^2 = (a + b \cdot \boldsymbol{X}_i)^2 \quad \Rightarrow \quad$ Supuesto B.

c) $\sigma_i^2 = \exp(a + b \cdot \boldsymbol{X}_i) \quad \Rightarrow \quad$ Supuesto C.

2.1.2. La gravedad de la heterocedasticidad λ

Se sustituye la varianza muestral del modelo base, $\hat{\sigma}_e^2$, en la ecuación $\hat{\sigma}_e^2 = a + b \cdot \overline{X}$, $\hat{\sigma}_e^2 = (a + b \cdot \overline{X})^2$, ó $\hat{\sigma}_e^2 = \exp(a + b \cdot \overline{X})$, dependiendo del supuesto sobre la varianza del error y siendo $\overline{X}$ el valor medio de la variable exógena. Dando valores a a en estas ecuaciones se obtienen valores de b, formándose parejas de valores (a, b) que hacen que la varianza se asigne, bien al término independiente, bien a la variable exógena, o bien a los dos. Por ejemplo, si a es igual a cero toda la varianza se le asignaría a la variable exógena, sería el caso de heterocedasticidad máxima; si b es igual a cero la varianza del error sería constante, es decir, estaríamos en el caso de homocedasticidad.

Para cada pareja de valores (a, b) que constituye un ensayo o punto experimental, se calcula la serie $t_i = a + b \cdot \boldsymbol{X}_i$ sustituyendo la pareja de valores (a, b). Si a dicha serie se le estima la desviación típica y se le calcula la media podemos construir una medida relativa de dispersión denominada coeficiente de variación

$$\text{C.V.} = \frac{\hat{\sigma}_t}{\overline{t}}$$

Esta medida de dispersión será representada por λ que indica la gravedad de la heterocedasticidad. La gravedad se establece, para cada uno de los supuestos anteriores, en siete niveles que suponen desde homocedasticidad hasta heterocedasticidad extrema. Los valores de a y b definirán esta gravedad. Se pueden ver los valores correspondientes a los parámetros de interés de este procedimiento en las tablas 2, 3, y 4 del anexo.

2.2. Generación del término de error heterocedástico y de cada muestra

Los pasos que se han seguido para generar cada muestra son los siguientes:

- Calcular el valor de σ_i^2, sustituyendo a, b y X en $\sigma_i^2 = a + b \cdot \boldsymbol{X}_i$, $\sigma_i^2 = (a + b \cdot \boldsymbol{X}_i)^2$ ó $\sigma_i^2 = \exp(a + b \cdot \boldsymbol{X}_i)$, según se trabaje con el primer, segundo o tercer supuesto, respectivamente.
- Generar un valor aleatorio de una distribución $N(0, 1)$.
- Generar la variable e_i a través de $e_i = \sigma_i \cdot z_i$.
- Generar cada punto muestral sustituyendo los valores de $\hat{\beta}_0$, $\hat{\beta}_1$, X_i y e_i en la ecuación

$$\boldsymbol{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \boldsymbol{X}_i + e_i.$$

Se han formulado tres supuestos de estructura de varianza del error con siete niveles de heterocedasticidad cada uno; por tanto, serán veintiún procesos distintos para cada uno de los cuales se generan dos mil muestras de tamaño cuarenta y siete.

2.3. Contrastes de heterocedasticidad

Para cada muestra generada se realizan los siguientes tests.

a) Contrastes específicos (Fomby, 1984):

- Test para cuando la varianza es una función lineal de las variables exógenas, $\sigma_i^2 = z_i'\boldsymbol{\alpha}$.
- Test para cuando la desviación típica es una función lineal de las variables exógenas, $\sigma_i = z_i'\boldsymbol{\alpha}$.
- Test para la heterocedasticidad multiplicativa, $\sigma_i^2 = \exp(z_i'\boldsymbol{\alpha})$.

b) Contrastes no específicos:

- Test de Breush-Pagan.
- Test de Goldfeld-Quandt.
- Test de Harvey.
- Test de Picos.
- Test de Correlación por Rangos (Spearman).
- Test de White.

2.4. Métodos de estimación

Para cada una de las muestras generadas se procede al cálculo de los siguientes estimadores. Por un lado, se calculan los estimadores por M.C.O. y por M.C.G. (teórico) a través de $\hat{\beta} = (X'X)^{-1}X'\boldsymbol{y}$ y $\tilde{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\boldsymbol{y}$, respectivamente. Por otro lado, se calculan tres estimadores para los que existe una forma funcional específica de heterocedasticidad (Fomby, 1984). A estos estimadores se les representará por *SA*, *SB* y *SC*, según correspondan al primer supuesto para la estructura de la varianza del error, al segundo o al tercero, respectivamente.

Por último, se calculan seis estimadores por M.C.G.F., $\tilde{\hat{\beta}} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}$. Estos estimadores se diferencian para una misma muestra en la matriz $\hat{\Omega}$. Si en la diagonal de dicha matriz situamos los elementos de los residuos por M.C.O. al cuadrado, se obtiene un estimador que representaremos por *R2*. Cuando la diagonal de $\hat{\Omega}$ esté compuesta por los elementos de la variable exógena que provoca la heterocedasticidad,

el estimador resultante se representará por *VE*, y si es el cuadrado de la variable exógena la que ocupa la diagonal de dicha matriz, el estimador resultante será representado por *VE2*. Si en los modelos de los tests de Breush-Pagan, Harvey y White estimamos sus respectivas variables estimadas, se pueden construir, a partir de éstas, tres matrices $\hat{\Omega}$ distintas sin más que sustituir en su diagonal cada una de estas variables. Los tres nuevos estimadores se denominarán *B-P*, *HA* y *W*, respectivamente.

En total tenemos once estimadores distintos. En la Tabla 5 del Anexo se ofrecen las notaciones de los estimadores por M.C.G.F. utilizados en el experimento.

2.5. Estimadores pretest

Cada estimador pretest se compone de un contraste y un procedimiento de estimación de Ω, para el caso en que se rechace la hipótesis nula de homocedasticidad. Hemos diseñado además, para cada contraste, varios estimadores pretest que corresponden a distintas formas de estimar Ω que son *R2, VE, VE2* y *W*. En los tests de Breush-Pagan y Harvey, además de los cuatro procedimientos «tipo» (*R2, VE, VE2* y *W*), se estimará con el procedimiento asociado a *B-P* y *HA*, respectivamente. Con los tests específicos se realizará algo similar. Así, en el test para cuando $\sigma_i^2 = z_i'\alpha$ se añade la estimación que corresponde a *SA*; al test para cuando $\sigma_i = z_i'\alpha$ se le agrega la estimación correspondiente a *SB* y al test para cuando $\sigma_i^2 = \exp(z_i'\alpha)$ se le une la estimación asociada a *SC*.

En la Tabla 2 (Anexo II) se recoge la nomenclatura utilizada para distinguir los estimadores pretest.

2.6. Cálculo de las curvas de potencia

En base a los resultados de los distintos tests se calcula empíricamente la probabilidad de cada uno de ellos de rechazar la hipótesis nula de homocedasticidad. Así, se representan las curvas de potencia de los contrastes de heterocedasticidad para cada supuesto concreto en función de los valores de λ. Se trabaja con un nivel de significación α igual a 0.05.

2.7. Cálculo de las curvas del E.C.M.

Se hallan los valores de los errores cuadráticos medios definidos en (1) para los distintos estimadores obtenidos en el experimento: M.C.O., M.C.G., M.C.G.F. y estimadores pretest. Posteriormente, se calculan los E.C.M. relativos al del estimador por M.C.G., dividiendo los primeros por los últimos con el fín de eliminar el efecto de dimensión

y bajo al supuesto de que éstos son los óptimos teóricos. Con estos nuevos valores se representan las curvas del E.C.M. en función del λ correspondiente para cada ensayo.

2.8. Tratamiento informático

Para poder llevar a cabo todos los pasos descritos en el diseño del experimento ha sido necesario realizar una serie de macros mediante el programa Econometric Views 2.0. Las macros que se han creado son las correspondientes a:

- La generación del término de error heterocedástico.
- La generación de cada muestra.
- La aplicación de la batería de tests de heterocedasticidad.
- La aplicación de los distintos métodos de estimación.
- El cálculo de los distintos estimadores pretest.
- El cálculo de las curvas de potencia.
- El cálculo de las curvas del E.C.M.

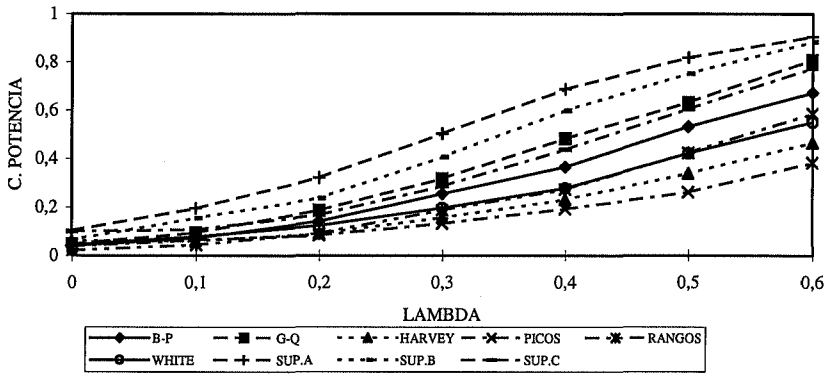
3. RESULTADOS

3.1. Estudio de las curvas de potencia

3.1.1. Supuesto $\sigma^2 = a + b \cdot X$

Las curvas de potencia correspondientes a los nueve tests llevados a cabo en el experimento se pueden ver en el Gráfico 1.

Hay que destacar dos aspectos. Primero, los tests específicos en presencia de homocedasticidad ($\lambda = 0$) tienen un comportamiento pésimo desde el punto de vista del error de tipo I, mientras que por el contrario, los tests no específicos ofrecen una probabilidad de error de tipo I más cercano al nivel de significación 0.05. Segundo, en presencia de heterocedasticidad la situación anterior se invierte, es decir, los tests generales poseen una probabilidad de error de tipo II muy elevado, mientras que los test específicos alcanzan probabilidades de rechazo de la hipótesis nula del 90%, caso del test específico para el supuesto que estamos considerando $\sigma^2 = a + b \cdot X$.

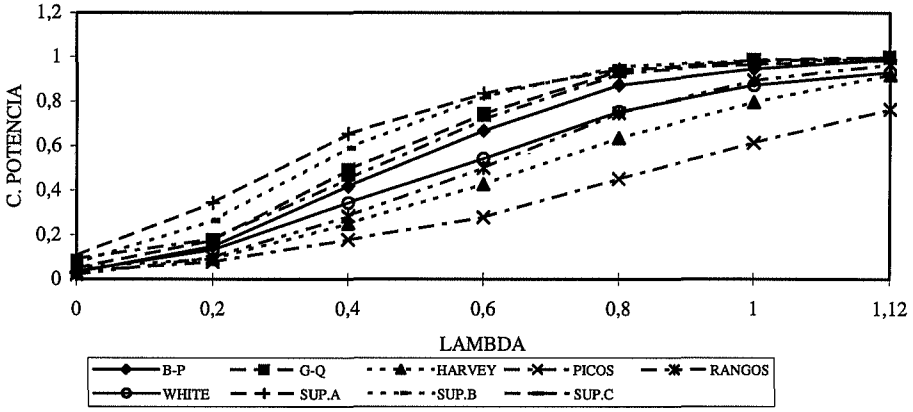


Fuente: Elaboración propia.

Gráfico 1. Curvas de potencia (Sup. A)

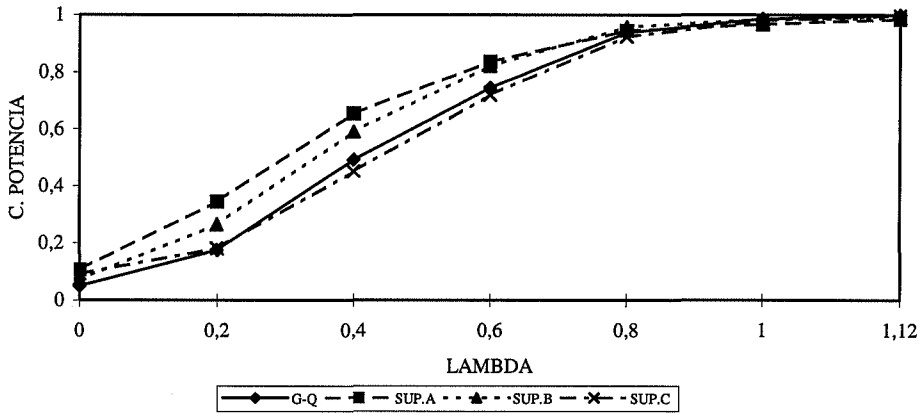
3.1.2. Supuesto $\sigma^2 = (a + b \cdot X)^2$

Las curvas de potencia de los tests se recogen en el Gráfico 2. Si se comparan las curvas de potencia de los Gráficos 1 y 2, para valores iguales de λ, podemos comprobar que se muestran igual de eficaces a la hora de detectar heterocedasticidad siendo cierta. De aquí se puede deducir que el supuesto del P.G.D. tiene poca influencia sobre el comportamiento de los contrastes de heterocedasticidad.



Fuente: Elaboración propia.

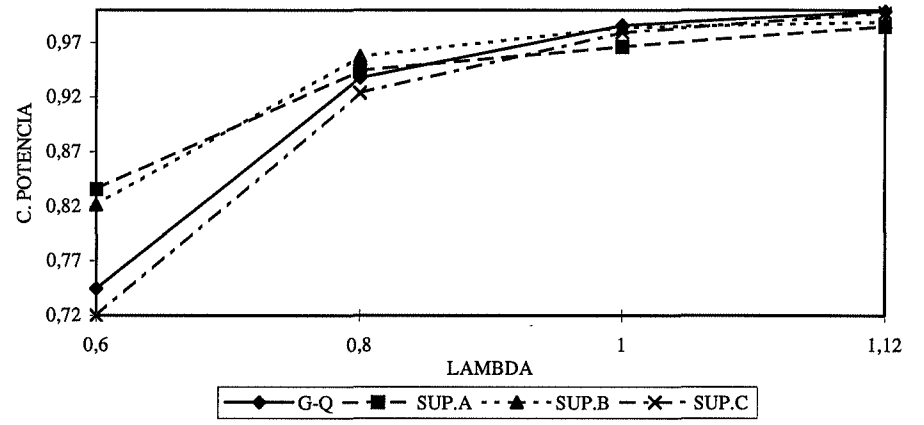
Gráfico 2. Curvas de potencia (Sup. B)



Fuente: Elaboración propia.

Gráfico 3. Curvas de potencia de los tests más potentes (Sup. B)

Siguiendo el criterio de máxima potencia, simplificaremos el número de curvas para analizar con mayor claridad la potencia de éstas, obteniendo así el Gráfico 3.



Fuente: Elaboración propia.

Gráfico 4. Curvas de potencia de los tests más potentes (Sup. B)

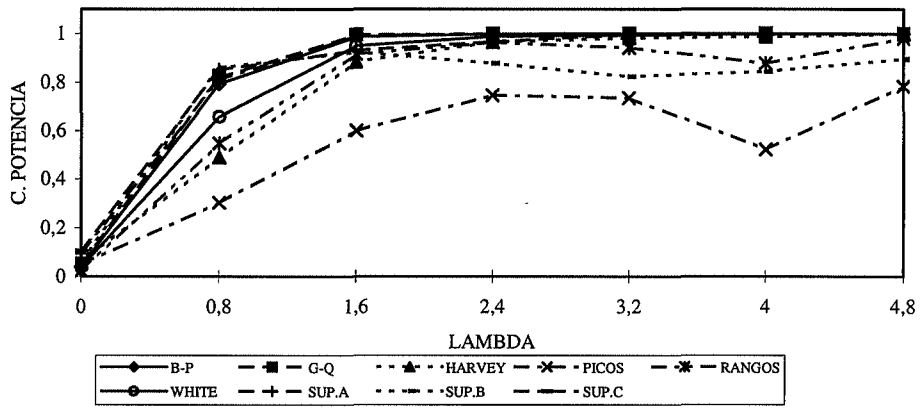
Se han eliminado todos los tests generales (excepto el de Goldfeld-Quandt), lo que viene a significar que estos contrastes son menos potentes que aquellos con formas funcionales específicas de heterocedasticidad; no obstante, hay que recalcar que los tests

no específicos cometen menos error de tipo I que los tests específicos. Centrándonos en el Gráfico 3, se observa que el test para cuando $\sigma^2 = a + b \cdot X$ parece dominar sobre los demás en un amplio espectro de valores de λ. Por otro lado, en caso de homocedasticidad, ocurre que la probabilidad de error de tipo I para el test de Goldfeld-Quandt es aproximadamente 0.05 (el nivel de significación).

Para analizar la probabilidad de error de tipo II del test de Goldfeld-Quandt, realicemos un zoom de la zona comprendida entre $\lambda = 0.6$ y $\lambda = 1.12$ del Gráfico 3. El resultado es el Gráfico 4 y en él se puede ver que la probabilidad de rechazar la hipótesis nula, siendo falsa, es muy elevada llegando al 99.9% que corresponde a una probabilidad de error de tipo II igual a 0.1. En definitiva, para valores extremos de λ y cuando la verdadera estructura de la varianza es $\sigma^2 = (a + b \cdot X)^2$, el test de Goldfeld-Quandt supone un contraste bastante fiable y eficaz.

3.1.3. Supuesto $\sigma^2 = \exp(a + b \cdot X)$

En el Gráfico 5 se muestran las curvas de potencia de los contrastes de heterocedasticidad. Lo primero que destaca es el excelente comportamiento de algunos test, llegándose a alcanzar probabilidades del 100% en la detección de la heterocedasticidad a partir de valores de $\lambda = 1.6$, lo que equivale a una probabilidad de error de tipo II nulo. Hay que destacar el hecho de que dentro de este grupo se encuentre el del supuesto A y también el de GQ. Otro aspecto significativo es que el tests de rangos que es no específico se muestre bastante eficaz, superando incluso al del supuesto B.



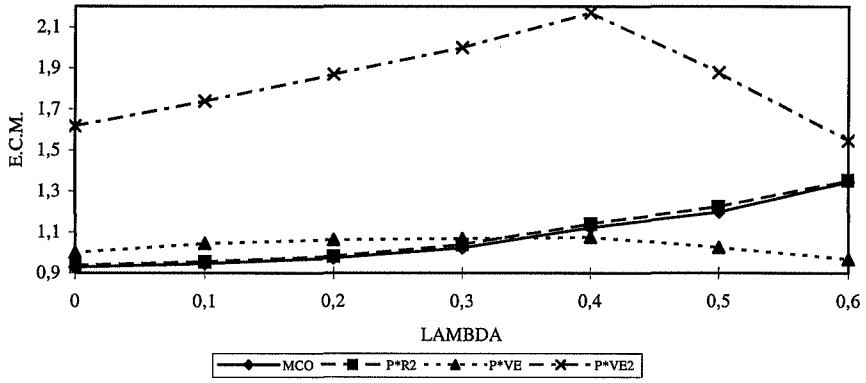
Fuente: Elaboración propia.

Gráfico 5. Curvas de potencia (Sup. C)

3.2. Estudio de la función de riesgo (E.C.M.) de los estimadores

3.2.1. Supuesto $\sigma^2 = a + b \cdot X$

A partir de las curvas del E.C.M. de los distintos estimadores pretest es posible obtener un gráfico que refleje el comportamiento característico del E.C.M. de todos ellos, debido a que los resultados obtenidos son muy similares entre pretests que contrastan el mismo contraste, debiéndose las diferencias al procedimiento de estimación. Así, hemos denominado *P*R2*, *P*VE* y *P*VE2* a los pretests que combinan cualquier contraste con la estimación *R2*, *VE* y *VE2*, respectivamente. Dicho gráfico es el siguiente:



Fuente: Elaboración propia.

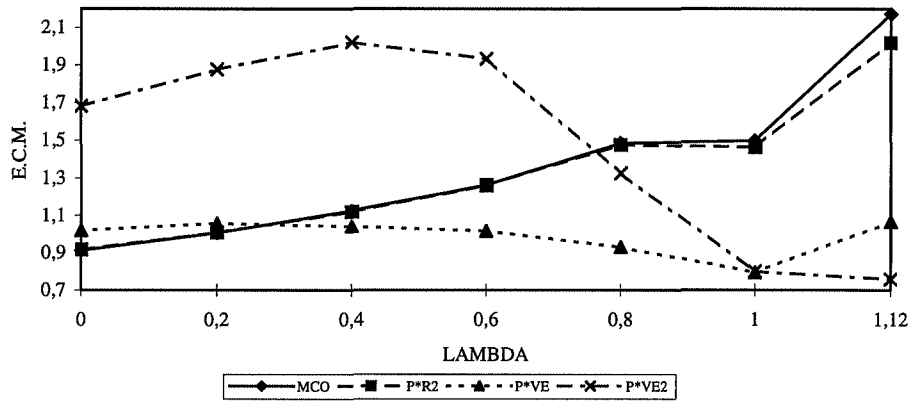
Gráfico 6. Error cuadrático medio de pretest (Sup. A)

El estimador $P * VE2$ ofrece un E.C.M. muy superior al del resto de los estimadores pretest para todos los valores ensayados de λ, por tanto en la búsqueda del estimador pretest óptimo, *P*VE2* queda descartado para el supuesto que se está considerando. Por otro lado, se puede observar que el E.C.M. del estimador *P*R2* tiene un comportamiento muy similar al del estimador por M.C.O.

Si se realiza un análisis más general del Gráfico 6 se observa que existe una primera zona de baja e intermedia heterocedasticidad donde la estimación por M.C.O. y por *P*R2* es óptima, sin embargo cuando la presencia de heterocedasticidad se acentúa lo mejor sería ir hacia un estimador como el $P * VE$, lo cual está en consonancia con el supuesto que se está tratando. Si ante la incertidumbre sobre el grado de heterocedasticidad, hubiera que decidirse por una sola forma de estimación a lo largo de todo λ, quizás lo más conveniente sería utilizar el estimador $P * VE$ debido a que su comportamiento es más homogéneo: en casos de máxima heterocedasticidad comete el mínimo E.C.M. y con poca heterocedasticidad su comportamiento, aunque no alcanza el de los M.C.O., podemos calificarlo como aceptable.

3.2.2. *Supuesto* $\sigma^2 = (a + b \cdot X)^2$

Para este supuesto, el gráfico característico que resume el comportamiento del E.C.M. de todos los pretests es el Gráfico 7. El E.C.M. que comete el estimador $P * VE2$ continúa siendo superior al del resto de los estimadores pretest, no obstante en este supuesto a partir de un determinado valor de λ el E.C.M. disminuye drásticamente, hasta tal punto que en términos de máxima heterocedasticidad dicho estimador es el que posee el mínimo E.C.M. Por otro lado, el E.C.M. del estimador $P * R2$ vuelve a tener un comportamiento muy similar al del estimador por M.C.O. El estimador $P * VE$ es el que tiene mayor bondad para un amplio intervalo de valores de λ, por tanto ante la incertidumbre sobre el grado de heterocedasticidad, dicho estimador sería el más adecuado ya que se comporta de manera más uniforme



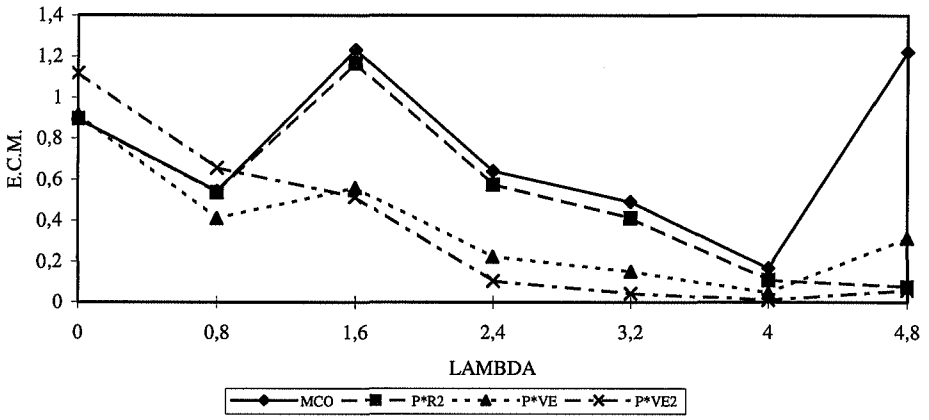
Fuente: Elaboración propia.

Gráfico 7. Error cuadrático medio de pretest (Sup. B)

Un estudio más amplio del Gráfico 7 nos lleva a diferenciar tres zonas. La primera, comprendida entre $\lambda = 0$ y $\lambda = 0.3$, en la que los estimadores óptimos son $P * R2$ y M.C.O. Una segunda zona entre $\lambda = 0.3$ y $\lambda = 1$, en la que domina el estimador $P * VE$. La última zona abarca el caso extremo de heterocedasticidad y el estimador $P * VE2$ resulta ser óptimo, siendo ésto lo que se debiera esperar teniendo en cuenta el supuesto de heterocedasticidad con el que se está trabajando.

3.2.3. *Supuesto* $\sigma^2 = \exp(a + b \cdot X)$

El Gráfico 8 es el representativo del comportamiento del E.C.M. de los pretests para el supuesto que se está considerando. El estimador $P * R2$ continua teniendo un comportamiento similar al estimador por M.C.O.

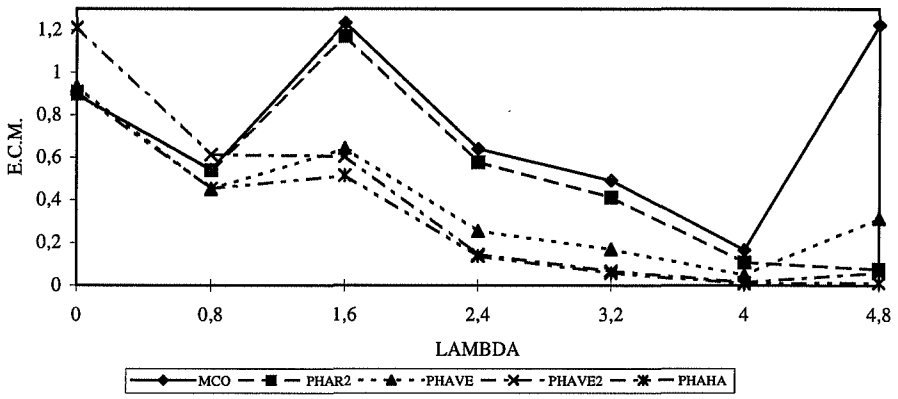


Fuente: Elaboración propia.

Gráfico 8. Error cuadrático medio de pretest (Sup. C)

Un aspecto interesante es la evolución del E.C.M. del estimador $P * VE2$. Para valores pequeños de λ tiene el mayor valor de la función de riesgo entre todos los estimadores y conforme aumenta la gravedad de la heterocedasticidad su E.C.M. disminuye, hasta que a partir de $\lambda = 1.6$, aproximadamente, se comporta como el estimador óptimo. Cuando la gravedad de la heterocedasticidad es muy leve el estimador $P * VE$ sería el más deseable y, cuando estamos en términos de heterocedasticidad media y extrema, es el estimador $P * VE2$ el que ofrece el mínimo valor en la función de riesgo.

A partir del test Harvey se ha obtenido el estimador pretest *PHAHA*. El comportamiento de su E.C.M. aparece reflejado en el siguiente gráfico.

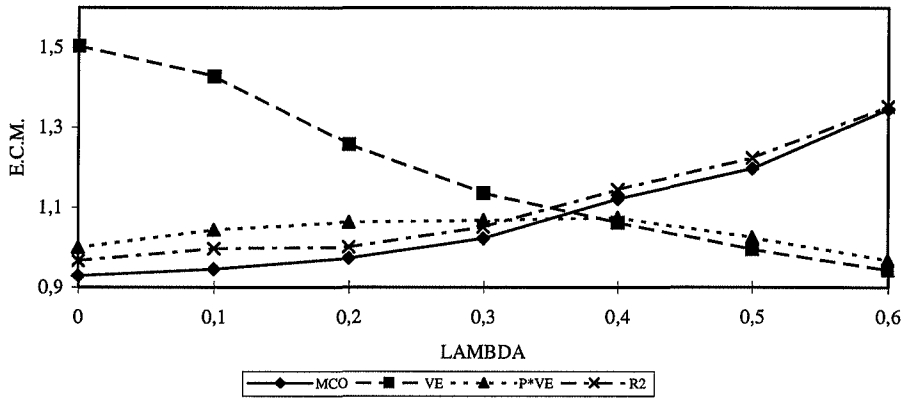


Fuente: Elaboración propia.

Gráfico 9. Error cuadrático medio de los pretest que utilizan el contraste de Harvey (Sup. C)

Como se puede comprobar, la utilización del estimador *PHAHA* nos asegura, sea cual sea el valor de λ, que estamos realizando la mejor estimación. Por tanto, independientemente de la gravedad de la heterocedasticidad, hemos encontrado un estimador óptimo para el supuesto que se está considerando.

Para completar el estudio del E.C.M. de los estimadores pretest, se van a analizar una serie de gráficos que se consideran de interés. El primero de ellos es el Gráfico 10, obtenido a partir del supuesto $\sigma^2 = a + b \cdot X$.

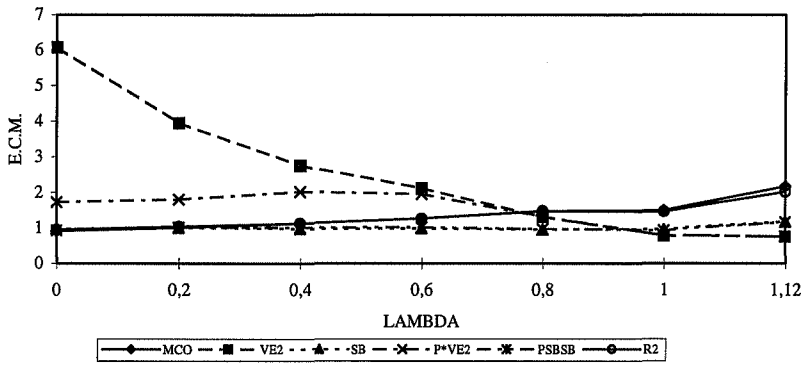


Fuente: Elaboración propia.

Gráfico 10. Comparación del error cuadrático medio de estimadores y pretest (Sup. A)

En dicho gráfico se representan las curvas del E.C.M. del estimador por M.C.O., de los estimadores por M.C.G.F. *VE* y *R2*, y del estimador pretest $P * VE$. Se puede comprobar que la curva del primero de ellos es ascendente con la gravedad de la heterocedasticidad; ésto es lo que se puede esperar del estimador por M.C.O. Es decir, este estimador resulta muy bueno en términos de homocedasticidad, incluso con poca heterocedasticidad, pero a medida que ésta aumenta, su E.C.M aumenta considerablemente. La evolución del E.C.M. del estimador *R2* es bastante similar a la del estimador por M.C.O., por tanto, cuando la heterocedasticidad es considerable el procedimiento de estimación mediante el cuadrado de los residuos es tan malo como el realizado por los M.C.O.

La curva que sigue el E.C.M. del estimador *VE* es descendente con la intensidad de la heterocedasticidad, es decir, *VE* alcanza sus mejores propiedades conforme la heterocedasticidad se acentúa. El estimador pretest $P * VE$ tiene características intermedias entre las del estimador por M.C.O. y las de *VE*, desde el punto de vista de la función de riesgo.



Fuente: Elaboración propia.

Gráfico 11. Comparación del error cuadrático medio de estimadores y pretest (Sup. B)

Observando el Gráfico 10 es fácil concluir que el estimador pretest $P * VE$ tiene un comportamiento más homogéneo que los otros dos (M.C.O. y *VE*), por tanto, si se desconoce la gravedad de la heterocedasticidad el estimador pretest sería preferentemente utilizado.

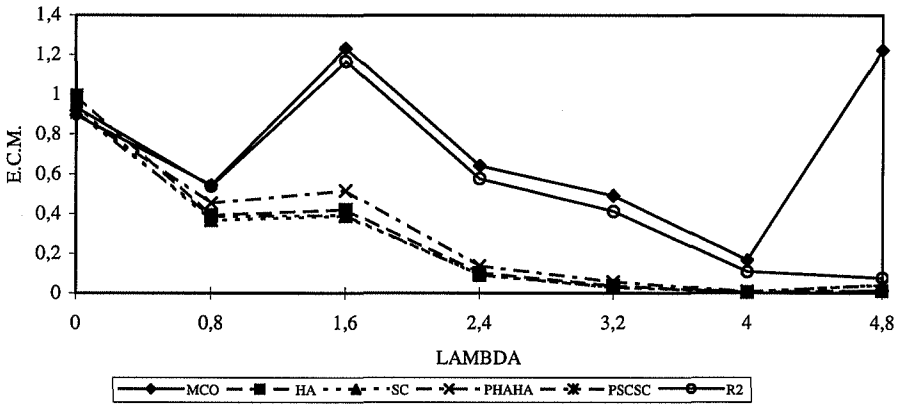
A partir del supuesto $\sigma^2 = (a + b \cdot X)^2$ se construye el Gráfico 11. En él se representan las curvas del E.C.M. de los siguientes estimadores: M.C.O., *VE2*, *SB*, $P * VE2$, *PSBSB* y *R2*. La curva del primero de ellos es ascendente con la heterocedasticidad y, como ya se ha dicho con anterioridad, es lo que se podía esperar de dicho estimador. Lo mismo ocurre con el estimador *R2*, así que también para este supuesto ($\sigma = a + b \cdot X$) la estimación según *R2* sería ineficiente, en presencia de heterocedasticidad.

La curva del E.C.M. del estimador *VE2* actúa justo al revés, es decir, es descendente con la heterocedasticidad; lo cual también es lógico si se tiene en cuenta que el supuesto considerado es $\sigma^2 = (a + b \cdot X)^2$. El estimador pretest $P * VE2$ tiene propiedades intermedias entre las del estimador por M.C.O. y las de *VE2*. Debido al comportamiento más uniforme que presenta $P * VE2$, se hace preferible su utilización sobre los otros dos en caso de incertidumbre sobre los posibles valores de λ.

Los estimadores *SB* y *PSBSB* ofrecen, para todo λ, un E.C.M. muy pequeño y bastante uniforme, de manera que serían los mejores estimadores (para el supuesto considerado), superando incluso al estimador $P * VE2$.

Por último, a partir del supuesto $\sigma^2 = \exp(a + b \cdot X)$ se construye el Gráfico 12 que contiene al estimador por M.C.O. y a los estimadores *R2, HA, SC, PHAHA* y *PSCSC*. Se puede comprobar que el E.C.M. de estos cuatro últimos estimadores desciende de manera considerable conforme la heterocedasticidad se hace más grave. Estos resulta-

dos se entienden fácilmente desde el punto de vista del supuesto considerado, ya que dichos estimadores provienen del test de Harvey y del test específico para este supuesto. Aún así, hay que destacar las propiedades tan excelentes (medidas en términos del E.C.M.) que ofrecen estos estimadores, especialmente cuando la heterocedasticidad es máxima.



Fuente: Elaboración propia.

Gráfico 12. Comparación del error cuadrático medio de estimadores y pretest (Sup. C)

4. CONCLUSIONES

En el presente trabajo se recoge una investigación sobre el comportamiento de algunos pretest de heterocedasticidad. Para ello se ha realizado un experimento Monte Carlo La parte sistemática del experimento se basa en una estructura estimada sobre datos reales de un modelo de consumo alimentario que resultó altamente heterocedástico. La perturbación elatoria se ha generado con el fín de introducir el problema de la heterocedasticidad como fuente de variación, diseñando a tal efecto tres supuestos distintos de estructura para la matriz de covarianzas del error. Para analizar la incidencia de la gravedad del problema, se han introducido siete niveles distintos de la misma, creando una metodología específica para ello.

Se exponen a continuación las conclusiones que se pueden extraer de los resultados expuestos. Dichas conclusiones entendemos que se pueden extender a modelos econométricos con una variable explicativa y un tamaño de muestra no muy distinto de T=50, que son los factores fijos con los que hemos trabajado por el momento, sin que ello suponga que no analicemos otros en un futuro próximo. En cuanto al modelo base

tomado para la parte sistemática, creemos que no supone gran perdida de generalidad, teniendo en cuenta además que las medidas de riesgo se calculan en términos relativos.

Concluimos, por tanto, lo siguiente:

1. Lo que define la potencia de un test de heterocedasticidad es la gravedad de ésta, siendo el proceso generador de datos poco relevante.
2. En términos generales, el test de Goldfeld-Quandt es bastante bueno, presentando un buen comportamiento tanto en caso de homocedasticidad como de heterocedasticidad.
3. Si se sospecha la existencia de heterocedasticidad, los tests más potentes son los tres específicos para cada uno de los supuestos sobre la estructura de la varianza del error y el de Goldfeld-Quandt.
4. Si se piensa que es muy poco probable que haya heterocedasticidad, los tests que mejor se comportan son los no paramétricos, es decir, los tests de Picos y Spearman.
5. El test de White sólo es eficiente cuando la gravedad de la heterocedasticidad es extrema. Los malos resultados obtenidos por este test pueden deberse a que en el experimento sólo se ha considerado una variable explicativa y el test de White está pensado para modelos con varias variables exógenas.
6. Desde el punto de vista del E.C.M., el comportamiento de los estimadores pretest depende sobre todo del procedimiento de estimación y en muy pequeña medida del test correspondiente.
7. En términos generales, el estimador pretest resultante de aplicar cualquier test con el procedimiento de estimación $R2$, que podemos representar en forma genérica por $P * R2$, se comporta de manera similar al estimador por M.C.O.
8. Cuando la heterocedasticidad no excede de $\lambda = 4$, el comportamiento de los estimadores pretest representados por $P * VE$ es bastante bueno, con la ventaja adicional de su facilidad de cálculo.
9. Los estimadores pretest $P * VE2$ no muestran su eficiencia hasta llegar a valores de $\lambda = 2.4$.
10. Los estimadores pretest concretos *PHAHA* y *PSCSC* son bastante buenos.

A modo de resumen final se enumeran las siguientes conclusiones:

1. Si se sospecha la existencia de heterocedasticidad grave, los estimadores pretest representados por $P*VE$ ofrecen un buen comportamiento y, dentro de éstos, *PHAVE* y *PSCVE* son los estimadores óptimos.

2. Si se piensa que la heterocedasticidad es muy leve o nula, los estimadores *PPVE* y *PSVE* son los mejores.

3. Si no se conoce nada sobre la estructura de la varianza del error lo más idóneo sería utilizar el estimador pretest *PG-QVE*, resultado de combinar el test de Goldfeld-Quandt con el procedimiento de estimación *VE*.

4. Queda patente la superioridad de los estimadores pretest sobre el resto de los estimadores que utilizan siempre el mismo procedimiento de estimación.

5. AGRADECIMIENTOS

Los autores agradecen las oportunas sugerencias de dos evaluadores anónimos.

6. REFERENCIAS

Breush, T.S. y Pagan, A.R. (1979).«A Simple Test for Heteroscedasticity and Ramdom Coefficient Variation», *Econometrica*, 47, 1287-1294.

Carroll, R.J. y Ruppert, D. (1981). «On Robust Tests for Heteroscedasticity», *Annals of Statistics*, 9, 205-209.

Fomby, T., Carter, R. y Johnson, S. (1984). *Advanced Econometric Methods*. Ed. Springer-Verlag. New York.

Glejser, H. (1969). «A New Test for Heteroscedasticity», *Journal of the American Statistical Association*, 64, 316-323.

Godfrey, L.G. (1978). «Testing for Multiplicative Heteroscedasticity», *Journal of Econometrics*, 8, 227-236.

Goldfeld, S.M. y Quandt, R.E. (1965). «Some Tests for Homoscedasticity», *Journal of the American Statistical Association*, 60, 539-547.

Harrison, M.J. y Mc Cabe, B.P.M. (1979). «A Test for Heteroscedasticity Based on Ordinary Least Squares Residuals», *Journal of the American Statistical Association*, 74, 494-499.

Harvey, A.C. (1976). «Estimating Regression Models with Multiplicative Heteroscedasticity», *Econometrica*, 44, 461-465.

Harvey, A.C. y Phillips, G.D. A. (1974). «A Comparison of the Power of Some Test for Heteroscedasticity in the General Linear Model», *Journal of Econometrics*, 2, 307-316.

Hausman, J.A. (1978). «Specification Tests in Econometrics», *Econometrica*, 46, 1251-1271.

Hendry, D.F. y Harrison, R.W. (1974). «Monte Carlo Methodology and the Finite Sample Behaviour of Ordinary and Two-Stage Least Squares», *Journal of Econometrics*, 2, 151-174.

Kloeck, T. y Van Dijk, H.K. (1978). «Bayesian Estimates of Ecuation System Parameters Application of Integration by Monte Carlo», *Econometrica*, 46, 1-19.

Mikhail, W.M. (1972). «Simulating the Small Sample Properties of Econometric Estimators», *Journal of the American Statistical Association*, 67, 620-624.

Mikhail, W.M. (1975). «A Comparative Monte Carlo Study of the Properties Econometrics Estimators», *Journal of the American Statistical Association*, 70, 91-104.

Mizon, G.E. y Hendry, D.F. (1980). «An Empirical Application and Monte Carlo Analysis of Dynamic Specification», *Review of Economic Studies*, 47, 21-45.

Sowey, E.R. (1973). «A Classified Bibliography of Monte Carlo Studies in Econometrics», *Journal of Econometrics*, 1, 377-395.

Spearman, C. (1904). «The Proof and Measurement of Association Between Two Things», *Am. J. Psychol.*, 15, 72-101.

Surekha, K. y Griffiths, W.E. (1984). «A Monte Carlo Comparison of Some Bayesian and Sampling Theory Estimators in Two Heteroscedastic Error Models», *Communications in Statistics B*, 13, 85-105.

White, H. (1980). «A Heteroscedastic-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity», *Econometrica*, 48, 817-838.

ANEXO

Tabla 1. Datos del modelo base en millones de pesetas (1990)

PROVINCIA	*GASTO TOTAL EN PRODUCTOS AGROALIMENTARIOS*	*RENTA FAMILIAR DISPONIBLE*
Alava	47108.6	307279
Albacete	51274.2	284369
Alicante	192572.5	1341225
Almería	68196.3	393382
Asturias	179565.6	1051136
Avila	28770.7	152069
Badajoz	76626.5	489155
Baleares	102282.2	867525
Burgos	61087.6	369639
Cáceres	59097.3	321460
Cádiz	155863.6	822891
Cantabria	79491.6	504772
Catellón	64343.7	475965
Ciudad Real	71549.1	398585
Córdoba	108998.9	602789
La Coruña	169989.5	1014977
Cuenca	31025.5	177677
Gerona	90795.4	727271
Granada	112631.9	606935
Guadalajara	23202.3	148198
Guipúzcoa	115635.9	705003
Huelva	63196.9	372708
Huesca	32567.8	211969
Jaén	91802	518609
León	82076.1	481207
Lérida	60650.6	389255
Lugo	61915.8	335107
Málaga	174493.7	1017954
Murcia	152302.1	967480
Navarra	86427.1	562873
Orense	66676.6	349990
Palencia	28967	183464
Las Palmas	106168.8	727844
Pontevedra	146753.9	856294
La Rioja	42082.9	306548
Salamanca	52498.8	310899
Sta. Cruz de Tenerife	92494.6	672427
Segovia	22568.7	138120
Sevilla	219127.2	1346684
Soria	19567.4	88928
Tarragona	77295.8	615884
Teruel	22905.7	144208
Toledo	76107.4	444165
Valladolid	73758.4	490250
Vizcaya	190063.6	1105062
Zamora	33169.5	203736
Zaragoza	138155.2	903283

Fuente: Instituto Nacional de Estadística y Banco Bilbao-Vizcaya.

Tabla 2. Valores de a, b y λ para el primer supuesto

a	b	λ	NIVEL
550000	0	0	0
455000	0.17	0.1	1
360000	0.34	0.2	2
270000	0.51	0.3	3
180000	0.67	0.4	4
90000	0.84	0.5	5
0	1	0.6	6

Fuente: Elaboración propia

Tabla 3. Valores de a, b y λ para el segundo supuesto

a	b	λ	NIVEL
742	0	0	0
615	2.3e-4	0.2	1
485	4.67e-4	0.4	2
365	6.85e-4	0.6	3
235	9.21e-4	0.8	4
110	1.15e-3	1	5
0	1.35e-3	1.12	6

Fuente: Elaboración propia

Tabla 4. Valores de a, b y λ para el tercer supuesto

a	b	λ	NIVEL
13	0	0	0
11.9	2e-6	0.8	1
11.1	3.45e-6	1.6	2
10.35	4.81e-6	2.4	3
9.4	6.54e-6	3.2	4
8	9.09e-6	4	5
0	2.36e-5	4.8	6

Fuente: Elaboración propia

Tabla 5. Nomenclatura de los estimadores por M.C.G.F.

Diagonal de $\hat{\Omega}$	*Estimadores por M.C.G.F.*
elementos de $z_i' \hat{\alpha}$	*SA*
elementos de $(z_i' c \hat{\alpha})^2$	*SB*
elementos de $\exp(z_i' \alpha^*)$	*SC*
residuos por M.C.O. al cuadrado	*R2*
variable exógena	*VE*
variable exógena al cuadrado	*VE2*
estimación de la variable estimada del modelo del test de Breush-Pagan	*B-P*
estimación de la variable estimada del modelo del test de Harvey	*HA*
estimación de la variable estimada del modelo del test de White	*W*

Fuente: Elaboración propia

Tabla 6. Nomenclatura de los estimadores pretest

Contraste	*Proced. estimación*	*Estimador pretest*
Test para cuando $\sigma_i^2 = z_i' \alpha$	*R2*	*PSAR2*
	VE	*PSAVE*
	VE2	*PSAVE2*
	W	*PSAW*
	SA	*PSASA*
Test para cuando $\sigma_i = z_i' \alpha$	*R2*	*PSBR2*
	VE	*PSBVE*
	VE2	*PSBVE2*
	W	*PSBW*
	SB	*PSBSB*
Test para cuando $\sigma_i^2 = \exp(z_i' \alpha)$	*R2*	*PSCR2*
	VE	*PSCVE*
	VE2	*PSCVE2*
	W	*PSCW*
	SC	*PSCSC*
Test de Breush-Pagan	*R2*	*PB-PR2*
	VE	*PB-PVE*
	VE2	*PB-PVE2*
	W	*PB-PW*
	BP	*PB-PB-P*
Test de Goldfeld-Quandt	*R2*	*PG-QR2*
	VE	*PG-QVE*
	VE2	*PG-QVE2*
	W	*PG-QW*
Test de Harvey	*R2*	*PHAR2*
	VE	*PHAVE*
	VE2	*PHAVE2*
	W	*PHAW*
	HA	*PHAHA*
Test de Picos	*R2*	*PPR2*
	VE	*PPVE*
	VE2	*PPVE2*
	W	*PPW*
Test de Spearman	*R2*	*PSR2*
	VE	*PSVE*
	VE2	*PSVE2*
	W	*PSW*
Test de White	*R2*	*PWR2*
	VE	*PWVE*
	VE2	*PWVE2*
	W	*PWW*

Fuente: Elaboración propia

ENGLISH SUMMARY

A COMPARATIVE ANALYSIS OF HETEROCEDASTICITY PRETEST ESTIMATORS IN ECONOMETRICS MODELS. A MONTE CARLO STUDY

R. DIOS PALOMARES
C. RODRÍGUEZ FONSECA
Universidad de Córdoba*

In this paper a Monte Carlo experiment has been designed in order to analyze the performance of the tests for heteroscedasticity and the ordinary least squares, estimated generalized least squares and preliminary test estimators in the presence of heteroscedasticity. The data generation process considered in the experiment is a model based on three structures of the variance of the error term with different levels of heteroscedasticity for each one. Different estimation procedures are analyzed and the Monte Carlo power of the tests for heteroscedasticicity and the efficiency of the different estimators are computed. The estimators are compared using the risk function criterion. The results show that the pretest estimators perform better than the estimators which always use the same estimation procedure.

Keywords: Monte Carlo method, heteroscedasticity, power, generalized least squarees, pretest estimator

AMS Classification: 62J05 , 62J20

JEL Classification: C12, C52

* Universidad de Córdoba. Departamento de Estadística, Econometría, Investigación Operativa y Organización de Empresas. Avda. Menéndez Pidal, s/n. 14004 Córdoba.

- Received March 1998.
- Accepted June 1999.

In this paper a Monte Carlo experiment has been designed in order to analyze the performance of the tests for heteroscedasticity and the ordinary least squares, estimated generalized least squares and preliminary test estimators in the presence of heteroscedasticity.

In the general linear model $\boldsymbol{y} = X\beta + \boldsymbol{e}$, the error vector e has a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix Ω. If $\Omega = \sigma_e^2 \boldsymbol{I}_T$ we would be in presence of a homoscedastic model, but in the practice there are cases in those that the homoscedasticity hypothesis can not be accepted since the variance-covariance matrix for $\boldsymbol{e}$ is

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdot & \cdot & 0 \\ 0 & \sigma_2^2 & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & \sigma_T^2 \end{pmatrix}$$

that is, the diagonal elements of Ω are not identical and, therefore, we are in the presence of a heteroscedastic model.

In the presence of a possible heteroscedastic error model the questions of interest are how to test for heteroscedasticity, how to model heteroscedasticity, and how to estimate the model optimaly.

Methods have been developed to test for heteroscedasticity. Their power will depend on the true structure of the heteroscedasticity, among another things.

Tests for heteroscedasticity frequently used can be broken down into two groups. The first one is formed by specific tests for each of the following assumptions about the variance structure of the disturbance term (Fomby, 1984): $\sigma_i^2 = z_i' \boldsymbol{\alpha}$, $\sigma_i^2 = (z_i' \boldsymbol{\alpha})^2$ and $\sigma_i^2 = \exp(z_i' \boldsymbol{\alpha})$. The second group consists of the general or nonspecific tests. By nonspecific, we mean tests which are designed to detect heteroscedasticity yet the investigator does not posses a priori knowledge of a specific form of heteroscedasticity. Some tests that have been suggested for this purpose are: Breush-Pagan test (1979), Goldfeld-Quandt test (1965), Harvey test (1976), Peak test (Goldfeld and Quandt, 1965) and the Range Correlation test (Spearman, 1904). The power of some of these tests has been questioned by several authors: Carrol and Ruppert (1981), Glejser (1969), Godfrey (1978), Harrison and Mc Cabe (1979), Harvey and Phillips (1974) and Hausman (1978).

The parameter estimates can be obtained using:

a) Ordinary Least Squares (O.L.S.) estimators $\hat{\beta} = (X'X)^{-1} X' \boldsymbol{y}$. If the model is homoscedastic then they are the best estimators.

b) Estimated Generalized Least Squares (E.G.L.S.) estimators $\tilde{\hat{\beta}} = (X' \hat{\Omega}^{-1} X)^{-1} X' \hat{\Omega}^{-1} \boldsymbol{y}$. If matrix Ω is properly estimated then these estimators are the best.

Estimated Generalized Least Squares (E.G.L.S.) estimation procedures can be divided in two groups, those dealing with specific functional forms of heteroscedasticity (Fomby, 1984) and those addressing unspecified functional forms, known as general estimation procedures. These general estimators use different methods to estimate matrix Ω. Though the large sample properties of the estimate generalized least squares estimators in heteroscedasticity models are well defined, their small sample properties are not.

This problem can be framed in the context of the pretest estimation. We define a preliminary-test estimator of heteroscedasticity as an estimator which involves both OLS and EGLS estimators. The choice between the ordinary least squares estimator and the estimated generalized least squares estimator is made on the basis of a test for heteroscedasticity, and the resulting estimator is a preliminary test estimator whose sampling properties are distinct from either those of OLS estimator or EGLS estimator. Thus, the pretest estimator is equal to OLS if homoscedasticity is accepted, and EGLS if homoscedasticity is rejected.

Thus, a test plus an estimation procedure for Ω define a pretest of heteroscedasticity. The pretest estimator is unbiased and its variance is between that of ordinary least squares and estimated generalized least squares.

The mean squared error or risk matrix is a criterion that may be used to measure the performance of the pretest estimators. It may be defined as

$$R(\beta, \hat{\hat{\beta}}) = \text{cov}(\hat{\beta}) + (\text{bias}\,\hat{\hat{\beta}})(\text{bias}\,\hat{\hat{\beta}})'$$

Likewise, the risk can be analyzed on the basis of severity of the heteroscedasticity in the Data Generating Process (D.G.P.). If a parameter denoted λ expresses this severity the question of interest is to know the performance of the pretest estimator over the whole range of the parameter space of λ. In order to know this performance we use the following function

$$R(\lambda, \beta, \hat{\hat{\beta}}) = E\left[(\hat{\hat{\beta}}_{(\lambda)} - \beta)(\hat{\hat{\beta}}_{(\lambda)} - \beta)'\right] + \left[E(\hat{\hat{\beta}}_{(\lambda)}) - \beta\right]\left[E(\hat{\hat{\beta}}_{(\lambda)}) - \beta\right]'$$

which only can be obtained when λ and β are known, such in a Monte Carlo simulation.

Monte Carlo experiments have been used to evaluate the new methods from a theorical point of view. Thus, Mikhail (1972) designs a Monte Carlo experiment to investigate the small sample properties of econometric estimators. Sowey (1973) presents a chronological classified bibliography of Monte Carlo studies in Econometrics. Hendry and Harrison (1974) study the finite sample behavior of ordinary and two-stage least esquares. Kloek and van Dijk (1978) discuss Monte Carlo integration for bayesian

estimation. Mizon and Hendry (1980) report a Monte Carlo analysis of tests of dynamic specification. Surekha and Griffiths (1984) investigate the behavior of bayesian estimators assuming two different heteroscedastic error models. Gonzalo (1994) and Stock and Watson (1993), among others, addressed the issue of efficient estimation of the cointegration vectors with the Monte Carlo model based on three structures of the variance of the error term with different levels of heteroscedasticity for each one.

The data generation process considered in the present experiment is a model based on three structures of the variance of the error term with different levels of heteroscedasticity for each one.

Different estimation procedures are analyzed and the Monte Carlo power of the tests for heteroscedasticicity and the efficiency of the different estimators are computed.

The estimators are compared using the risk function criterion. The results show that the pretest estimators perform better than the estimators which always use the same estimation procedure.

The main conclusions are the following:

- When heteroscedasticity exists, the variance of the EGLS estimators is not always smaller than that of the LS estimators, but it depends on the success at choosing the matrix $\hat{\Omega}$
- The power of a test for heteroscedasticity depends on the severity of it, and the DGP is hardly relevant.
- From the point of view of the risk function, the performance of the pretest estimators depends on the estimation procedure whereas the corresponding test has a little influence.

Finally, we conclude the abstract with the following remarks:

- If one suspects heteroscedasticity is severe, the pretest estimators denoted $P * EV$ performs well, which arises of combining any test with the estimation procedure denoted *EV*.
- If one believes that heteroscedasticity does not exist or is mild, the simple pretest estimators denoted *PPEV* and *PSEV* are the best estimators.
- If one does not know anything about the structure of the variance of the error term, the best thing is to use the pretest estimator denoted *PGQEV*, which arises of combining the Goldfeld-Quandt test with the estimation procedure denoted *EV*.
- It is clear that the pretest estimators perform better than the estimators which always use the same estimation procedure.

QÜESTIIÓ, vol. 23, 3, p. 465-488, 1999

INDEPENDENCIA ENTRE LAS CUESTIONES EN EL ANÁLISIS FACTORIAL DE TABLAS DISYUNTIVAS INCOMPLETAS CON PREGUNTAS CONDICIONADAS

A. ZÁRRAGA
B. GOITISOLO
Universidad del País Vasco-Euskal Herriko Unibertsitatea*

El análisis de correspondencias múltiples (ACM) estudia la relación entre varias variables cualitativas definidas sobre una misma población. Sin embargo, una de las principales fuentes de información son las encuestas donde es frecuente encontrar cierto número de datos ausentes y de preguntas condicionadas. Escofier (Escofier 1981) propone analizar la tabla disyuntiva incompleta sustituyendo la marginal real de la tabla sobre los individuos por una marginal impuesta constante. El análisis de la tabla disyuntiva incompleta está asociado a unas tasas de inercia pequeñas que no deben ser interpretadas como partes de información explicada por los ejes. Se estudiará el caso en el cual las cuestiones son independientes dos a dos y se propondrá una corrección a estas tasas de inercia en el análisis con marginal modificada de una tabla disyuntiva incompleta.

Independence between questions in the factor analysis of incomplete disjunctive tables with conditioned questions

Palabras clave: Análisis de correspondencias múltiples, tabla disyuntiva incompleta, independencia entre variables cualitativas, valores propios, tasas de inercia

Clasificación AMS: 62H25

* Universidad del País Vasco-Euskal Herriko Unibertsitatea. Dpto. de Economía Aplicada III. Fac. CCEE y EE. Avda. Lehendakari Aguirre, 83. 48015 Bilbao. E-mails: az@alcib.bs.ehu.es y bg@alcib.bs.ehu.es.
Este trabajo ha sido financiado por el Proyecto de Investigación PB98-0149 de la Dirección General de Enseñanza Superior del Ministerio Español de Educación y Ciencia y el Proyecto UPV 038.321-HA041/99 de la Universidad del País Vasco (UPV/EHU).

- Recibido en febrero de 1998.
- Aceptado en julio de 1999.

1. INTRODUCCIÓN

Las razones por las que existen datos ausentes pueden ser diversas, revelando problemas que deberán ser tratados también de diferentes formas según se indica en §4 y en (Escofier & Pagès 1992).

- Las no respuestas pueden estar causadas por un olvido involuntario del individuo y no tener un significado especial. Suelen representar una proporción muy pequeña de los datos y afectar por igual a todas las cuestiones e individuos.
- Una segunda razón para la existencia de la no respuesta corresponde a una actitud particular del entrevistado, deseo de no revelar cierta información (por ejemplo, los ingresos, la ideología política, etc). Este tipo de no respuesta no se reparte de forma aleatoria en la tabla de datos sino que afecta más a determinadas cuestiones y grupos de individuos.
- Otra razón por la que aparecen las tablas disyuntivas incompletas —definidas en §2— muy frecuente en las encuestas se debe a la existencia de preguntas condicionadas, es decir, aquéllas a las cuales un individuo debe contestar o no dependiendo de cual haya sido su respuesta a una cuestión anterior. Por ejemplo, se le pregunta si sabe o no inglés; a continuación, y sólo si ha respondido saber inglés, se le pregunta su nivel de inglés en determinados aspectos. En una encuesta pueden existir varios grupos de preguntas condicionadas (los que saben inglés, francés, sólo los que tienen familiares y se relacionan con ellos contestarán la frecuencia con que lo hacen, etc). En este caso, la no respuesta se agrupa en un determinado número de cuestiones (aquellas cuya respuesta está condicionada por una pregunta anterior) y caracteriza a determinados grupos de individuos.

Tras la definición de las tablas disyuntivas incompletas y la notación básica utilizada —§2—, se verá el problema que plantea la aplicación del análisis de correspondencias múltiples clásico a este tipo de tablas —§3— y algunas posibles alternativas dependiendo de las razones de la ausencia —§4—. La solución propuesta para el caso de preguntas condicionadas (el análisis con marginal modificada) se desarrolla —§5 a §9— insistiendo en las diferencias con el análisis de correspondencias clásico (por ejemplo, en la elección del origen —§6— y número de ejes —§9—) y en los diferentes resultados según se posea una proporción pequeña de los datos ausentes o esta proporción sea elevada (relaciones entre los factores —§8—).

En §10 se estudia el caso de independencia entre las cuestiones y a partir de los resultados obtenidos se propone una corrección a las tasas de inercia calculadas en el análisis general con marginal modificada de la tabla disyuntiva incompleta.

2. DEFINICIÓN DE LAS TABLAS DISYUNTIVAS INCOMPLETAS Y NOTACIÓN

Se considera la tabla de datos que recoge en forma lógica y disyuntiva las respuestas de un conjunto de individuos a un conjunto de preguntas o cuestiones, poseyendo cada una de ellas un conjunto finito de modalidades de respuesta. En el análisis de correspondencias múltiples clásico se impone a todos los individuos la obligación de pertenecer a alguna de las modalidades de cada cuestión y se denomina a la tabla así obtenida tabla disyuntiva completa (Z). Se dirá que tal tabla de datos es disyuntiva incompleta (Z^*) cuando los individuos no dan respuesta a una o más de las cuestiones preguntadas.

Tabla 1

Tabla disyuntiva completa (Z) o Tabla disyuntiva incompleta (Z^*)

	$q=1$		...	q	...	$q=Q$
	$j=1$ ...	$j=J_1$		j		
1	1	0				
2	0	0				
3	0	1				
⋮						
i				z_{ij}		
⋮						
n						

donde:

$\mathcal{Q} = \{1, \ldots, q, \ldots, Q\}$ es el conjunto de variables a las cuales debe responder el individuo

$\mathcal{J}_q = \{1, \ldots, j, \ldots, J_q\}$ es el conjunto de modalidades de la variable $q \in \mathcal{Q}$

$\mathcal{J} = \{1, \ldots, j, \ldots, J\}$ es el conjunto de modalidades de todas las variables
$= \cup_{q=1}^{Q} \mathcal{J}_q$

$\mathcal{I} = \{1, \ldots, i, \ldots, n\}$ es el conjunto de individuos

$z_{ij} = \begin{cases} 1 & \text{si el individuo } i \in \mathcal{I} \text{ responde la modalidad } j \in \mathcal{J} \\ 0 & \text{en otro caso} \end{cases}$

$z_{i.} = \sum_{j \in \mathcal{J}} z_{ij}$ es el número de cuestiones a las que responde el individuo $i \in \mathcal{I}$

$z_{.j} = \sum_{i \in \mathcal{I}} z_{ij}$ es el número de individuos que eligen la modalidad $j \in \mathcal{J}$

Se denotará $z_{.j}^{q}$ cuando interese dejar constancia de la variable $q \in \mathcal{Q}$ a la que pertenece dicha modalidad

$z = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} z_{ij}$ es el total de la tabla

En las tablas disyuntivas incompletas –al igual que en las completas analizadas mediante el ACM clásico– las variables siguen estando definidas a través de un conjunto de modalidades a las que el individuo debe responder sobre su pertenencia ($z_{ij} = 1$) o no ($z_{ij} = 0$).

Pero, en ocasiones, esas modalidades correspondientes a una misma variable no están definidas en forma completa, es decir, el individuo puede no pertenecer a ninguna de ellas; en otras ocasiones a pesar de estar definidas en forma completa el individuo puede no revelar a que modalidad pertenece; en ambos casos:

$$z_{ij} = 0 \qquad i \in \mathcal{I} \qquad \forall j \in \mathcal{J}_q \quad q \in \mathcal{Q}$$

Por ello será necesario definir también una variable que toma el valor 1 si el individuo $i \in \mathcal{I}$ responde a la cuestión $q \in \mathcal{Q}$ y 0 en caso contrario:

$$z_{i.}^{q} = \sum_{j \in \mathcal{J}_q} z_{ij} \qquad \forall q \in \mathcal{Q} \qquad \forall i \in \mathcal{I}$$

Y se denotará por z_q el número de individuos que han respondido a la cuestión $q \in \mathcal{Q}$:

$$z_q = \sum_{j \in \mathcal{J}_q} z_{.j} \qquad \forall q \in \mathcal{Q}$$

En resumen, las tablas disyuntivas incompletas se caracterizan porque en ellas dejan de cumplirse algunas de las relaciones que se dan en las completas:

$$\begin{aligned} z_{i.}^{q} &= 1 && \forall q \in \mathcal{Q} && \forall i \in \mathcal{I} \\ z_q &= n && \forall q \in \mathcal{Q} \\ z_{i.} &= Q && \forall i \in \mathcal{I} \\ z &= nQ \end{aligned}$$

Como ya es sabido, en todo análisis de correspondencias se definen ((Escofier & Pagès 1992), (Lebart, Morineau & Tabard 1977) y (Abascal & Grande 1989) entre otros) las frecuencias relativas conjuntas y marginales como:

$$\begin{aligned} f_{ij} &= \frac{z_{ij}}{z} && \forall i \in \mathcal{I} && \forall j \in \mathcal{J} \\ f_{i.} &= \frac{z_{i.}}{z} = \sum_{j \in \mathcal{J}} f_{ij} && \forall i \in \mathcal{I} \\ f_{.j} &= \frac{z_{.j}}{z} = \sum_{i \in \mathcal{I}} f_{ij} && \forall j \in \mathcal{J} \end{aligned}$$

y los perfiles fila $i, i \in \mathcal{I}$:

$$\frac{z_{ij}}{z_{i.}} \qquad \forall j \in \mathcal{J}$$

y perfiles columna $j, j \in \mathcal{J}$:

$$\frac{z_{ij}}{z_{.j}} \qquad \forall i \in \mathcal{I}$$

que forman las nubes $\mathcal{N}(\mathcal{I}) \in \Re^J$ y $\mathcal{N}(\mathcal{J}) \in \Re^n$ respectivamente.

3. PROBLEMA QUE PLANTEA LA APLICACIÓN DEL ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES HABITUAL A UNA TABLA DISYUNTIVA INCOMPLETA

El problema que plantea este tipo de tablas es el mismo independientemente de la razón para esa ausencia de datos: la marginal sobre $\mathcal{I}$ ya no es constante.

Podría pensarse en aplicar directamente el análisis de correspondencias simples a esta tabla, (notar que originalmente se creó para el estudio de tablas de contingencia donde las marginales no son constantes). Sin embargo, cuando se posee una tabla disyuntiva incompleta, la distancia χ^2 y los pesos definidos en el análisis clásico no se ajustan a los objetivos, ya conocidos de un análisis de correspondencias.

La aplicación de la distancia χ^2 entre dos perfiles fila i e $i' \in \mathcal{I}$ sería:

$$\begin{aligned} d^2(i,i') &= \sum_{j\in\mathcal{J}} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 = \\ &= \sum_{j\in\mathcal{J}} \frac{z}{z_{.j}} \left(\frac{z_{ij}}{z_{i.}} - \frac{z_{i'j}}{z_{i'.}} \right)^2 \end{aligned} \tag{1}$$

Si los individuos i e $i' \in \mathcal{I}$ no contestan al mismo número de preguntas, entonces $z_{i.}, i \in \mathcal{I}$ difiere de $z_{i'.}, i' \in \mathcal{I}$ y por tanto la distancia χ^2 aumenta también con las respuestas comunes. Este es, por tanto, un concepto de distancia no deseable puesto que no reflejaría la similaridad entre individuos –en términos de modalidades comunes elegidas– buscada en un análisis de correspondencias.

La distancia χ^2 entre dos perfiles columna j y $j' \in \mathcal{J}$ sería:

$$\begin{aligned} d^2(j,j') &= \sum_{i\in\mathcal{I}} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 = \\ &= \sum_{i\in\mathcal{I}} \frac{z}{z_{i.}} \left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}} \right)^2 \end{aligned} \tag{2}$$

En esta distancia χ^2 cada individuo tendría una ponderación distinta dependiendo del número de respuestas elegidas. No parece lógico asignar menos importancia a aquellos individuos que responden a la totalidad de las preguntas frente a quienes no lo hacen.

Por tanto, la aplicación directa del análisis de correspondencias clásico no es adecuada al estudio de las tablas disyuntivas incompletas.

4. POSIBLES ALTERNATIVAS AL PROBLEMA DE LAS TABLAS DISYUNTIVAS INCOMPLETAS: ANÁLISIS FACTORIAL CON MARGINAL MODIFICADA

Una vez descartada la aplicación del análisis clásico, se debe buscar otra alternativa para el estudio de las tablas disyuntivas incompletas que se adapte mejor. Se busca un método que minimice la influencia de la no respuesta sobre el análisis.

Una solución evidente desde el punto de vista analítico sería eliminar aquellos individuos que no responden a todas las cuestiones, obteniendo de esta forma una tabla disyuntiva completa. Con esta alternativa perderíamos la información referente a esos individuos en el resto de las cuestiones; información que puede no ser importante cuando la no respuesta se debe a un descuido y por tanto representa una pequeña proporción sobre el total, pero que alteraría los resultados cuando revela una actitud particular o implica a un gran número de individuos (caso de las preguntas condicionadas).

Una práctica habitual es crear para cada variable con datos ausentes una modalidad de no respuesta, obteniendo de esta forma una tabla disyuntiva completa a la que se puede aplicar el análisis clásico. Esta solución puede ser adecuada cuando la no respuesta se debe a una actitud particular del individuo (deseo de no revelar determinada información por ejemplo); sin embargo, cuando la no respuesta se debe a un descuido involuntario la modalidad de no respuesta no tendría una interpretación adecuada y perturbaría los resultados.

En los casos en los que la no respuesta se debe a la existencia de preguntas condicionadas ya se ha indicado en la introducción que caracteriza a un grupo de individuos. A pesar de ello, la inclusión de una modalidad de no respuesta en cada cuestión no sería adecuada, puesto que se estaría creando una serie de modalidades todas ellas con el mismo perfil e idéntico a una de las modalidades de la pregunta condicionante (no saben inglés) o a una combinación lineal de ellas («no tienen familiares» y «tienen pero no se relacionan»). Esto podría perturbar los resultados hasta el punto de llegar a crear uno de los primeros ejes del análisis, como así ocurre en la aplicación a la Encuesta de Condiciones de Vida de 1989 de la Comunidad Autónoma de Euskadi presentada en (Goitisolo & Zárraga 1998a).

Escofier (Escofier 1981) propone sustituir la marginal real de la tabla ($f_{i.} = z_{i.}/z, i \in \mathcal{I}$ que no es constante), por una marginal constante $g_{i.} = 1/n, i \in \mathcal{I}$ en todo el análisis.

Posteriormente (Benali 1985), (Benali 1988), (Benali & Escofier 1987) y (Escofier 1990) utilizan también esta técnica, pero siempre para el caso de tablas disyuntivas

incompletas donde la no respuesta se debe a una omisión involuntaria y el caso de modalidades de efectivo débil. En el cálculo de la distancia entre dos individuos (y por tanto, en la inercia de la nube), las modalidades tienen una ponderación inversa a su efectivo. En consecuencia, las modalidades muy raras pueden influir demasiado; Benali y Escofier proponen eliminarlas y tratar a esos individuos como si no hubieran dado respuesta a la cuestión.

Se analizará en detalle lo adecuado de esta sustitución y las consecuencias sobre el análisis, tanto para los casos de omisión involuntaria y de modalidades raras como para el caso de preguntas condicionantes, en el que centraremos nuestro interés.

5. NUBE DE INDIVIDUOS: $\mathcal{N}(\mathcal{I})$

El punto $i \in \mathcal{I}$ se representa en $\Re^J$ por el perfil $\frac{f_{ij}}{g_{i.}} = n\frac{z_{ij}}{z}, i \in \mathcal{I}, j \in \mathcal{J}$. Este perfil es diferente del perfil obtenido en el análisis de correspondencias múltiples clásico $(z_{ij}/Q,\ i \in \mathcal{I}, j \in \mathcal{J})$ al ser el efectivo total de la tabla (z) distinto de nQ.

La distancia cuadrática propuesta entre dos individuos i e $i' \in \mathcal{I}$ es:

$$d^2(i,i') = \sum_{j\in\mathcal{J}} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{g_{i.}} - \frac{f_{i'j}}{g_{i'.}}\right)^2 = \tag{3}$$
$$= \frac{n^2}{z} \sum_{j\in\mathcal{J}} \frac{1}{z_{.j}} (z_{ij} - z_{i'j})^2$$

Se comprueba que únicamente las respuestas diferentes hacen aumentar la distancia, sin tener en cuenta si ambos individuos responden al mismo número de cuestiones o no.

La ponderación de cada modalidad es, al igual que en correspondencias múltiples con datos completos, el inverso de su efectivo (la distancia aumenta en mayor proporción cuando la modalidad poseída por sólo uno de los individuos es rara).

Considerar la distancia anterior entre los puntos i e $i' \in \mathcal{I}$ es equivalente a buscar la distancia euclídea habitual en un espacio dotado de métrica $1/f_{.j},\ j \in \mathcal{J}$.

Cada punto $i \in \mathcal{I}$ está dotado de un peso $g_{i.} = 1/n,\ i \in \mathcal{I}$, que a pesar de no venir representado por la marginal $f_{i.},\ i \in \mathcal{I}$ –del ACM clásico– coincide con el peso que se asigna a los individuos en correspondencias múltiples habitual. Este peso constante significa que todos los individuos tienen la misma importancia, independientemente del número de cuestiones que han respondido. Es por tanto, más adecuado que $f_{i.}, i \in \mathcal{I}$.

La coordenada j-ésima del centro de gravedad de la nube es:

$$G_I(j) = f_{.j} = \frac{z_{.j}}{z} \qquad \forall j \in \mathcal{J}$$

que coincide con la correspondiente al centro de gravedad de la nube de individuos en correspondencias múltiples habitual.

Este será también el origen de los ejes de máxima inercia que se han de buscar.

6. NUBE DE MODALIDADES: $\mathcal{N}(\mathcal{J})$

El punto $j \in \mathcal{J}$ se representa en $\Re^n$ por el perfil $\frac{f_{ij}}{f_{.j}} = \frac{z_{ij}}{z_{.j}}$, $i \in \mathcal{I}$, $j \in \mathcal{J}$, es decir, el mismo que en el A.C.M. clásico.

La distancia cuadrática propuesta entre dos modalidades j y $j' \in \mathcal{J}$ es:

$$d^2(j,j') = \sum_{i\in\mathcal{I}} \frac{1}{g_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2 = \tag{4}$$
$$= n \sum_{i\in\mathcal{I}} \left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}}\right)^2$$

Esta distancia es semejante a la utilizada cuando la tabla es completa y apropiada también para el caso de preguntas condicionadas. Equivale a considerar la distancia euclídea en un espacio dotado de métrica $1/g_{i.}, i \in \mathcal{I}$.

Cada punto $j \in \mathcal{J}$ está dotado de un peso $f_{.j} = z_{.j}/z$, $j \in \mathcal{J}$, coincide con el asignado en correspondencias múltiples de tablas completas y supone que cada modalidad tiene una importancia proporcional a la población que representa. Las modalidades tienen un peso en la construcción de los ejes tanto menor cuanto menor sea su efectivo.

La coordenada i-ésima del centro de gravedad de la nube es:

$$G_J(i) = f_{i.} = \frac{z_{i.}}{z} \qquad \forall i \in \mathcal{I}$$

Sin embargo, en (Goitisolo & Zárraga 1998b) se justifica la elección como origen del punto O_J (cuya coordenada i – ésima es $g_{i.}, i \in \mathcal{I}$). La diferencia entre los puntos O_J y G_J conlleva que en el análisis las nubes han de ser siempre centradas, no existiendo la equivalencia conocida en correspondencias múltiples con tablas completas entre los análisis de las nubes centradas y no centradas.

7. OBTENCIÓN DE LOS FACTORES DE AMBAS NUBES

Calcular la sucesión de ejes $(u_s, s \in \mathcal{S} = \{1, \ldots, s, \ldots, S\}, S \leq J)$ que maximizan la inercia proyectada de la nube $\mathcal{N}(\mathcal{I})$ equivale a:

$$
\begin{array}{lll}
\text{maximizar:} & {u_s}^T M X^T P X M u_s & \\
\text{sujeto a:} & {u_s}^T M u_s = 1 & \\
 & {u_s}^T M u_t = 0 & \forall t < s
\end{array} \tag{5}
$$

donde:

- X es una matriz $(n \times J)$ de término general:

$$x_{ij} = \frac{f_{ij}}{g_{i.} f_{.j}} - 1 \qquad i \in \mathcal{I} \qquad j \in \mathcal{J} \tag{6}$$

  Al igual que en análisis de correspondencias clásico cada elemento de esta matriz contiene las desviaciones entre la tabla de datos $f_{ij}, i \in \mathcal{I}, j \in \mathcal{J}$ y una tabla de término general que corresponde a la hipótesis de independencia. La diferencia con el análisis clásico radica en que la frecuencia relativa marginal correspondiente a las filas es impuesta en función del número de filas en lugar de obtenida a partir de los datos.

- M es una matriz diagonal correspondiente a la métrica del espacio:

$$m_j = f_{.j} \qquad j \in \mathcal{J} \tag{7}$$

- P es la matriz (también diagonal) de pesos:

$$p_i = g_{i.} \qquad i \in \mathcal{I} \tag{8}$$

Se puede demostrar (Escofier & Pagès 1992) que la nube definida por las filas de la matriz X, con la métrica y los pesos considerados, es isomorfa de la definida en §5 como objetivo del estudio. Ambas nubes mantienen las mismas distancias entre dos puntos cualesquiera.

La resolución de este problema lleva a la diagonalización de la matriz $X^T P X M$ de orden $(J \times J)$ cuyo término general es:

$$a_{jj'} = n \sum_{i \in \mathcal{I}} \frac{f_{ij}\, f_{ij'}}{f_{.j}} - f_{.j'} \qquad j, j' \in \mathcal{J} \tag{9}$$

Las proyecciones de la nube de individuos sobre los ejes de máxima inercia resultantes son:

$$F_s = X M u_s \qquad s \in \mathcal{S}$$

Su i-ésima coordenada adopta la expresión:

$$
\begin{aligned}
F_s(i) &= \textstyle\sum_{j \in \mathcal{J}} \left(\frac{n\, f_{ij}}{f_{.j}} - 1 \right) f_{.j}\, u_{sj} = \\
&= n \textstyle\sum_{j \in \mathcal{J}} f_{ij}\, u_{sj} - \sum_{j \in \mathcal{J}} f_{.j}\, u_{sj} \qquad i \in \mathcal{I}, \quad s \in \mathcal{S}
\end{aligned} \tag{10}
$$

Al diagonalizar la matriz XMX^TP cuyo término general es:

$$d_{ii'} = n\sum_{j\in\mathcal{J}} \frac{f_{ij}\, f_{i'j}}{f_{.j}} - f_{i'.} - f_{i.} + \frac{1}{n} \qquad i, i' \in \mathcal{I}$$

se obtienen los ejes $v_s, s \in \mathcal{S}$[1] que maximizan la inercia proyectada de la nube $\mathcal{N}(\mathcal{J})$ y tras premultiplicar dicha matriz por X^TP las proyecciones $G_s, s \in \mathcal{S}$ de dicha nube cuya j-ésima coordenada puede expresarse:

$$\begin{aligned} G_s(j) &= \sum_{i\in\mathcal{I}} \left(\frac{n\, f_{ij}}{f_{.j}} - 1\right) \frac{1}{n} v_{si} = \\ &= \sum_{i\in\mathcal{I}} \frac{f_{ij}}{f_{.j}} v_{si} - \frac{1}{n}\sum_{i\in\mathcal{I}} v_{si} \qquad j \in \mathcal{J}, \quad s \in \mathcal{S} \end{aligned} \tag{11}$$

8. RELACIONES ENTRE LOS FACTORES

Los factores de ambas nubes se relacionan a través de las expresiones:

$$F_s = \frac{1}{\sqrt{\lambda_s}} XMG_s \qquad s \in \mathcal{S} \tag{12}$$

$$G_s = \frac{1}{\sqrt{\lambda_s}} X^TPF_s \qquad s \in \mathcal{S} \tag{13}$$

En el análisis de correspondencias con marginal modificada, igual que ocurre en el análisis clásico para la cantidad $f_{i.}, i \in \mathcal{I}$, los factores $F_s, s \in \mathcal{S}$ están centrados para la cantidad $g_{i.}, i \in \mathcal{I}$:

$$\sum_{i\in\mathcal{I}} g_{i.}\, F_s(i) = 0 \qquad s \in \mathcal{S}$$

Aplicando la fórmula de transición (13):

$$\begin{aligned} G_s(j) &= \frac{1}{\sqrt{\lambda_s}} \sum_{i\in\mathcal{I}} \left(\frac{f_{ij}}{f_{.j}\, g_{i.}} - 1\right) g_{i.}\, F_s(i) = \\ &= \frac{1}{\sqrt{\lambda_s}} \sum_{i\in\mathcal{I}} \frac{f_{ij}}{f_{.j}} F_s(i) \qquad j \in \mathcal{J} \qquad s \in \mathcal{S} \end{aligned} \tag{14}$$

que coincide con la relación baricéntrica del análisis de correspondencias.

[1] Las relaciones de dualidad entre ambos espacios, que se verifican en todo análisis de correspondencias, permiten establecer que los subespacios de ajuste, asociados a valores propios no nulos, son de idéntica dimensión.

Sin embargo, a diferencia del análisis de correspondencias clásico los factores $G_s, s \in \mathcal{S}$ no están centrados por la cantidad $f_{.j}, j \in \mathcal{J}$ porque el análisis se hace tomando como origen un punto diferente al centro de gravedad.

$$\sum_{j\in\mathcal{J}} f_{.j}\, G_s(j) = \sum_{i\in\mathcal{I}} v_{si}\, f_{i.} - \sum_{i\in\mathcal{I}} g_{i.}\, v_{si} \qquad s \in \mathcal{S}$$

Si la marginal impuesta difiere de la marginal propia de la tabla esta cantidad es distinta de cero.

Por ello los factores $F_s, s \in \mathcal{S}$ no pueden interpretarse como el baricentro de los $G_s, s \in \mathcal{S}$ como en análisis clásico. Según la fórmula de transición (12):

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j\in\mathcal{J}} \left(\frac{f_{ij}}{f_{.j}\, g_{i.}} - 1 \right) f_{.j}\, G_s(j) =$$

$$= \frac{1}{\sqrt{\lambda_s}} \left\{ \sum_{j\in\mathcal{J}} \frac{f_{ij}}{g_{i.}}\, G_s(j) - \sum_{j\in\mathcal{J}} f_{.j}\, G_s(j) \right\} \qquad i \in \mathcal{I} \qquad s \in \mathcal{S} \tag{15}$$

El segundo sumatorio corresponde a la proyección del centro de gravedad de $\mathcal{N}(\mathcal{J})$, que al no haber sido tomado como origen de los ejes es diferente de 0.

Benali y Escofier (Benali & Escofier 1987) afirman que «Este término, en la práctica es casi nulo, lo que permite interpretar como en correspondencias múltiples clásico la abscisa de un individuo como el baricentro de las modalidades que ha elegido». Hacen referencia a tablas disyuntivas incompletas en las que el efectivo de datos ausentes representa una proporción reducida en relación al total (caso de datos ausentes por olvido distribuidos de forma aleatoria a lo largo de la tabla). Sin embargo, puede alterar los resultados y su interpretación cuando se considera nulo en una tabla de datos en la cual la proporción de no respuesta es grande o corresponde a ciertos grupos de individuos (caso de tablas de datos con preguntas condicionadas) como se ha podido comprobar en la aplicación a la Encuesta de Condiciones de Vida de 1989 de la Comunidad Autónoma de Euskadi presentada en (Goitisolo & Zárraga 1998a).

Lo cierto es que este segundo sumatorio es el mismo para todos los individuos, aunque difiere para los distintos ejes, por lo que se podría trasladar los factores $F_s(i), s \in \mathcal{S}, i \in \mathcal{I}$ de tal forma que en la representación superpuesta de ambas nubes un individuo siga estando representado en el baricentro de las modalidades que posee:

$$F_s^*(i) = \frac{1}{\sqrt{\lambda_s}}\, n \sum_{j\in\mathcal{J}} f_{ij}\, G_s(j) \qquad i \in \mathcal{I} \qquad s \in \mathcal{S} \tag{16}$$

9. NÚMERO DE EJES

En el análisis de correspondencias de una tabla disyuntiva completa el número de ejes S es igual al número de modalidades activas menos el número de variables, porque todas

las modalidades correspondientes a cada una de las variables se encuentran restringidas al mismo hiperplano. En el análisis de correspondencias con marginal modificada de una tabla disyuntiva incompleta, las modalidades de una misma cuestión no cumplen ningún tipo de restricción por lo que pueden existir tantos ejes como número de modalidades activas exista en el análisis. Si existen cuestiones con datos completos, sus modalidades mantendrán la misma restricción que en el análisis clásico por lo que la cantidad de ejes disminuirá en ese número de cuestiones completas.

La existencia de preguntas condicionadas en el análisis también reduce la cantidad de ejes, puesto que los individuos que han de responder a una pregunta condicionada vienen determinados por la respuesta a una modalidad (o combinación de ellas) anterior.

10. ESTUDIO DE LA ASOCIACIÓN ENTRE LAS CUESTIONES

El estudio de la independencia entre dos variables cualitativas q y $q' \in \mathcal{Q}$ con $\mathcal{J}_q$ y $\mathcal{J}_{q'}$ modalidades respectivamente se realiza habitualmente a través de su tabla de contingencia donde han de ser clasificados todos los individuos. Para ello, en el caso de que alguna de las cuestiones no haya sido respondida por todos los individuos, será necesario tener presente las modalidades de no respuesta. La tabla de contingencia tendrá la forma:

Tabla 2

	$1 \ldots j' \ldots J_{q'}$ a'
1	
$\vdots$	
j	$b_{jj'}^{qq'}$
$\vdots$	
J_q	
a	

donde a y a' representan las modalidades de no respuesta de las cuestiones q y $q' \in \mathcal{Q}$ respectivamente.

Un elemento $b_{jj'}^{qq'}; j \in \mathcal{J}_q; j' \in \mathcal{J}_{q'}; q, q' \in \mathcal{Q}$ de esta tabla indica el número de individuos que pertenecen simultáneamente a las modalidades j y j' de las cuestiones q y $q' \in \mathcal{Q}$ respectivamente. Es decir:

$$b_{jj'}^{qq'} = \sum_{i \in \mathcal{I}} z_{ij} z_{ij'} = \text{Card}\{i : z_{ij} = z_{ij'} = 1 | i \in \mathcal{I}\} \quad j \in \mathcal{J}_q \quad j' \in \mathcal{J}_{q'} \quad q, q' \in \mathcal{Q}$$

Analizar la independencia entre ambas cuestiones lleva a comparar los términos:

(17) $\frac{b_{jj'}^{qq'}}{n}$ y $\frac{b_{jj}^{qq}\, b_{j'j'}^{q'q'}}{n^2}$ $\quad j \in \mathcal{J}_q \quad j' \in \mathcal{J}_{q'} \quad q, q' \in \mathcal{Q}$

que, en función de la tabla disyuntiva incompleta, equivale a comparar:

(18)

$$\frac{\sum_{i \in \mathcal{I}} z_{ij} z_{ij'}}{n} \quad \text{y} \quad \frac{\sum_{i \in \mathcal{I}} z_{ij}^2 \sum_{i \in \mathcal{I}} z_{ij'}^2}{n^2} \qquad j \in \mathcal{J}_q \quad j' \in \mathcal{J}_{q'} \quad q, q' \in \mathcal{Q}$$

Sin embargo, la no respuesta en el caso de cuestionarios con preguntas condicionadas es determinada por una cuestión diferente. Si, por ejemplo, las dos cuestiones han sido condicionadas por la misma pregunta –de tal forma que ambas tienen en común el total de las respuestas ausentes– no existe independencia. Puesto que el interés se centra en la asociación entre las modalidades respondidas, parece más apropiado no tener en consideración la relación entre las modalidades de no respuesta a la hora de analizar la independencia entre ambas cuestiones.

En ACM clásico el análisis simultáneo de la dependencia entre las Q cuestiones se realiza a través de tabla de Burt definida en función de la tabla disyuntiva completa (Z) de la siguiente forma:

$$B = Z^T Z$$

La tabla B de término general $b_{jj'}^{qq'}$ está formada por Q^2 bloques. El bloque (q, q'), de orden $(J_q, J_{q'})$ es la tabla de contingencia que cruza las respuestas a las cuestiones q y $q' \in \mathcal{Q}$.

El elemento $b_{jj'}^{qq'}$, $j, j' \in \mathcal{J}$ es nulo si las dos modalidades pertenecen a la misma cuestión y representa el número de individuos que eligen una determinada modalidad (denominado también por $z_{.j}$ o $z_{.j}^q$, $j \in \mathcal{J}, q \in \mathcal{Q}$ -en §2-) si las modalidades j y $j' \in \mathcal{J}$ coinciden.

Por analogía se define la matriz B^* de orden $(J \times J)$ que será denominada pseudo-tabla de Burt y puede expresarse en función de la tabla disyuntiva incompleta:

$$B^* = {Z^*}^T Z^*$$

La diferencia entre B y B^* radica en que en la matriz B^* el total de cada una de las subtablas que la forman no es constante, sino la cantidad de individuos que responden a las cuestiones q y $q' \in \mathcal{Q}$ (no necesariamente coincidente con el número total de individuos encuestados).

En (Goitisolo & Zárraga 1998b) se demuestra la equivalencia entre los análisis factoriales de la tabla disyuntiva incompleta y la pseudo-tabla de Burt asociada, obteniéndose

factores proporcionales asociados a valores propios que se relacionan mediante:

$$\lambda_s^{B^*} = \left(\frac{z}{n}\lambda_s\right)^2 \qquad s \in \mathcal{S} \tag{19}$$

donde $\lambda_s^{B^*}$ son los valores propios correspondientes al análisis de la pseudo-tabla de Burt y λ_s los del análisis de la tabla disyuntiva incompleta.

10.1. Independencia entre las cuestiones

Al igual que en el análisis de correspondencias múltiples con datos completos, el análisis de la tabla disyuntiva incompleta está asociado a unas tasas de inercia pequeñas que no deben ser interpretadas como partes de información explicada por los ejes. A continuación se estudia el caso en el cual las cuestiones son independientes dos a dos y se propone una corrección a estas tasas de inercia en el análisis con marginal modificada de una tabla disyuntiva incompleta.

La matriz de diagonalización en el análisis con marginal modificada de la tabla disyuntiva incompleta, tiene un término general (recogido en la ecuación (9)) que puede ser expresado en función de los elementos de la tabla disyuntiva incompleta de la siguiente forma:

$$a_{jj'} = \frac{n}{z}\left(\frac{1}{z_{.j}}\sum_{i\in\mathcal{I}} z_{ij}\, z_{ij'} - \frac{1}{n} z_{.j'}\right) \qquad j, j' \in \mathcal{J} \tag{20}$$

Notar que:

- $\sum_{i\in\mathcal{I}} z_{ij}\, z_{ij'} = 0$ si: $j, j' \in \mathcal{J}_q$ y $j \neq j'$ $q \in \mathcal{Q}$
  porque un individuo no puede pertenecer a dos modalidades de la misma cuestión
- $\sum_{i\in\mathcal{I}} z_{ij}\, z_{ij'} = z_{.j}$ si: $j, j' \in \mathcal{J}_q$ y $j = j'$ $q \in \mathcal{Q}$
  número de individuos que pertenecen a una determinada modalidad
- $\sum_{i\in\mathcal{I}} z_{ij}\, z_{ij'} = \frac{z_{.j}\, z_{.j'}}{n}$ si: $j \in \mathcal{J}_q$, $j' \in \mathcal{J}_{q'}$, $q \neq q'$ $q, q' \in \mathcal{Q}$ y además existe independencia entre ambas cuestiones

y que por tanto:

- $a_{jj'} = -\frac{z_{.j'}}{z}$ si: $j, j' \in \mathcal{J}_q$ y $j \neq j'$ $q \in \mathcal{Q}$
- $a_{jj'} = \frac{n - z_{.j'}}{z}$ si: $j, j' \in \mathcal{J}_q$ y $j = j'$ $q \in \mathcal{Q}$

- $a_{jj'} = 0$ si: $j \in \mathcal{J}_q$, $j' \in \mathcal{J}_{q'}$, $q \neq q'$ $q, q' \in \mathcal{Q}$ y existe independencia entre ambas cuestiones.

En consecuencia, la matriz A_{Z^*} que se ha de diagonalizar tiene la forma:

$$A_{Z^*} = \begin{bmatrix} A^1_{Z^*} & & & 0 \\ & \ddots & & \\ & & A^q_{Z^*} & \\ & & & \ddots \\ 0 & & & A^Q_{Z^*} \end{bmatrix}$$

donde cada submatriz $A^q_{Z^*}$ es:

$$A^q_{Z^*} = \begin{bmatrix} \frac{n - z^q_{.1}}{z} & \frac{-z^q_{.2}}{z} & \frac{-z^q_{.3}}{z} & \cdots & \frac{-z^q_{.J_q}}{z} \\ \frac{-z^q_{.1}}{z} & \frac{n - z^q_{.2}}{z} & \frac{-z^q_{.3}}{z} & \cdots & \frac{-z^q_{.J_q}}{z} \\ \frac{-z^q_{.1}}{z} & \frac{-z^q_{.2}}{z} & \frac{n - z^q_{.3}}{z} & \cdots & \frac{-z^q_{.J_q}}{z} \\ & & & \ddots & \\ \frac{-z^q_{.1}}{z} & \frac{-z^q_{.2}}{z} & \frac{-z^q_{.3}}{z} & \cdots & \frac{n - z^q_{.J_q}}{z} \end{bmatrix}$$

Siendo $z^q_{.j}$ equivalente al $z_{.j}$ anterior.

Para diagonalizar la matriz A_{Z^*} se ha de resolver la ecuación:

$$|A_{Z^*} - \lambda_s I| = 0 \qquad s = 1, \ldots, J$$

Debido a la estructura diagonal por bloques de la matriz A_{Z^*}, los valores λ_s (incluyendo los valores nulos) que solucionan esa ecuación son los resultantes de resolver el sistema de Q ecuaciones siguiente:

$$|A^q_{Z^*} - \lambda^q_s I_{J_q}| = 0 \qquad \forall q \in \mathcal{Q} \qquad s = 1, \ldots, J_q$$

siendo I_{J_q} la matriz identidad de orden $J_q, q \in \mathcal{Q}$.

Las matrices $A^q_{Z^*}, q \in \mathcal{Q}$ pueden expresarse como:

$$A^q_{Z^*} = \tfrac{n}{z} I_{J_q} - E^q$$

siendo:

$$E^q = \begin{bmatrix} \frac{z^q_{.1}}{z} & \frac{z^q_{.2}}{z} & \frac{z^q_{.3}}{z} & \cdots & \frac{z^q_{.J_q}}{z} \\ \vdots & & & \ddots & \vdots \\ \frac{z^q_{.1}}{z} & \frac{z^q_{.2}}{z} & \frac{z^q_{.3}}{z} & \cdots & \frac{z^q_{.J_q}}{z} \end{bmatrix}$$

Como se puede comprobar fácilmente, esta matriz (de orden J_q) es de rango 1 y su traza es z_q/z (número de individuos que responden la cuestión $q \in \mathcal{Q}$ entre el número total de respuestas de los n individuos a las Q cuestiones), por ello tiene un valor propio $\mu^q = z_q/z$ y $(J_q - 1)$ valores propios nulos.

A través de la relación entre los valores propios de $A^q_{Z^*}$ y de esta matriz E^q, $q \in \mathcal{Q}$:

$$\begin{aligned} A^q_{Z^*}\, u_s &= \lambda^q_s\, u_s \\ \tfrac{n}{z}\, I_{J_q}\, u_s - E^q\, u_s &= \lambda^q_s\, u_s \\ E^q\, u_s &= \left(\tfrac{n}{z} - \lambda^q_s\right)\, u_s \\ E^q\, u_s &= \mu^q_s\, u_s \end{aligned}$$

donde $s = 1, \ldots, J_q$; se obtienen los valores propios de cada matriz $A^q_{Z^*}$:

$$\lambda^q_s = \begin{cases} \dfrac{n}{z} & \text{si } s = 1, \ldots, J_q - 1 \\ \dfrac{n}{z} - \dfrac{z_q}{z} & \text{si } s = J_q \end{cases} \qquad \forall q \in \mathcal{Q}$$

y de la matriz A_{Z^*} que resultan ser:

$$\lambda_s = \begin{cases} \dfrac{n}{z} & \text{si } s = 1, \ldots, J - Q \\ \dfrac{n}{z} - \dfrac{z_q}{z} & \text{si } s = (J - Q + 1), \ldots, J \end{cases} \qquad \forall q \in \mathcal{Q} \tag{21}$$

Donde existen tantos valores propios nulos como cuestiones a las que responden todos los individuos ($z_q = n$, $q \in \mathcal{Q}$).

Al igual que en el caso de tablas disyuntivas completas, la independencia entre cuestiones no se refleja en una inercia nula y no se debe, obviamente, a asociaciones entre las cuestiones sino a un efecto de estructura o de construcción de la tabla disyuntiva

incompleta. Cada uno de los J ejes estaría recogiendo una inercia trivial que evidentemente engorda los valores propios del análisis de la tabla disyuntiva incompleta cuando no existe la independencia.

10.2. Tasas de inercia

Cuando no existen datos ausentes, la aplicación del análisis de correspondencias simples a la tabla de Burt lleva a la obtención de los mismos factores del análisis de la tabla disyuntiva completa. En el caso particular de dos cuestiones, ambos análisis producen los mismos factores que el análisis de correspondencias simples de la tabla de contingencia. Sin embargo, los valores propios que se obtienen en los tres análisis son diferentes y dan lugar a distintas tasas de inercia proyectada sobre cada uno de los ejes.

Se pone de manifiesto de esta forma que el análisis de correspondencias múltiples (bien se realice a través de la tabla disyuntiva completa o bien a partir de la tabla de Burt) estudia las relaciones de dependencia entre cada par de cuestiones y se revela el escaso interés de los valores propios como medida de la información explicada por cada uno de los factores.

Benzécri (Benzécri 1979) propone por ello, basándose en la equivalencia entre el análisis de correspondencias de la tabla disyuntiva completa y de la tabla de contingencia cuando el número de cuestiones es dos, la siguiente corrección de los valores propios:

$$\lambda_s^* = \left(\frac{Q}{Q-1}\right)^2 \left(\lambda_s - \frac{1}{Q}\right)^2 \qquad s \in \mathcal{S}$$

para aquellos valores λ_s superiores a $1/Q$, siendo λ_s los valores propios resultantes del análisis de la tabla disyuntiva completa. Al estar los valores propios λ_s comprendidos entre 0 y 1, el término $\left(\frac{Q}{Q-1}\right)^2$ permite obtener unos valores propios corregidos comprendidos también entre 0 y 1, que hacen posible su comparación con otros análisis. Esta modificación de los valores propios lleva a definir las tasas de inercia proyectada:

$$\tau_s = \frac{\lambda_s^*}{\sum_s \lambda_s^*} \qquad s \in \mathcal{S}$$

donde la suma del denominador se extiende a los valores λ_s superiores a $1/Q$.

Esta corrección encuentra su justificación, para el caso de más de dos cuestiones, en la equivalencia entre los análisis de la tabla disyuntiva completa y de la tabla de Burt. En éste último análisis se introducen en la diagonal principal tablas que cruzan una cuestión consigo misma haciendo incrementar la inercia total y donde el resto de las tablas de contingencia aparecen dos veces (Greenacre 1993).

Una razón alternativa para calcular las tasas de inercia corregidas se encuentra en el estudio del caso particular donde las Q cuestiones son independientes dos a dos. A pesar del nulo interes del análisis de correspondencias clásico cuando se conoce (en ocasiones únicamente tras los resultados del análisis) la independencia de las cuestiones, su aplicación proporciona una inercia total no nula y $(J-Q)$ factores con valores propios asociados iguales a $1/Q$ (Zárraga 1989), demostrando que los valores propios del análisis (exista o no independencia) recogen una inercia trivial debida a un efecto de estructura o de construcción de la tabla disyuntiva completa.

Cuando existen datos ausentes, el análisis del caso en el cual las cuestiones son independientes, en la forma definida, revela que las tasas de inercia calculadas como los valores propios entre la inercia total, no son una buena medida de la asociación entre las cuestiones recogida por cada eje, por ello se propone, en el caso general en que las cuestiones no son independientes, calcular los valores propios y las tasas de inercia de la siguiente forma:

$$\lambda_s^* = \left(\frac{z}{z-n}\right)^2 \left(\lambda_s - \frac{n}{z}\right)^2 \qquad \forall \lambda_s > \frac{n}{z} \qquad s = 1, \ldots, J$$

$$\tau_s^* = \frac{\lambda_s^*}{\sum_{\lambda_s^* > 0} \lambda_s^*} \qquad s = 1, \ldots, J$$

donde λ_s es el valor propio obtenido en el análisis con marginal modificada de la tabla disyuntiva incompleta cuando no existe independencia. Estos nuevos valores propios y tasas de inercia coinciden, si la tabla es disyuntiva completa, en la que $z = nQ$, con los propuestos por Benzécri (1979).

REFERENCIAS

Abascal, E. & Grande, I. (1989). *Métodos multivariantes para la investigación comercial. Teoría, aplicaciones y programación BASIC*, Ariel economía.

Benali, H. (1985). *Stabilité de l' analyse en composantes principales et de l' analyse des correspondances multiples en présence de certains types de perturbations. Méthodes de dépouillement d' enquêtes. Thèse de troisième cycle*, Université de Rennes I.

Benali, H. (1988). «Données Manquantes et Modalités à Faible Effectif en Analyse des Correspondances Multiples et Conditionnelle», *Data Analysis and Informatics V*, 311-318.

Benali, H. & Escofier, B. (1987). «Stabilité de l' analyse factorielle des correspondances multiples en cas de données manquantes et de modalités à faibles effectifs». *Revue de Statistique Appliquée*, XXXV (1), 41-51.

Benzécri, J.P. (1979). «Sur le calcul des taux d' inertie dans l' analyse d' un questionnaire, addendum et erratum à», *Les Cahiers de l'Analyse des Données*, IV (3), 377-378.

Escofier, B. (1981). «Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte», *INRIA*.

Escofier, B. (1990). «Traitement des Variables Incomplètes en Analyse des Correspondances Multiples», *Revue de Modulad*, 5, 13-27.

Escofier, B. & Pagès, J. (1992). *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*, Servicio Editorial Universidad del País Vasco.

Goitisolo, B. & Zárraga, A. (1998a). «Application of the Incomplete Disjunctive Tables Study to the C.A.V. Living Conditions Survey, in *Analyses Multidimensionnelles des donnes. IV^{éme} Congrès International NGUS'*97, Fernandez-Aguirre, K. and Morineau, A., 301-313.

Goitisolo, B. & Zárraga, A. (1998b). «Equivalence between the Incomplete Disjunctive Table and the Associated Burt Pseudo-table Analysis», in *Analyses Multidimensionnelles des donnes. IV^{éme} Congrès International NGUS'*97, Fernandez-Aguirre, K. and Morineau, A., 227-238.

Greenacre, M.J. (1990). «Some Limitations of Multiple Correspondence Analysis», *Computational Statistics Quartely*, 3, 249-256.

Greenacre, M. (1993). *Correspondence Analysis in Practice*, Academic Press.

Lebart, L., Morineau, A. & Tabard, N. (1977). *Techniques de la description statistique*, Dunod.

Zárraga, A. (1989). *Análisis de correspondencias múltiples por bandas. Aplicación al estudio de una gran encuesta.* Tesis Doctoral, Universidad del País Vasco.

ENGLISH SUMMARY

INDEPENDENCE BETWEEN QUESTIONS IN THE FACTOR ANALYSIS OF INCOMPLETE DISJUNCTIVE TABLES WITH CONDITIONED QUESTIONS

A. ZÁRRAGA
B. GOITISOLO
Universidad del País Vasco-Euskal Herriko Unibertsitatea*

Multiple Correspondence Analysis (MCA) studies the relationship between several categorical variables defined with respect to a certain population. However, one of the main sources of information are those surveys in which it is usual to find a certain number of absent data and conditioned questions that do not need to be answered by the whole population. In these cases, the data codification in a complete disjunctive table requires the inclusion of non-answer categories that can alter the results.
Escofier (Escofie 1981) suggests the analysis of the incomplete disjunctive table (IDT) by substituting the real marginal of the table about the individuals for a constant imposed marginal. As in the analysis of multiple correspondences with complete data, the analysis of the incomplete disjunctive table is related to small percentages of inertia that cannot be regarded as a part of the information expounded by the axes. This paper will examine the case in which the questions are independent two by two and will propose the correction to these percentages of inertia in a modified marginal analysis of an incomplete disjunctive table.

Keywords: Multiple correspondence analysis, incomplete disjunctive table, independence between categorical variables, eigenvalues, percentages of inertia

AMS Classification: 62H25

* Universidad del País Vasco-Euskal Herriko Unibertsitatea. Dpto. de Economía Aplicada III. Fac. CCEE y EE. Avda. Lehendakari Aguirre, 83. 48015 Bilbao. E-mails: az@alcib.bs.ehu.es y bg@alcib.bs.ehu.es.
This work was supported by Dirección General de Enseñanza Superior del Ministerio Español de Educación y Ciencia and Universidad del País Vasco (UPV/EHU) under research grants PB98-0149 and UPV 038.321-HA041/99

-Received February 1998.
-Accepted July 1999.

1. INTRODUCTION

In surveys, there can be absent answers for either unwilling slips by the surveyed, a will to hide certain information, the existence of conditioned questions answered by that person or not, depending on his answer to a previous question.

2. DEFINITION OF INCOMPLETE DISJUNCTIVE TABLES AND THEIR NOTATION

Table 1 gathers the answers by a certain colectivity of individuals $\mathcal{I}$ into a group of questions $\mathcal{Q}$, each of them with a finite group of answer categories $\mathcal{J}_q$, $q \in \mathcal{Q}$. Each element z_{ij} of the table assumes value 1 if individual i answers category j and 0 otherwise. It is said that it is an incomplete disjunctive table - I.D.T.- when for some i and j $z_{ij} = 0$, $i = \{1, \ldots, i, \ldots, n\}, j = \{1, \ldots, j, \ldots, J\}$.

A variable $z_{i.}^{q}$ is defined with value 1 if the individual i answers q and with value 0 otherwise, and z_q as the number of individuals responding to question q, $q \in \mathcal{Q}$.

3. PROBLEM BROUGHT ABOUT BY THE APPLICATION OF STANDARD M.C.A. TO AN I.D.T.

The distance χ^2 between two row profiles i and i' (equation 1) also increases with the common answers, when individuals i and i' do not answer the same number of questions.

In the distance χ^2 between two column profiles j and j' (equation 2) each member could have a different mass according to the number of answers previously chosen.

Therefore, the direct application of the standard M.C.A. is not appropriate to the study of an I.D.T.

4. POSSIBLE ALTERNATIVES TO THE I.D.T. PROBLEM: A FACTOR ANALYSIS WITH A MODIFIED MARGINAL

Either the elimination of those people not answering all the questions, or the creation for each absent data variable of a non-answer category would allow for a complete disjunctive table, but that could alter the results as well. As for the I.D.T. cases with a number of absent data, Escofier (1981) proposes the substitution of the real marginal of the table ($f_{i.}$ which is not constant) for a constant marginal $g_{i.} = 1/n$ in the whole analysis.

5. CLOUD OF INDIVIDUALS

Point i, provided with a weight $g_{i.}$, is represented in $\Re^J$ by the profile $f_{ij}/g_{i.}$. The distance between both i and i' (equation 3) increases only with different answers. This is equivalent to looking for the Euclidean distance in a space provided with a metric $1/f_{.j}$.

The weighting of each category is the inverse to its mass.

The j-th gravity centre coordinate of the cloud is $f_{.j}$. That will also originate the axes of maxima inertia to be searched for.

6. CLOUD OF CATEGORIES

Point j, provided with a weight $f_{.j}$, is represented in $\Re^n$ by the profile $f_{ij}/f_{.j}$.

The distance between categories j and j' (equation 4) is similar to that used when the table is complete. It is equivalent to the Euclidean distance in a space provided with a metric $1/g_{i.}$.

G_J, the i-th gravity centre coordinate of the cloud is: $f_{i.}$. However, (Goitisolo & Zárraga 1998*b*) justifies the election of point O_J (whose i-th coordinate is $g_{i.}$) as origin. The difference between the points O_J and G_J implies that the clouds must always be centered during the analysis.

7. COMPUTATION OF FACTORS IN BOTH CLOUDS

Calculating the succession of the axes (u_s) that maximize the inertia projected on the $\mathcal{N}(\mathcal{I})$ cloud is equivalent to maximizing the expression (5) with the matrices X, M, and P, whose general terms appear in the equations (6, 7 and 8).

The solution to this problem leads to the diagonalization of the matrix $X^T P X M$. The projection of the cloud of individuals on the maxima inertia resulting axes are: $F_s = XMu_s$. Its i-th coordinate assumes the expression (10).

Once the matrix $XMX^T P$ is diagonalized, the axes v_s that maximize the projected inertia about the cloud $\mathcal{N}(\mathcal{J})$ are obtained. And the projetions G_s, whose j-th coordinates are gathered in (11), are also obtained, just after premultiplication of the matrix by $X^T P$.

8. RELATIONSHIPS BETWEEN THE FACTORS

The factors of both clouds are related either with the expressions 12 and 13 or for every coordinate, according to the expressions 14 and 15.

The second summatory of equation 15 corresponds to the gravity center projection of $\mathcal{N}(\mathcal{J})$, which, not having been regarded as the origin of the axes, is different from 0.

Benali and Escofier (1987) suggest that «This term is nearly invalid in practice, which, like in a classic M.C.A., allows one to consider the abscissa of an individual to be the baricentre of the categories chosen». However, that can alter the results and its interpretation when regarded as invalid in a table of data, in which the non-answer proportion is in fact very large. That is exactly what happened in the application to the 1989 Life Conditions in the Basque Country Survey presented in (Goitisolo & Zárraga 1998*a*).

The truth is that this second summatory is similar to all the individuals for which the factors $F_s(i)$ could be shifted so that an individual could still be represented in the baricentre of all his categories during the superposed representation in both clouds (equation 16).

9. NUMBER OF AXES

The categories of the same question do not feature any kind of restriction and there can be as many axes as numbers of active categories in the analysis.

10. STUDY OF THE ASSOCIATION BETWEEN THE QUESTIONS

The study of the independence between two categorical variables q and $q'(q, q' \in \mathcal{Q})$ usually takes place with its contingency table (Table 2, where a and a' represent the non-answer categories of the questions q and q' respectively).

Analyzing the independence between both questions leads to comparing the terms in either equation 17 or 18.

When there are conditioned questions, the association between the categories answered has to be searched for.

In analogy with Burt's table, the matrix B^* is defined and will be named Burt's pseudo-table. It can be expressed according to the incomplete disjunctive table $B^* = {Z^*}^T Z^*$

(Goitisolo & Zárraga 1998*b*) shows the equivalence between the factor analysis Z^* and B^* and they obtain the proportional factors associated to eigenvalues related to equation 19, where $\lambda_s^{B^*}$ are the eigenvalues corresponding to Burt's pseudo-table analysis and λ_s those in the I.D.T.

10.1. Independence between the questions

The I.D.T. analysis is related to small percentages of inertia that cannot be regarded as parts of the information expounded by all the axes.

When there is independence between the questions, the diagonalization matrix in the analysis with an I.D.T. modified marginal does have a general term gathered in equation 9. Its expression according to I.D.T. elements appears in equation 20. Equation 21 gathers the eigenvalues of the analysis. There are as many invalid eigenvalues as questions answered by all the individuals.

10.2. Percentages of inertia

The relationship between the eigenvalues of the B^* and Z^* analyses, and the analysis of the case when the questions are independent two by two reveals that the percentages of inertia calculated as the ratio between the eigenvalues and the total inertia are not an appropriate measure of the association between the questions gathered by each axis. That is why we propose to calculate the eigenvalues and the percentages of inertia as follows:

$$\lambda_s^* = \left(\frac{z}{z-n}\right)^2 \left(\lambda_s - \frac{n}{z}\right)^2 \qquad \forall \lambda_s > \frac{n}{z} \qquad s = 1, \ldots, J$$

$$\tau_s^* = \frac{\lambda_s^*}{\sum_{\lambda_s^* > 0} \lambda_s^*} \qquad s = 1, \ldots, J$$

λ_s is the eigenvalue attained in the analysis with a modified I.D.T. marginal, when there is no independence. These new eigenvalues and percentages of inertia do agree if the table is a complete disjunctive one, in which $z = nQ$, with those proposed by Benzécri (1979).

Estadística Oficial

QÜESTIIÓ, vol. 23, 3, p. 491-505, 1999

UTILISATION D'INFORMATIONS AUXILIAIRES DANS LES ENQUÊTES PAR SONDAGE

Y. TILLÉ
Laboratoire de Statistique d'Enquête*

La notion de représentativité est apparue dès la naissance de la théorie des sondages à la fin du dix-neuvième siècle. Pourtant, ce concept qui est appliqué autant aux plans par quotas qu'aux plans probabilistes est largement galvaudé. Après avoir rappelé quelques éléments de l'histoire de la théorie des sondages, nous rappelons quelques techniques de base de plans aléatoires et à choix raisonnés. Nous montrons ensuite que le concept de plan équilibré permet de lever les ambiguïtés fondamentales de la notion de représentativité.

Use of Auxiliary Information in Survey Sampling

Mots clés: Sondage, plans équilibrés, représentativité

AMS Classification: 62D05

*Laboratoire de Statistique d'Enquête. CREST - ENSAI. École Nationale de la Statistique and de l'Analyse de l'Information rue Blaise Pascal, Campus de Ker Lann 35170 Bruz, France, email: tille@ensai.fr

–Reçu en juillet de 1999.
–Accepté en octobre de 1999.

1. SONDAGE ET REPRÉSENTATIVITÉ

La théorie des sondages est un ensemble d'outils statistiques permettant l'étude d'une population au moyen de l'examen d'une partie de celle-ci. Le sondage s'oppose au recensement qui est l'étude exhaustive de la population. La théorie des sondages vise à justifier ce processus d'extrapolation (illustré en figure 1.). Nous verrons cependant que l'extrapolation de la partie au tout est une démarche qui a été rejetée par les statisticiens jusqu'au début du vingtième siècle. Les arguments visant à valider cette extrapolation ne sont pas encore toujours clairs. La justification la plus couramment utilisée est la «représentativité» de l'échantillon.

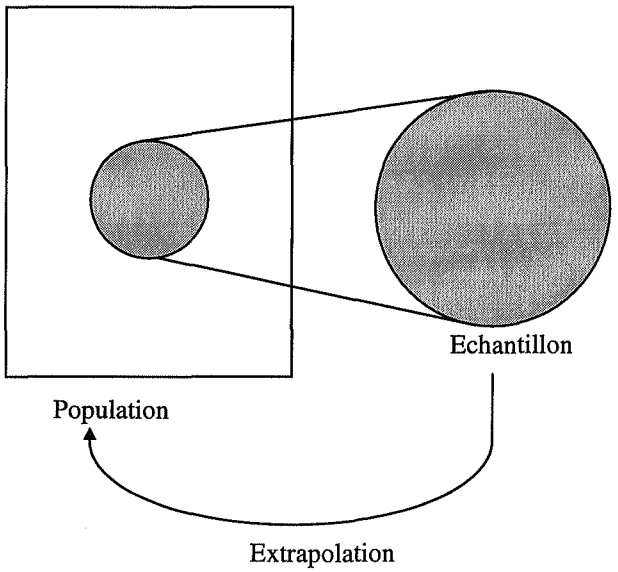


Figure 1. Extrapolation de l'échantillon à la population

On dit souvent qu'un échantillon est représentatif de la population s'il en constitue un modèle réduit. Un bon échantillon devrait «ressembler» autant que possible à la population à étudier de sorte que certaines catégories apparaissent en même proportion dans l'échantillon et dans la population. Pourtant, cette théorie couramment véhiculée par les médias et même par certains ouvrages de méthodologie est incorrecte: un échantillon pour être valide ne doit pas être représentatif (ou sens où nous venons de le définir).

Il est en effet souvent souhaitable d'effectuer des tirages à probabilités inégales ou de surreprésenter certaines parties de la population. Pour estimer de manière précise

une fonction d'intérêt, il faut aller chercher l'information de manière judicieuse plutôt que d'accorder la même importance à chaque unité. Prenons un exemple, si on veut estimer la production de fer d'un pays et qu'on sait que cette production est assurée, d'une part, par deux entreprises sidérurgiques gigantesques qui occupent des milliers de travailleurs et, d'autre part, par plusieurs centaines des petites entreprises artisanales de moins de cinquante travailleurs, va-t-on sélectionner chaque unité avec une même probabilité? Non, bien sûr. On va commencer par s'enquérir de la production des deux grandes entreprises (qui seront donc sélectionnées d'office dans l'échantillon). Ensuite, on sélectionnera les petites de manière aléatoire selon un plan de sondage à déterminer. Cet exemple simple va à l'encontre de l'idée de représentativité et montre bien qu'il faut aller chercher l'information là où elle se trouve, et que le concept de représentativité n'est pas pertinent.

2. ORIGINE DE LA THÉORIE DES SONDAGES

Le développement de la méthodologie statistique (Étymologiquement, science de l'État) est indissociable de l'émergence des États modernes au dix-neuvième siècle. Une des personnalités les plus marquantes de la statistique officielle du dix-neuvième siècle est le belge Adolphe Quetelet (1796-1874) qui fut d'abord attiré par l'idée d'utiliser des données partielles, mais s'est rapidement rallié à l'idée selon laquelle l'utilisation de données partielles est incompatible avec la déontologie statistique. Depuis lors, Quetelet a toujours considéré l'exactitude comme un principe de base de la science statistique. Celui-ci eut une grande influence dans le développement de la statistique officielle. Il organisa le premier Congrès International de la Statistique à Bruxelles en 1853. Il a vraisemblablement contribué à faire admettre par toute la communauté scientifique que l'utilisation de sondages n'est pas une méthode statistique valide.

Au dix-neuvième siècle, l'établissement d'un appareil statistique fut une nécessité dans l'édification des grands États modernes. À cette époque, l'objectif du statisticien était surtout de réaliser des énumérations. La préoccupation majeure était d'inventorier les ressources des nations. Dans ce contexte, le recours à l'échantillonnage fut unanimement rejeté comme une procédure inexacte et donc foncièrement anti-scientifique. Tout au long de ce siècle, les discussions des statisticiens portent essentiellement sur la méthode à appliquer pour obtenir des données fiables et sur la présentation, l'interprétation et éventuellement la modélisation (par un ajustement) de ces données.

En 1895, le Norvégien A.N. Kiaer, directeur du Bureau Central de la Statistique de Norvège, présente au Congrès de l'Institut International de Statistique (IIS) à Berne un travail intitulé «Observations et expériences concernant des dénombrements représentatifs» relatif à un sondage réalisé en Norvège. Kiaer sélectionne d'abord un échantillon de villes et de communes. Ensuite, dans chacune de ces communes, il ne sélectionne

qu'une partie des individus selon la première lettre de leurs noms de famille. Il applique donc un plan à deux degrés mais le choix des unités n'est pas aléatoire. Kiaer défend l'intérêt de l'utilisation de données partielles pour peu qu'elles soient produites au moyen d'une «méthode représentative». Selon cette méthode, l'échantillon doit être une représentation de la population à taille réduite. La notion de représentativité de Kiaer est donc liée à la méthode des quotas. L'intervention de Kiaer est suivie d'un débat houleux, les actes du congrès de l'IIS rendent compte d'une longue polémique. Examinons de plus près l'argumentation de deux des opposants à la méthode de Kiaer (voir Procès-verbal de l'Assemblée Générale de l'IIS, 1896).

> M.V. Mayr [...] C'est surtout dangereux de se déclarer pour ce système des investigations représentatives au sein d'une assemblée de statisticiens. On comprend que pour des buts législatifs ou administratifs un tel dénombrement restreint peut être utile - mais alors il ne faut pas oublier qu'il ne peut jamais remplacer l'observation statistique complète. Il est d'autant plus nécessaire d'appuyer là-dessus, qu'il y a parmi nous dans ces jours un courant au sein des mathématiciens qui, dans beaucoup de directions, voudraient plutôt calculer qu'observer. Mais il faut rester ferme et dire: pas de calcul là où l'observation peut être faite.
>
> M. Milliet. Je crois qu'il n'est pas juste de donner par un voeu du congrès à la méthode représentative (qui enfin ne peut être qu'un expédient) une importance que la statistique sérieuse ne reconnaîtra jamais. Sans doute, la statistique faite avec cette méthode ou, comme je pourrais l'appeler, la statistique, Pars pro toto, nous a donné ça et là des renseignements intéressants ; mais son principe est tellement en contradiction avec les exigences que doit avoir la méthode statistique, que, comme statisticiens, nous ne devons pas accorder aux choses imparfaites le même droit de bourgeoisie, pour ainsi dire, que nous accordons à l'idéal que scientifiquement nous nous proposons d'atteindre.

Le contenu de ces réactions peut se résumer ainsi: comme la statistique est par définition exhaustive, renoncer au dénombrement complet c'est nier la mission même de la science statistique. La discussion ne porte donc pas sur la méthode proposée par Kiaer mais sur la définition de la science statistique. Kiaer ne désarme pourtant pas et continue à défendre la méthode représentative en 1897 au congrès de l'IIS à Saint-Petersbourg, en 1901 à Budapest et en 1903 à Berlin. Après cette date, la question ne sera plus mentionnée au congrès de l'IIS. Kiaer obtient cependant l'appui d'Arthur Bowley (1869-1957) qui jouera ensuite un rôle déterminant dans le développement des sondages. Bowley (1906) présente une vérification empirique pour l'application du théorème central limite à l'échantillonnage. Celui-ci fut le véritable promoteur des techniques de sondage aléatoire, il développe les plans stratifiés avec allocations proportionnelles et utilise la formule de décomposition de la variance.

En 1924, une commission (composée de Arthur Bowley, Corrado Gini, Adolphe Jensen, Lucien March, Verrijn Stuart, et Frantz Zizek) est créée afin d'évaluer la pertinence de l'utilisation de la méthode représentative. Les résultats de cette commission intitulés «Reports on the representative method in statistics» sont présentés au congrès

de l'IIS de 1925 à Rome. La commission accepte le principe du sondage pour autant que la méthodologie soit respectée. Plus de trente ans après la communication de Kiaer, l'idée de l'échantillonnage est donc officiellement acceptée. La commission jettera les bases des recherches futures: deux méthodes sont clairement distinguées «la sélection aléatoire» et la «sélection raisonnée». Ces deux méthodes correspondent à deux démarches scientifiques fondamentalement différentes. D'une part, la validation des méthodes aléatoires est basée sur le calcul des probabilités qui permet de construire des intervalles de confiance pour certains paramètres. D'autre part, la validation des méthodes par sélection raisonnée ne peut être donnée que par l'expérimentation en comparant les estimations obtenues à des résultats de recensement. Les méthodes aléatoires sont donc validées par un argument strictement mathématique tandis que les méthodes par choix raisonnés sont validées par une démarche expérimentale.

Depuis la publication de ce rapport, l'opposition entre ces deux types de plans de sondage est restée pleinement d'actualité. Dans un article récent, Brewer (1999) oppose encore les plans probabilistes stratifiés aux plans obtenus par une méthode de choix raisonnés. Les méthodes probabilistes sont plus largement utilisées en statistique officielle, tandis que les méthodes à choix raisonnés (et plus particulièrement la méthode des quotas), sont largement utilisées (en Europe) dans les instituts privés de statistique. Une des ambiguïtés majeures du terme représentativité est qu'il est appliqué indifféremment aux plans probabilistes et aux plans à choix raisonnés.

3. INFORMATION AUXILIAIRE

La notion d'information auxiliaire regroupe toute information extérieure à l'enquête proprement dite permettant d'augmenter la précision des résultats d'un sondage. De manière générale, on appelle information auxiliaire toute information connue sur la population. Cette information peut être la connaissance des valeurs d'une ou de plusieurs variables sur toutes les unités de la population ou simplement d'une fonction de ces valeurs. Pour la plupart des enquêtes, une information auxiliaire est disponible. Elle peut être donnée par un recensement ou tout simplement par la base de sondage. On peut citer comme exemple d'information auxiliaire: le total d'un caractère sur la population, des sous-totaux selon des sous-populations, des moyennes, des proportions, des variances, les valeurs d'un caractère sur toutes les unités de la base de sondage. La notion d'information auxiliaire englobe donc toute donnée issue de recensement.

Les variables dont au moins une fonction des valeurs est connue sont alors appelées variables auxiliaires. L'objectif principal consiste donc à mettre à profit toutes ces informations pour obtenir des résultats précis. L'information auxiliaire peut être utilisée à deux moments: à l'étape de la conception du plan de sondage et à l'étape de l'estimation des paramètres (voir figure 2.).

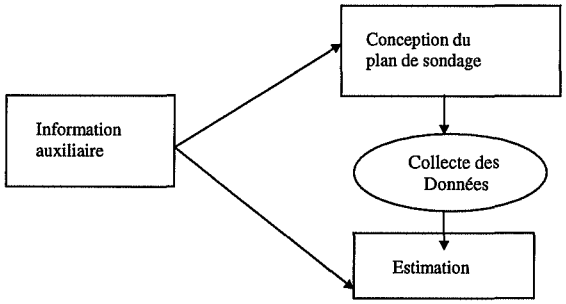


Figure 2. Les deux étapes de l'utilisation de l'information auxiliaire

Quand l'information auxiliaire est mise à profit pour concevoir le plan de sondage, on cherche un plan qui fournit des estimateurs précis pour un prix donné ou qui est peu coûteux pour des critères de précision donnés. Pour ces raisons, on utilisera des plans à probabilités inégales, par grappes ou à plusieurs degrés. Quand l'information est utilisée à l'étape de l'estimation, elle sert à «recaler» les résultats du sondage sur l'information auxiliaire du recensement. Les estimateurs sont alors basés sur deux sources d'informations: l'information auxiliaire connue sur toute la population, et l'information concernant les variables d'intérêt connue uniquement sur les unités sélectionnées dans l'échantillon (voir figure 3.). La méthode générale de calage (en anglais: *calibration*) de Deville et Särndal (1992) permet d'utiliser des informations auxiliaires en modifiant les poids affectés aux unités de manière à ce que les estimateurs de totaux calculés dans l'échantillon soient égaux aux totaux de la population pour toutes les variables auxiliaires connues. Nous limiterons cependant, par la suite, à l'utilisation de l'information auxiliaire à l'étape de la planification.

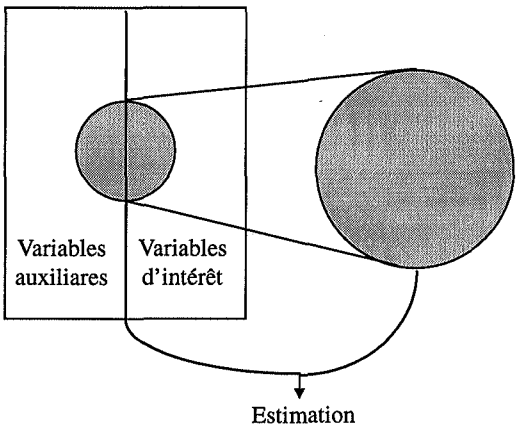


Figure 3. Estimation avec information auxiliaire

4. PLAN DE SONDAGE ET OBJECTIF D'ESTIMATION

Considérons une population de taille N, et supposons que les unités d'observation peuvent être désignées par un numéro d'ordre $k \in \{1, \ldots, k, \ldots, N\} = U$. On s'intéresse à une variable d'intérêt y dont la valeur prise sur l'unité k est notée y_k, pour tout $k \in U$. L'objectif est d'estimer le total de ces valeurs

$$Y = \sum_{k \in U} y_k,$$

au moyen d'un échantillon de cette population. La taille de la population est un total particulier qui s'obtient quand $y_k = 1, k \in U$. Dans ce cas,

$$N = \sum_{k \in U} 1.$$

La moyenne de la variable y dans la population peut alors s'écrire comme un rapport de deux totaux

$$\overline{Y} = \frac{Y}{N},$$

qui seront estimés séparément.

Un échantillon est un sous-ensemble non-vide de U et un plan de sondage est une loi de probabilité $p(.)$ sur tous les échantillons $s \subset U, \#s = n$, telle que

$$p(s) \geqslant 0, \text{ pour tout } s \subset U, \text{ tel que } \sum_{s \subset U} p(s) = 1.$$

Si S est l'échantillon aléatoire tel que $Pr(S = s) = p(s)$, on note I_k la variable aléatoire indicatrice qui prend la valeur 1 si $k \in S$ et 0 sinon, pour tout $k \in U$. De plus, on note $\pi_k = E(I_k) = Pr(k \in S)$, la probabilité d'inclusion d'ordre un, c'est-à-dire la probabilité que l'unité k soit sélectionnée dans l'échantillon. Enfin, on note $\pi_{l\ell} = E(I_k I_\ell) = P(k \in S \text{ et } \ell \in S), k \neq \ell$, la probabilité d'inclusion d'ordre deux, c'est-à-dire la probabilité que deux unités distinctes k et ℓ soient sélectionnées conjointement dans l'échantillon.

Le total Y peut s'estimer sans biais par l'estimateur d'Horvitz-Thompson.

$$\hat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

La variance de l'estimateur de Horvitz-Thompson est donnée par

$$\text{Var}\left[\hat{Y}_\pi\right] = \sum_{k \in S} \sum_{k \in S} \frac{y_k\, y_\ell}{\pi_k\, \pi_\ell} \Delta_{k\ell},$$

où

$$\Delta_{k\ell} = \begin{cases} \pi_k(1-\pi_k) & \text{si} \quad k = \ell \\ \pi_{k\ell} - \pi_k \pi_\ell & \text{si} \quad k \neq \ell \end{cases}$$

La précision de l'estimateur de Horvitz-Thompson ne dépend donc du plan qu'au travers des probabilités d'inclusion à l'ordre un et deux.

5. INFORMATION AUXILIAIRE DANS LES PLANS PROBABILISTES

L'introduction d'information auxiliaire peut avoir deux objectifs: l'amélioration de la précision pour un coût donné, ou l'amélioration de l'organisation de l'enquête. L'impact de l'introduction de l'information auxiliaire dans le plan sur la précision des estimateurs se fera, soit sur les probabilités d'inclusion d'ordre un, soit sur les probabilités d'inclusion d'ordre deux, soit sur les deux en même temps.

5.1. Plans à probabilités inégales

Les plans à probabilités inégales consistent à introduire un «effet» sur les probabilités d'inclusion d'ordre un. Ces plans s'avèrent particulièrement intéressants quand les variables sont liées par un effet de taille. Par exemple, pour des entreprises, des variables comme le chiffre d'affaires, le nombre de travailleurs, sont liées par un tel effet. Si une variable auxiliaire x permet de mesurer approximativement cet effet, il est particulièrement intéressant de sélectionner les unités d'observation avec des probabilités d'inclusion proportionnelles à cette variable auxiliaire. Le gain de précision sera alors très important. L'idée même du tirage à probabilités inégales va à l'encontre de la notion de représentativité telle que nous l'avons définie précédemment.

5.2. Plans stratifiés

La technique classique de stratification permet presque toujours d'améliorer la précision d'un estimateur. La stratification consiste à partitionner la population en strates, puis à sélectionner un plan aléatoire simple de taille fixe dans chaque strate. Pour pouvoir réaliser un tel tirage, il est nécessaire de disposer d'une information auxiliaire qui permet d'affecter chaque unité à une strate. L'estimateur d'Horvitz-Thompson présente l'intéressante propriété d'être naturellement «calé» sur les tailles des strates. En effet, si on estime les tailles des strates à partir de l'échantillon, on estime ces tailles sans biais avec une variance nulle.

Rappelons également qu'en stratification, les probabilités d'inclusion ne doivent pas nécessairement être égales d'une strate à l'autre. La stratification optimale de Neyman consiste d'ailleurs à surreprésenter les unités dans les strates où la dispersion est plus importante. Il est intéressant de constater que la stratification optimale de Neyman infirme l'idée de la représentativité. Il n'est pas du tout nécessaire que les unités soient «représentées» dans l'échantillon de manière proportionnelle aux effectifs des strates dans la population.

5.3. Plans à plusieurs degrés

Les plans à plusieurs degrés visent plutôt à une économie de moyens. Un premier échantillonnage est appliqué sur des unités primaires (par exemple des communes). On sélectionne ensuite des unités secondaires (par exemple des ménages) dans les unités primaires sélectionnées. Un plan classique consiste à sélectionner les unités primaires à probabilités inégales proportionnelles aux nombres d'unités secondaires (nombre de ménages dans la commune). Ensuite, on sélectionne un nombre fixe d'unités secondaires dans les unités primaires sélectionnées. Un tel plan présente l'intérêt d'être facile à gérer en terme de répartition de travail entre les enquêteurs. En établissant le formulaire, on constate que le premier degré d'échantillonnage contribue beaucoup plus à la variance des estimateurs que la seconde. Il est donc important de «soigner» le tirage des unités primaires.

6. INFORMATION AUXILIAIRE DANS LES PLANS À CHOIX RAISONNÉS

La méthode des quotas est la méthode empirique la plus utilisée. Le principe est le suivant: on divise la population en un certain nombre de sous-populations selon une ou plusieurs variables catégorielles. Ensuite, on demande aux enquêteurs d'interroger un nombre d'individus proportionnel à chacune de ces sous-populations. Les enquêteurs sont libres de choisir les personnes à interroger. Ce sont donc les enquêteurs qui construisent le plan de sondage. Le plan de sondage et les probabilités d'inclusion sont inconnus. Les avantages de cette méthode sont nombreux: il n'est pas nécessaire de disposer de la base de sondage. Les seules informations utiles sont les effectifs de certaines catégories de la population. De plus, le problème des refus de réponse ne se pose pas puisque l'enquêteur peut choisir lui-même les individus à interroger.

La technique des quotas marginaux consiste à demander aux enquêteurs de sélectionner un certain nombre d'unités de manière à vérifier conjointement les effectifs de plusieurs variables catégorielles, classe d'âge, profession, niveau d'études, etc. L'enquêteur reçoit une «feuille» de quotas indiquant l'effectif à atteindre pour chaque modalité de chaque variable. L'enquêteur peut choisir assez librement les premières personnes à

interroger, mais verra ses choix de plus en plus contraints au fur et à mesure que sa feuille de quotas se remplit.

Si on peut considérer les plans stratifiés comme la version probabiliste des plans par quotas sur une variable catégorielle, il n'existait pas de version probabiliste des plans par quotas marginaux. Un des intérêts de la technique des quotas est qu'elle élude le problème de la non-réponse. Cependant comme un remplacement est organisé d'office par les enquêteurs, le biais dû aux non réponses reste présent dans l'enquête. Comme le gestionnaire d'enquête ne sait pas qui a refusé de répondre, il est impossible de réaliser une correction de ce biais de non-réponse.

7. PLANS ÉQUILIBRÉS: LA SYNTHÈSE

Les plans équilibrés présentent à la fois les avantages des plans par quotas et des plans probabilistes. De manière générale, on se réfère à la définition suivante.

Définition 1. *Un plan de sondage $p(s)$ est dit équilibré pour les variables auxiliaires $x_1, \ldots, x_p$, si et seulement si il vérifie les équations d'équilibrage:*

$$\sum_{k \in S} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj}.$$

où x_{kj} est la valeur prise par la variable j sur l'unité k.

Un plan équilibré estime exactement les totaux des variables auxiliaires avec l'estimateur naturel d'Horvitz-Thompson. Il peut être à probabilités inégales et les variables $x_1, \ldots, x_p$, peuvent être catégorielles ou quantitatives. Si les plans par quotas étaient probabilistes, ils seraient donc équilibrés pour les variables de quotas. Examinons quelques cas particuliers:

Exemple 1. *Un plan de taille fixe est équilibré sur la variable auxiliaire $x_k = \pi_k, k \in U$. En effet,*

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in S} 1 = \sum_{k \in U} \pi_k = n.$$

Exemple 2. *Supposons que le plan soit stratifié et que dans chaque strate $U_h, h = 1, \ldots, H$, de taille N_h on sélectionne un plan simple sans remise de taille fixe n_h, alors le plan est équilibré sur les variables δ_{kh} de valeurs*

$$\delta_{h\ell} = \begin{cases} 1 & si \quad k \in U_h \\ 0 & si \quad k \notin U_h \end{cases}$$

En effet,

$$\sum_{k \in S} \frac{\delta_{hk}}{\pi_k} = \sum_{k \in S} \frac{N_H \, \delta_{hk}}{n_h} = N_H,$$

pour $h = 1, \ldots, H$.

La notion de plan équilibré éclaircit les ambiguïtés soulevées par le concept galvaudé de représentativité. Un plan équilibré peut être à probabilités inégales. De plus, il n'est jamais équilibré en soi mais pour un ensemble quelconque de variables particulières. La définition formelle énoncée ci-dessus correspond donc à l'exigence de rigueur du statisticien. Le concept de plan équilibré est ancien. Il est déjà présent dans Thionet (1953). Il a été longuement discuté dans le cadre de l'inférence basée sur un modèle par Royall et Herson (1973) par exemple. Plus récemment, une procédure de tirage équilibré a été appliquée pour le tirage de l'échantillon-maître par Ardilly (1991). Cependant jusqu'à présent, aucune procédure simple ne permettait de sélectionner un échantillon équilibré pour un ensemble de variables. Récemment, Deville et Tillé (1999) ont proposé une méthode permettant de sélectionner des échantillons équilibrés sur un ensemble de variables auxiliaires. Le tirage à probabilités inégales de taille fixe, la stratification en sont des cas particuliers. La méthode permet également d'utiliser plusieurs critères de stratification dans le même plan. La méthode a été implémentée sous SAS, et servira probablement à la sélection des unités primaires dans de nombreux plans de sondages.

RÉFÉRENCES

Ardilly, P. (1991). «Echantillonnage représentatif optimum à probabilités inégales», 23, 91-113.

Bowley, A.L. (1906). «Address to the economic and statistics section for the British Association of Advancement of Sciences», *Journal of the Royal Statistical Society*, 69, 540-558.

Brewer, K.R.W. (1999). «Design-based or prediction based inference? Stratified random vs stratfied balanced sampling», *International Statistical Review*, 67, 35-47.

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

Deville J.-C. et Tillé, Y. (1999). *Balanced sampling by means of the cube method.* Manuscrit non-publié, ENSAI, Paris.

Horvàth, R.A. (1974). «Les idées de Quetelet sur la formation d'une discipline moderne et sur le rôle de la théorie des probabilités», in *Mémorial Adolphe Quetelet*, N°3, Académie Royale des Sciences de Belgique.

Royall, R. et Herson, J. (1973). «Robust estimation in finite populations I», *Journal of the American Statistical Association*, 68, 880-889.

Stigler, S.M. (1986). *The History of Statistics*, Cambridge-London, Harvard University Press.

Thionet, P. (1953). «La théorie des sondages», *Etudes théoriques*, N°5, Paris, INSEE.

— «Procès-verbal de l'Assemblée Générale de l'Institut International de Statistique, N°13», *Séance du vendredi matin 30 août, Bulletin de l'Institut International de Statistique*, Berne, 9, livre 1, 1896, pp. LXXXVIII-XCVII.

ENGLISH SUMMARY

USE OF AUXILIARY INFORMATION IN SURVEY SAMPLING

Y. TILLÉ
Laboratoire de Statistique d'Enquête*

The concept of representativeness appears with the creation of the theory of survey sampling at the end of the 19th century. Nevertheless, this concept which is applied at the same time for quota sampling and for random sampling is dramatically overworked. After a brief presentation of the history of survey sampling, we give a short overview of the basic techniques of planning for purposive selection and for random sampling. Furthermore it is shown that the concept of balanced sampling allows to remove the ambiguity of the notion of representativeness.

Keywords: Sampling, balanced sampling, representativeness

AMS Classification: 62D05

*Laboratoire de Statistique d'Enquête. CREST - ENSAI. École Nationale de la Statistique and de l'Analyse de l'Information rue Blaise Pascal, Campus de Ker Lann 35170 Bruz, France, email: tille@ensai.fr

–Received July 1999.
–Accepted October 1999.

The sampling theory allows to study a population by means of subset of this population called sample. The sampling theory aims at justifying this extrapolation process (see Figure 1). The idea of extrapolation from the sample to the population was rejected till the beginning of the 20th century. The arguments used were not always very clear? The most common justification was the representativeness of the sample.

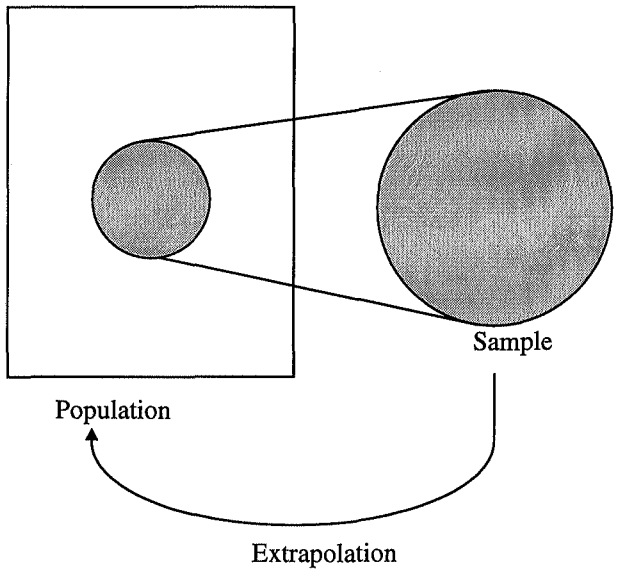


Figure 1. Extrapolation from the sample to the population

Usually a sample is said to be representative when it is a «small-scale» model of the population. A «good» sample should be very similar to the target population. That is some categories appears with the same proportion in the sample and in the population. Kiaer already advocated for the use of representative samples in 1895 at the congress of the International Statistical Institute (Berne). Nevertheless a rapid examination of the modern sampling theory shows that it is often more efficient to select units with unequal probabilities and that the intuitive idea of representativeness is actually false.
Consider a population $U = \{1, \ldots, k, \ldots, N\}$. We are interested to estimate the total of the values y_k, for all $k \in U$. Thus the objective is to estimate

$$Y = \sum_{k \in U} y_k.$$

Suppose also that a random sample S is selected. Let us define by $\pi_k = E(I_k) = Pr(k \in S)$ the first order inclusion probabilities. An unbiased estimator of Y is given

by the Horvitz-Thompson estimator

$$\hat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Moreover suppose that the values of p auxiliary variables $x_1, \ldots, x_p$, are known on all the units of the population.

Definition 1. *A sampling design is said to be balanced on the auxiliary variables $x_1, \ldots, x_p$, if and only if it satisfies*

$$\sum_{k \in S} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj},$$

for $j = 1, \ldots, p$, where x_{kj} is the value taken of variable j for unit k.

The concept of balanced sampling generalises most of the sampling techniques that allows to use auxiliary information at the estimation stage. Moreover it allows us to use unequal inclusion probabilities, and it removes the ambiguity of the concept of representativeness. Recently Deville and Tillé (1999) have proposed a general method to select a random sample balanced on a large number of auxiliary variables.

QÜESTIIÓ, vol. 23, 3, p. 507-524, 1999

MUESTREO Y RECOGIDA DE DATOS EN EL ANÁLISIS DE REDES SOCIALES

J.M. VERD PERICÁS
J. MARTÍ OLIVÉ
Universitat Autònoma de Barcelona*

El artículo revisa las propuestas que dentro de la perspectiva del Análisis de Redes Sociales han realizado diversos autores en relación al muestreo y la recogida de datos. Estos aspectos, resueltos de modo satisfactorio en la perspectiva individualista-atomista, plantean no pocos problemas en la perspectiva de redes sociales. Resulta especialmente problemática la posibilidad de realizar muestras representativas de las relaciones existentes en una población. Aún en el caso de conocer con antelación todos los actores y sus relaciones entre ellos, la elección de una muestra representativa de actores no garantiza una muestra representativa de relaciones. La alternativa tomada en el artículo es la de realizar el análisis de grupos sociales tomados como «poblaciones». Una aplicación práctica de esta última opción es presentada como ejemplo.

Data collection and sampling in Social Network Analysis

Palabras clave: Redes sociales, datos relacionales, muestreo

Clasificación AMS: 92H30, 92G99

*Grup d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (QUIT). Departament de Sociologia. Universitat Autònoma de Barcelona. Edifici B. 08193 Bellaterra (Barcelona). E-mails: Joel.Martí@uab.es, JoanMiquel.Verd@uab.es.

–Recibido en julio de 1999.
–Aceptado en octubre de 1999.

1. INTRODUCCIÓN: ¿QUÉ ES EL ANÁLISIS DE REDES SOCIALES?

El análisis de redes sociales ha experimentado en los últimos años una creciente popularidad en el mundo de las ciencias sociales como alternativa (y en otros casos como complemento) al análisis de tipo individualista-atomista. Frente al estudio tradicional centrado en la consideración de los atributos individuales y la construcción de categorías basadas en estos atributos, el análisis de redes sociales aboga por tomar las relaciones entre actores como el «material» sobre el cual se construye y se organiza el comportamiento social de los actores.

Con este enfoque, el punto de partida del análisis deja de ser el individuo y pasan a serlo las *relaciones*. Como apuntan Wellman y Berkowitz (1988), las relaciones entre personas estructuran la asignación de recursos, y esta estructuración se refleja en redes de poder y dependencia, de este modo resulta más fructífero analizar las pautas de interacción entre las diferentes unidades (generalmente personas), que analizar las características individuales de las unidades que se consideran.

Debe apuntarse que este análisis se ha generalizado yendo más allá de las relaciones entre personas, pudiéndose aplicar a cualesquiera unidades entre las cuales sea posible concebir algún tipo de *interacción* (con un sentido más amplio que el que tiene en el lenguaje habitual): entre grupos o instituciones en una organización social, entre palabras o frases en una estructura lingüística, etc. Las interacciones consideradas pueden ser infinitas, entendiéndose como relación, por ejemplo, (entre personas) haber estudiado en un mismo centro (aunque nunca se haya tenido contacto personal directo), o (entre países) tener un mismo sistema de gobierno. Es decir, la definición de relación puede ir desde cualquier tipo de contacto directo entre las unidades en que se está interesado hasta el hecho de compartir una determinada característica.

1.1. Conceptos fundamentales de redes sociales

Además del concepto de *relación*, que acabamos de definir, existe en el Análisis de Redes Sociales un conjunto de conceptos clave en torno a los cuales se sistematiza el trabajo de los/as diversos/as autores/as agrupables bajo este enfoque. Según Wasserman y Faust (1994) son los siguientes:

Actor: Son las entidades entre las cuales se establecen los vínculos que se pretenden analizar. Puede tratarse de individuos, empresas u otras unidades de carácter colectivo. El nombre utilizado no implica que estas entidades necesariamente tengan la capacidad de volición o de actuar.

Lazo relacional: Son los vínculos existentes entre pares de actores. La gama y tipo de lazos es muy diverso: opiniones de carácter personal (amistad, respeto, preferencia),

transmisión de recursos (transacciones económicas, información), interacción entre individuos (hablar, escribirse), conexión física (una carretera, un puente), pertenencia o afiliación a una misma organización, relación de parentesco, etc.

Díada: Una díada consiste en un par de actores y los posibles vínculos entre ellos. Los vínculos se contemplan siempre como una propiedad de una pareja de actores, y nunca como una característica individual. Por lo tanto la díada es el nivel mínimo al cual puede realizarse el análisis.

Tríada: Subconjunto de tres actores y sus posibles vínculos. Importantes métodos y modelos se basan en ellas para su análisis, particularmente los interesados en la transitividad y en el equilibrio de las relaciones.

Subgrupo: Puede definirse como un subconjunto superior a tres de actores y sus relaciones entre ellos. Existen diferentes criterios para delimitarlos.

Grupo: Sistema de actores que ha sido delimitado por razones conceptuales, teóricas o empíricas, lo cual permite ser tratado como un conjunto finito. Se trata del conjunto de actores cuyos vínculos serán analizados.

Red social: Conjunto finito de actores y de relaciones definidas entre ellos.

1.2. El desarrollo de la perspectiva

Wasserman y Faust (1994) agrupan las aportaciones que han ayudado al desarrollo del Análisis de Redes Sociales alrededor de tres grandes motivaciones: las empíricas, las teóricas y las matemáticas. En relación a las motivaciones empíricas estos autores citan el trabajo pionero de Moreno en los años treinta, que con la creación del *sociograma* obtuvo una forma de visualizar las relaciones y la estructura dentro de un pequeño grupo; algo más tarde (años cincuenta) también la necesidad en psicología social de representar las estructuras de comunicación en pequeños grupos llevó a los investigadores a representar gráficamente a los actores y las líneas de comunicación entre ellos. En relación a las motivaciones teóricas deben citarse conceptos como los de *clique, rol*, o *estatus social*, que han llevado a los investigadores a identificarlos y definirlos en base a las redes con las que trabajaban. Finalmente, entre las motivaciones matemáticas, pueden citarse los desarrollos de la teoría de grafos, que proporciona tanto una representación apropiada como un conjunto de conceptos de utilidad para el análisis, así como el de algunos modelos de probabilidad utilizados para comparar redes teóricas con redes empíricas y modelos algebraicos utilizados para representar redes multirrelacionales.

Puede afirmarse que desde mediados de los años setenta se ha venido produciendo una institucionalización del enfoque de Redes Sociales, especialmente tras la creación de la *International Network for Social Network Analysis* y de la revista *Social Networks*,

surgidos en el ámbito de la sociología anglosajona. De todos modos, a pesar de que el enfoque vaya tomando cada vez más un carácter homogéneo en relación a las bases epistemológicas y a los instrumentos técnicos utilizados, conviven bajo este término enfoques que continúan siendo diferentes. Burt (1980; 1982; 1987) prefiere hablar de Análisis de Redes y de Modelos de Estructura Reticular o Modelos Reticulares, términos que acentúan el hecho de que no todos estos modelos se utilizan en la descripción de «relaciones sociales». De hecho, aquello que define el Análisis de Redes Sociales es el nuevo protagonismo que se confiere al concepto de *estructura*, entendiendo como tal el conjunto de relaciones entre las unidades estudiadas. Leinhardt (1977: xxx) insiste en la importancia de las *reglas estructurales*, puesto que son éstas las que «influyen haciendo que determinadas pautas de comportamiento sean probables, mientras otras pautas son menos probables». Otros autores (Adler Lomnitz, 1994; Requena, 1994) utilizan el concepto de red social o estructura de relaciones como metáfora o idea-motor, sin que se adopten los instrumentos matemáticos habituales en este enfoque, como el tratamiento matricial o la representación mediante grafos.

Esta situación no debe extrañarnos si tenemos en cuenta que el *paradigma* que podríamos definir como Análisis de Redes Sociales proviene de la convergencia de diferentes escuelas, situadas ellas mismas en distintas ramas del conocimiento científico. No haremos aquí un repaso histórico de estas diferentes tradiciones. Buenos textos que trazan los orígenes del Análisis de Redes Sociales pasando revista a las aportaciones realizadas desde diversas tradiciones son los de Scott (1991) en lengua inglesa o Lozares (1996) y Rodríguez (1995) en español.

2. OBTENCIÓN Y TRATAMIENTO DE LOS DATOS RELACIONALES

2.1. Tipos de datos

Las particularidades de la medición en el Análisis de Redes Sociales evidencian unas características distintivas que la alejan del marco analítico habitual en las ciencias sociales. Como Wasserman y Faust (1994) han señalado, la peculiaridad de este tipo de enfoque es el uso de información relacional o estructural con el objetivo de estudiar o comprobar teorías, dejándose de lado datos de carácter atributivo[1] como actitudes, opiniones o variables factuales.

[1]La incorporación de este tipo de datos a los modelos supone una complicación formal notable, ello suele inclinar a los/as autores/as a utilizar esta información de modo complementario, sin introducirla en el propio análisis de las redes.

Los datos relacionales expresan contactos, transacciones, lazos, conexiones, vínculos, servicios dados o recibidos, comunicaciones entre grupos a partir de agentes, etc. En definitiva, conectan pares de actores entre sí. Los datos son la información y la medida de esa relación. Precisamente expresan los lazos de funcionamiento entre distintos agentes.

Esta información que obtenemos sobre una red permite a la vez un tratamiento formal y una interpretación sustantiva, sin que un aspecto pueda ser separado del otro. El *contenido* constituye «la materialidad sociológica de la relación» (Lozares, 1996: 109), implica un tipo de comportamiento o acción que ha sido elegida como problemática de investigación. La *forma* es la expresión abstracta de la relación, mide tanto la fortaleza de la relación como su configuración global en forma de red. La representación formal de la relación entre un actor I y un actor J puede expresarse como z_{ij} siendo el contenido de la relación la sustantividad material de z.

De todos modos, existen diferentes niveles de medida de los datos, puesto que podríamos proponernos medir la direccionalidad[2] y la intensidad de la relación. Scott (1991: 48) establece cuatro niveles principales de medida en función de lo que él llama *directionality* y *numeration*, tal como se puede observar en el siguiente cuadro:

Tabla 1

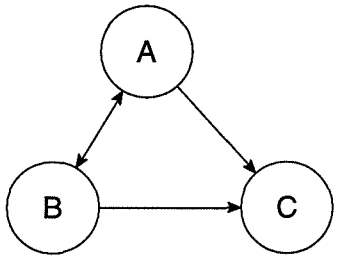
		Direccionalidad	
		No dirigido	Dirigido
Numeración	Binaria	1	3
	Valorada	2	4

La forma más simple de presentar los datos relacionales corresponde al tipo 1, en que la relación es no dirigida (véase la nota 2) y binaria (simplemente se recoge si la relación existe o no existe, indicándose con un 1 su existencia y con un 0 su no existencia). El tipo 2 corresponde a las relaciones no dirigidas pero valoradas, en que los valores de la relación indican la fortaleza de la relación más que la mera presencia[3]. El tipo 3

[2]En el caso de relaciones simétricas (no dirigidas) —por ejemplo, «ser hermano»— el orden de los subíndices es indiferente; si estuviésemos representando relaciones asimétricas (dirigidas) —por ejemplo, «prestar dinero»— este orden indica la dirección de la relación, por lo tanto la expresión z_{ij} significaría que I presta dinero a J.

[3]Esta valoración podría combinarse con un signo positivo o negativo, de modo que el signo «+» indicase una relación de carácter positivo y un signo «—» una relación de carácter negativo, por ejemplo en las relaciones comerciales entre un grupo de países. De todos modos podría argumentarse que ello supone representar dos relaciones diferentes en una misma red, en el ejemplo anterior podría establecerse una red de países deudores y otra red de países creditores.

corresponde a las relaciones dirigidas (gráficamente representadas mediante una flecha) y de carácter binario. Y el tipo 4 corresponde a las relaciones dirigidas y valoradas. Siempre es posible simplificar los datos con que trabajamos, convirtiendo las relaciones valoradas en binarias y las dirigidas en no dirigidas; aunque con estas conversiones perdemos información de carácter descriptivo, ello nos permite calcular ciertos índices que de otro modo no podríamos obtener.



	A	B	C
A	–	1	1
B	1	–	1
C	0	0	–

Figura 1. Ejemplo de red dirigida y binaria. Sociograma y sociomatriz

2.2. Muestreo y delimitación de la red

Las especiales características de los datos relacionales implican, en la práctica, la imposibilidad de realizar algún tipo de muestreo, ello se debe a que es necesario para el análisis contar con información de todos los actores y todos los lazos que componen la red social. Scott (1991: 60-63) propone un tipo de muestreo basado en la identificación de redes parciales –por ejemplo, política, económica, religiosa– entre las cuales se debería realizar un muestreo de actores siguiendo los métodos tradicionales en la investigación por encuestas. Pero como este mismo autor reconoce, citando a Alba (1982), una muestra representativa de agentes no ofrece, por sí sola, una muestra representativa de relaciones.

El resto de «criterios de muestreo» en el Análisis de Redes Sociales son en realidad criterios de delimitación de la «población».[4] Esta delimitación no está exenta de problemas, puesto que en la mayoría de relaciones siempre podemos encontrar argumentos que nos lleven a ampliar nuestra red hasta el infinito.

En la delimitación de los límites de la red existen dos aproximaciones principales. Según el enfoque *realista* los límites de la red social deben ser definidos por los propios

[4]Es decir, se selecciona un determinado grupo de actores que se supone que forman una unidad y se contabilizan todas las relaciones existentes entre ellos.

actores pertenecientes a la red que se desea analizar; esta postura parte de la suposición de que los actores son conscientes de la pertinencia a un determinado grupo y de que son capaces de identificarlo. Según el enfoque *nominalista* los límites de la red deben ser fijados por el propio investigador, no se da por hecho que los actores sean «conscientes» de una definición que se ha fijada de forma externa a ellos mismos.

Laumann, Marsden y Prensky (1983) han realizado una tipología partiendo de estas dos posturas, que llaman *perspectivas metateóricas*. Estas dos posturas son consideradas en relación a los tres conjuntos de componentes *–focos definicionales* en la terminología de los autores– que los/as investigadores/as suelen considerar para delimitar las redes sociales: los actores, las relaciones entre éstos y las actividades en que se ven envueltos.

Cruzando el número de perspectivas metateóricas con el número de focos definicionales (incluyen entre éstos el criterio basado en una combinación de los tres apuntados) Laumann, Marsden y Prensky obtienen una tipología formada por ocho estrategias de delimitación de las redes. Esta tipología (1983: 25) la presentamos en el siguiente cuadro:

Tabla 2

	Foco definicional			
Perspectiva metateórica	Atributo de los actores	Relación	Participación en actividades	Combinación de focos
Realista	I	III	V	VII
Nominalista	II	IV	VI	VIII

El criterio I es el adoptado en la mayor parte de investigaciones. En este caso la definición del grupo de actores se realiza siguiendo la definición social o institucional de pertenencia, es decir se toman todos los actores que tienen reconocida la pertenencia al grupo que se desea estudiar. Es el criterio seguido cuando se seleccionan los miembros de una fábrica, de una parroquia o los miembros de una determinada clase en una escuela.

Según el criterio II, el grupo se delimita también siguiendo una determinada característica de los actores, pero en este caso sin que esta característica coincida con una definición social o institucional reconocida. Es el criterio usual utilizado en el estudio de las élites empresariales, en que se estudian los directores, consejeros, etc. de un grupo de compañías que el analista considera, según su propio juicio, suficientemente representativas.

El criterio III, basado en la relación entre actores, es el utilizado para la identificación empírica del concepto de *grupo primario* o *grupo de pares* (Gil Calvo, 1985). La conec-

tividad completa entre sus miembros, intereses y actividades comunes, y el sentimiento subjetivo de pertenencia son condiciones que debe cumplir un grupo primario, es por ello que son los propios actores los que fijan los límites del grupo.

El criterio IV es el utilizado cuando se delimita un grupo siguiendo el método de la «bola de nieve». Son las relaciones que tienen un determinado número de actores escogidos inicialmente las que van determinando la inclusión de otros actores en el grupo que se pretende analizar. En este caso es el analista quien decide el número final de miembros incluidos en el grupo, puesto que la incorporación de nuevos miembros siguiendo este método puede no detenerse nunca. Uno de los criterios recomendados es que se llegue a un determinado nivel en que la mayor parte de los nuevos candidatos hayan sido ya citados anteriormente por otros miembros del grupo.

El criterio V es el que se sigue cuando se delimita un determinado grupo en función de las actividades en que los actores participan conjuntamente. Este criterio es el seguido cuando se desea identificar el grupo formado por los miembros influyentes en una determinada comunidad (considerando, por ejemplo, la participación en determinadas actividades públicas).

El criterio VI ha sido raramente utilizado. Es el criterio seguido en algunos trabajos que han indagado las relaciones dentro de una comunidad científica. En alguno de estos casos la identificación de los miembros del grupo se ha realizado sobre la base de la publicación de artículos en relación con un tema concreto elegido por el analista, independientemente de que los autores se conozcan entre sí.

El criterio VII es el utilizado en algunas definiciones de comunidad étnica: se combina el criterio de atributo de los actores (una determinada herencia común), junto con una interacción conjunta elevada (matrimonios, amistad).

Finalmente, el criterio VIII ha sido utilizado en la identificación de élites nacionales. Estos estudios combinan los criterios II y IV, con una selección inicial basada en alguna característica distintiva de los actores, a la que se le han añadido miembros mediante el sistema de «bola de nieve».

Como comentan Laumann, Marsden y Prensky (1983), no existe un criterio que sea claramente superior a otro. La bondad del criterio de delimitación elegido vendrá dada por la capacidad de obtener una determinada estructura de relaciones sociales con un contenido sociológico sustantivo. En este sentido señalan algunos problemas en relación a alguno de los criterios señalados. Uno de ellos es la llamada *partial system fallacy* que se produce cuando se considera que una determinada relación define las relaciones sociales en un determinado grupo, de modo que sólo se tiene en cuenta un subconjunto de actores, cuando no sabemos si existen otras relaciones más relevantes que afectan a la totalidad del grupo. Otro problema que se da en el caso de la delimitación según la participación en actividades conjuntas, es el de la elección de actividades sin ninguna

justificación clara, de modo que la selección de actores de la red tiene escaso contenido sociológico.

2.3. Recogida de los datos

El criterio de delimitación de la red que elijamos influirá, evidentemente, sobre el método de recogida de datos que utilicemos. Los datos relacionales pueden obtenerse mediante cuestionario, documentos, archivos, por observación o también por otros métodos etnográficos.

Cuando recogemos los datos relacionales mediante cuestionario —el más habitual de los métodos— tenemos diferentes alternativas (Rodríguez, 1995). El cuestionario puede contener preguntas acerca de los lazos, las relaciones, las similitudes..., tanto de los que afectan a la persona a la que entrevistamos, como respecto a otras personas sobre las cuales el/la entrevistado/a posea información. Existen tres formatos de preguntas que pueden ser utilizados en los cuestionarios (Wasserman y Faust, 1994). Por un lado podemos demandar a nuestro informante que elija los actores con quien tiene una determinada relación de entre un número cerrado de personas —poseemos un listado de todos los miembros de la red que queremos analizar— o darle la opción de citar a quien quiera —no tenemos delimitada la red y su información nos servirá para ello. Podemos solicitar un número fijo de personas —fijamos el número de relaciones que consideraremos por persona— o dejar al entrevistado libertad en su elección —tendremos un número diferente de lazos relacionales para cada actor. Finalmente, podemos solicitar al informante que ordene sus preferencias —nos puede servir para valorar las relaciones, dando diferentes puntuaciones en función del orden que se nos facilite— o podemos tratar de igual modo —no valorar— las relaciones con todos y cada uno de los actores mencionados.

Los cuestionarios se administran (cara a cara, telefónicamente, por correo) del mismo modo que en cualquier investigación que tome el punto de vista atributivo. En este punto no existen diferencias entre ambas perspectivas.

2.4. Niveles de análisis

Una vez obtenidos los datos queda finalmente aplicar los instrumentos formales de análisis habituales en Redes Sociales. Como ha señalado Lozares (1996: 108) estas herramientas son «la teoría de los grafos, operando a partir de productos cartesianos con los grafos como representación, y la teoría matricial, a partir de sociomatrices como matriz de datos inicial». De todos modos, pueden aplicarse otros métodos, compartidos con los enfoques de carácter atributivo (Wigand, 1988).

De todos modos la aplicación de estos instrumentos no determina el tipo de análisis que se realizará, puesto que como señalábamos al inicio, existen diferentes aproximaciones al Análisis de Redes, cada una de las cuales se concentra en unas determinadas características de la red, explicándolas desde unos posicionamientos teóricos concretos.

Burt (1980) ha realizado una tipología a partir de la distinción entre la aproximación *relacional* y la *posicional* a las redes sociales. Según este autor, la concepción del actor en cada una de estas aproximaciones es diferente. El enfoque relacional aborda la implicación de los actores en un determinada relación sin necesidad de explicar el resto de relaciones en las que participa; mientras que para el enfoque posicional el actor es uno más en un sistema interconectado de actores, de modo que deben considerarse todas las relaciones en que está implicado. La segunda dimensión que considera Burt en su tipología es el grado de agregación de los actores que se toma como unidad de análisis:

Tabla 3

Grado de agregación de los actores en la unidad de análisis

Aproximación analítica	Actor	Diversos actores como subgrupo de la red	Diversos actores/ subgrupos como sistema estructurado
Relacional	Red personal	Grupo de actores relacionados por relaciones cohesivas	Estructura del sistema (densidad, transitividad)
Posicional	Posición en la red	Actores estructuralmente equivalentes	Estructura del sistema (estratificación de posiciones de estatus)

Pasemos a explicar cada una de los modelos a que da lugar la tipología de Burt:

Análisis centrado en un solo actor: Desde la perspectiva relacional los análisis que se centran en un solo actor han dado lugar al desarrollo de redes egocentradas, mientras que desde la perspectiva posicional han dado como resultado modelos interesados en la posición en la red del actor. En el segundo caso las relaciones que el actor en que estamos interesados tiene con el resto de la red son tan importantes como las que no tiene, mientras que en el primer caso sólo se analizan las relaciones que posee.

Análisis centrado en subgrupos: Desde la perspectiva relacional los análisis interesados en la localización de subgrupos trabajan principalmente con el concepto de *clique*, definido en términos generales como un conjunto de actores conectados unos con otros mediante lazos fuertes. Desde la perspectiva posicional la identificación de los sub-

grupos se realiza en términos de equivalencia estructural, es decir se considera que los actores que tienen relaciones similares con el resto de la red forman un conjunto con rasgos estructurales equivalentes.

Análisis centrado en las relaciones de actores/subgrupos con la red completa: Desde la perspectiva relacional estos análisis se interesan por la densidad y la transitividad de la red, quedándose a veces en el nivel más bajo de las *díadas* y *tríadas*; mientras que en otras ocasiones se extienden al nivel de toda la red los análisis surgidos a este nivel más bajo —por ejemplo mediante el *censo de tríadas* (Holland y Leinhardt, 1978). Desde la perspectiva posicional el interés se centra en las pautas relacionales que unen a los actores en diferentes posiciones en la red, permite observar el grado de centralización (si todas las relaciones pasan por un actor central en la red) y jerarquización de las relaciones.

La elección de una u otra perspectiva dependerá básicamente de la elección metodológica y del problema sustantivo y teórico investigado. Para algunos propósitos la aproximación relacional será preferible, mientras que en otras situaciones la posicional será más interesante.

3. UNA APLICACIÓN PRÁCTICA[5]

A continuación vamos a presentar un análisis de redes sociales aplicado a lo que denominaremos como redes de «formación distribuida» en las empresas. Se trata de un análisis que se inscribe en una investigación más amplia y, por ello, en primer lugar describiremos brevemente los objetivos de la misma; en el segundo, desarrollamos los objetivos y supuestos del análisis de redes que realizamos; en el tercero, resumimos algunas apuntes del diseño; en el cuarto, presentaremos brevemente los resultados obtenidos.

3.1. Investigación en la que se inscribe el estudio presentado

El análisis de redes sociales realizado pertenece a un estudio mucho más amplio llevado a cabo por el grupo de estudios QUIT del Departamento de Sociología de la UAB, entre 1995 y 1998 (QUIT, en prensa). El objetivo global de esta investigación general ha consistido en detectar los elementos de la formación, bien de naturaleza institucional, reglada o formal, o bien de naturaleza más amplia, difusa e informal, con el fin de

[5]Un desarrollo más amplio de este apartado fue presentado en el VI Congreso Español de Sociología (C. Lozares, P. López, J. Martí, J.M. Verd, «La red formativa en una empresa mediana de textil», Federación Española de Sociología, Universidad de A Coruña, A Coruña, setiembre 1998).

valorar su eficacia que como recursos tienen en la inserción, permanencia y promoción en el empleo.

En este estudio se trabajaba a distintos niveles de análisis: desde el nivel más contextual, en el que se realizaron análisis de datos secundarios sobre formación y empleo y análisis del discurso de los agentes sociales, hasta el nivel más microsociológico, en el que se realizaron entrevistas en profundidad a varios trabajadores y trabajadoras sobre su trayectoria formativa y laboral. Pero quizás el nivel en el que más se centró el análisis fue el de la empresa puesto que, al fin y al cabo, es el ámbito donde se movilizan los recursos formativos en relación al empleo.

Para ello se realizaron estudios de caso en dos empresas del Vallès Occidental que respondían a tipologías empresariales de la comarca: una empresa mediana textil y una empresa grande de servicios. Entre otras técnicas aplicadas, en cada una de estas empresas se pasó un cuestionario a todos los trabajadores con preguntas relativas a las funciones que realizaban en la empresa y a la formación que movilizaban en su desempeño, así como a su historial formativo y laboral. En este cuestionario, como se explicará más adelante, también se incluyeron preguntas relativas a redes sociales con el fin de hacer un análisis aplicado a la empresa.

3.2. Objetivos y supuestos del análisis de redes realizado

¿Por qué este análisis? Uno de los objetivos de la investigación presentada, y concretamente el que aquí abordaremos, era el de identificar lo que denominaremos como formación distribuida entre los trabajadores de la empresa, constituida por un tejido y flujo de conocimientos, de saber hacer, de experiencia, actitudes y aptitudes compartidas. Esta formación en la empresa se manifiesta por la comunicación y la relación entre los/as trabajadores/as en la realización de tareas conjuntas o en equipo, por el mimetismo mutuo de los/as trabajadores/as en sus actitudes y prácticas, por los consejos y orientaciones comunicados, etc. Esta formación es esencial en la vida de la empresa y sirve para un mejor cumplimiento y definición de las tareas y funciones del puesto de trabajo, permite resolver situaciones de emergencia e interpretar las situaciones. En realidad es una reserva de formación acumulada en y de la empresa y que al mismo tiempo se transmite. Con relación a este contenido formativo nos planteamos dos subobjetivos: 1) El primero, consiste precisamente en analizar la estructura de esta red de formación distribuida, examinar los elementos centrales y más densos, su esquema de relación, sus puentes o intermediarios y los sujetos de su apropiación, 2) El segundo, consiste en examinar la superposición u homologación de esta estructura a la organización productiva o a la jerarquía de autoridad. De no haber tal correspondencia, al menos parcial, asistiríamos a un cierto grado de participación formativa informal con un contenido o estructura, parcial o completamente, autónomos con relación a las formas institucionales u organizativas de la empresa, más pautadas, formales y explicitadas.

La manera más intuitiva y más apropiada de modelizar esta formación invisible consiste en idearla a partir de la imagen comunicacional y relacional entre los trabajadores bajo la forma de redes que expresan los flujos de la transmisión de este «stock invisible».

3.3. Recogida de datos y diseño del análisis

La red social de la que se partía era una red cerrada, puesto que se consideraron como miembros de la red a todas aquellas personas que trabajaban en la empresa (por lo tanto, correspondería a lo que antes se ha definido como perspectiva realista basada en los atributos de los actores). Ello conlleva ciertas ventajas de muestreo, fundamentalmente la de no presentar problemas de delimitación de la población: en una empresa hay 44 trabajadores y en la otra cerca de 300; pero conlleva también inconvenientes fuertes: a saber, el condicionante de tener que entrevistar a todos y cada uno de los individuos para poder completar la red (si los individuos que faltasen ocupasen una posición periférica, ello sería un problema menor, puesto que podríamos analizar como se estructura el resto de la red; pero si los individuos que faltan son individuos centrales en el sentido que acumulan muchas relaciones, entonces su presencia es fundamental para completar la red y poderla analizar correctamente). Y este fue el problema con el que nos encontramos: en la empresa pequeña, de 44 trabajadores, se entrevistó a todos/as los/as trabajadores/as; en cambio, en la empresa de servicios apenas se obtuvo un porcentaje de respuesta del 50%, por lo que nos fue imposible construir la red de relaciones. Por lo tanto, el análisis se pudo realizar únicamente en la empresa pequeña.

Como se ha comentado anteriormente, la toma de datos se realizó mediante cuestionario. Con el fin de cubrir los objetivos propuestos, este cuestionario contenía, además de otras preguntas (analizables en una matriz de datos convencional de individuos por variables), tres preguntas relativas a las redes sociales dentro de la empresa (analizables en una matriz de relaciones de individuos por individuos). Las preguntas fueron las siguientes[6]:

1. ¿A quién o a quiénes pide orientación o consejos sobre cuestiones de trabajo?
2. ¿A quién o a quiénes ha dado orientaciones o consejos sobre cuestiones de trabajo?
3. ¿Cómo se llama Vd.?

En cada una de las dos primeras preguntas, cada persona entrevistada podía incluir el nombre de varias, una o ninguna persona de entre los/as trabajadores/as de la empresa.

[6] Además de estas tres preguntas se hicieron otras dos referidas al proceso de inserción laboral en la empresa con el fin de examinar si dicha red era fundamental en el proceso de inserción, que no se analizan en el presente artículo.

De esta forma, se obtienen dos sociogramas (uno para cada pregunta) en los que se proyectan todos los miembros de la empresa y las relaciones de donación o bien de solicitud de orientación que se establecen entre ellos. Estos sociogramas se pueden representar en la correspondiente sociomatriz de individuos por individuos, en la que el valor 1 indica la presencia de una relación y el valor 0, su ausencia. Se trata, en este caso, de una red dirigida (respeta el sentido de la relación) y binaria o no valorada (puesto que no se solicita que los nombres se ordenen por preferencias o intensidad de las relaciones).

3.4. El análisis y sus resultados

A partir de aquí, el procedimiento de análisis es relativamente simple y, aunque no lo vamos a desarrollar, se esbozarán brevemente los resultados. Partimos de dos indicadores (solicitud y donación) que, en principio, parecen ser complementarios pero que, en la práctica, se observa que ofrecen connotaciones y dimensiones bien distintas con relación al concepto de formación participada.

Así, en cuanto al indicador de DONACIÓN de orientación o consejo, la estructura de la red se asemeja fuertemente al organigrama formal de la empresa. Es decir, las personas que más orientaciones dan son las que más vinculadas están a puestos de dirección y control del proceso productivo, y las personas que más orientaciones o consejos reciben son aquéllas asignadas a los puestos de producción directa. Es, por lo tanto, un tipo de formación que se asocia a la cualificación formal, establecida y reconocida por la empresa.

En cambio, cuando abordamos el indicador de SOLICITUD de consejo, la estructura de la red es distinta. Si bien los puestos directivos continúan ocupando un papel predominante (esta vez no como donantes, sino como receptores), se observa un flujo muy denso de relaciones que ya no está vinculado a las posiciones de jerarquía, sino a la proximidad entre los trabajadores/as en la cadena productiva es decir, entre trabajadores/as de una misma sección o de secciones próximas en dicha cadena productiva. Esta red nos muestra, pues, unos flujos de formación informal (que formarían parte del denominado «curriculum oculto» de cada trabajador/a) que, aunque sean fundamentales en el día a día del proceso de producción, no son reconocidos por la empresa en términos de cualificación formal ni en términos salariales.

Para terminar, medir la riqueza atribuida a este tipo de formación distribuida o participada con únicamente dos indicadores no deja de ofrecer una visión parcial de un tal contenido. Pero, por los resultados, el procedimiento es esperanzador, puesto que es fácilmente imaginable el aumento del peso y validez de las conclusiones si se hubieran añadido otras redes en la misma dirección.

BIBLIOGRAFÍA

Adler Lomnitz, L. (1994). *Redes sociales, cultura y poder: Ensayos de antropología latinoamericana*, México D.F.: FLACSO.

Alba, R.D. (1982). «Taking Stock of Network Analysis: A Decade's Results», *Research in the Sociology of Organizations*, 1, 39-74.

Burt, R.S. (1980). «Models of network structure», *Annual Review of Sociology*, 6, 79-141.

Burt, R.S. (1982). *Toward a Structural Theory of Action: Network Models of Social Structure, Perception and Action*. Nueva York: Academic Press.

Burt, R.S. (1987). «Social Contagion and Innovation. Cohesion versus Structural Equivalence», *American Journal of Sociology*, 92, 1287-1335.

Gil Calvo, E. (1985). *Los depredadores audiovisuales: Juventud urbana y cultura de masas*, Madrid: Tecnos.

Holland, P.W. y Leinhardt, S. (1978). «An omnibus test for social structure using triads», *Sociological Methodology and Research*, 7, 227-56.

Laumann, E.O., Marsden, P.V. y Prensky, D. (1983). «The Boundary Specification Problem in Network Analysis», en Burt y Minor (ed): *Applied Network Analysis*, Beverly Hills, California: Sage Publications, 1983.

Leinhardt, S. (1977). «Social Networks. A Developing Paradigm», en Leinhardt (ed.) *Social Networks: A Developing Paradigm*. Nueva York: Academic Press.

Lozares, C. (1996). «La teoría de redes sociales», *Papers*, 48, 103-126.

QUIT (en prensa). *¿Más formación y más empleo?* Madrid: Consejo Económico y Social.

Requena, F. (1994). *Amigos y redes sociales. Elementos para una sociología de la amistad*, Madrid: Centro de Investigaciones Sociológicas.

Rodríguez, J.A. (1995). *Análisis estructural y de redes*. Madrid: CIS.

Scott, J. (1991). *Social Network Analysis. A Handbook*. Londres: Sage Publications.

Wasserman, S. y Faust, K. (1994). *Social Network Analysis*, Cambridge: Cambridge University Press.

Wellman, B. y Berkowitz, S.D. (1988). «Introduction: Studying social structures», en Wellman y Berkowitz (ed.): *Social Structures: A Network Approach*, Cambridge, Massachussets: Cambridge University Press, 1988.

Wigand, R.T. (1988). «Communication Network Analysis: History and Overview», en Goldhaber y Barnett (ed.): *Handbook of Organizational Communication*, Norwood: Abex.

ENGLISH SUMMARY

DATA COLLECTION AND SAMPLING IN SOCIAL NETWORK ANALYSIS

J.M. VERD PERICÁS
J. MARTÍ OLIVÉ
Universitat Autònoma de Barcelona*

The article reviews the proposals made by different authors inside the perspective of Social Networks Analysis regarding sampling and data collection. These aspects, satisfactorily resolved under the individualist-atomist perspective, raise some problems under the social network perspective. It is specially problematic the possibility of making representative sampling of the relations existing in a population. Even in the case of knowing all the actors and the relations among them in advance, the choice of a representative sample of actors does not guarantee a representative sampling of relations. The option taken in the article is to make the analysis of social groups considered as «populations». A practical application of this last option is offered as an example.

Keywords: Social networks, relational data, sampling

AMS Classification: 92H30, 92G99

* Grup d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (QUIT). Departament de Sociologia. Universitat Autònoma de Barcelona. Edifici B. 08193 Bellaterra (Barcelona). E-mails: Joel.Martí@uab.es, JoanMiquel.Verd@uab.es.

- Received July 1999.
- Accepted October 1999.

In the recent years, the analysis of social networks has experienced a growing popularity in the field of social sciences as an alternative to the individualist-atomist analysis. Opposite to the traditional perspective, centred on considering the individual attributes and the construction of categories based on these attributes, the social network analysis pleads for taking the relations between actors as the «material» over which the social behaviour is constructed and organised.

As Wasserman and Faust (1994) have pointed out, the particularity of this kind of approach is the use of relational or structural information with the purpose of studying or verifying theories, leaving aside data of an attributive characteristic such as attitudes, opinions or factual variables. Relational data express contacts, transactions, ties, connections, links, given or received services, communications among groups starting from actors, etc. In short, they connect pairs of actors between them.

The particularities of the measurement in the Social Network Analysis show some distinctive characteristics, moving it away from the habitual analytical framework of social sciences. Data express the functioning links between different actors; they are, at the same time, information and measurement of that relation. However, this collection may range from its simplest form (neither directionality nor intensity of the relations are collected) to the most complex one (both directionality and intensity of the relation are collected).

It turns out to be specially problematic the possibility of collecting data of a relational kind by means of sampling. The choice of a representative sample of actors does not guarantee a representative sample of relations. The hypothetical making of a representative sample would not just mean knowing all the individuals in a population but also, all the existing relations among them. But even so, a selection of the most significant actors, owing to the relations they have, means the breaking in the existing structure of relations.

As a result, the only feasible alternative in view to the impossibility of obtaining a representative sample of the relations is to select a concrete group of actors who are supposed to form a unity and give them a population treatment: counting all relations existing among them. Thus, when in the social network analysis we speak about «sampling criteria», we actually refer to «boundary criteria of population».

This boundary is not free from problems since in most of relations we can always find arguments leading us to widen our network ad infinitum. There are two main approaches to the demarcation of the network boundaries. According to the realist approach, the limits of the social network must be defined by actors themselves belonging to the network to be analyzed; this position starts from the supposition that actors are aware of belonging to a concrete group and they are able to identify it. According to the nominalist approach, the boundaries of the network must be defined by the researcher

himself/herself, it is not supposed that actors are «aware» of a definition which has been set externally to them.

Laumann, Marsden and Prensky (1983) have made a typology starting from these two positions, which they call metatheoretical perspectives. These two positions are considered in relation to the three groups of components –defining focuses according to the authors' terminology– that researchers usually consider to delimit the social networks: actors, relations among them and the activities they are involved.

The practical application we are presenting follows criteria combining the realist perspective with the defining focus centred on the actors' attributes –which, on the other hand, are the criteria adopted in most of the researches. In this case, the demarcation of the boundaries in the network is made after a social or institutional definition without the imposition of a group definition by the analyst. That is, all actors being socially or institutionally recognised as belonging to the group to study, are considered –in our case, all the actors who are recognised as members of one concrete company.

Under our point of view, despite the used delimitation criteria being one of the least problematic, the practical experience shows the added difficulties that collecting data of a relational characteristic mean. Out of the two companies chosen at the beginning to make the analysis, only in one –the smallest one, made up by 44 workers–, it was possible to interview all its members; in the second one –made up by 300 workers–, the response percentage was 50%, a fact making impossible to construct the relations' network. The fact of having to interview all actors considered as members of the network is a strong condition in the social networks analysis, only percentages close to 100% allow to make the network and only as long as the no responses affect individuals who have peripheral positions in the structure of relations.

QÜESTIIÓ, vol. 23, 3, p. 525-541, 1999

LA ARTICULACIÓN ENTRE LO CUANTITATIVO Y LO CUALITATIVO: DE LAS GRANDES ENCUESTAS A LA RECOGIDA DE DATOS INTENSIVA

V. BORRÀS
P. LÓPEZ
C. LOZARES
Universitat Autònoma de Barcelona*

Son casi innumerables las reflexiones que se han hecho en el campo de la metodología de las ciencias sociales, sobre la dicotomía, real o inexistente, entre las perspectivas de análisis cuantitativo y cualitativo, mientras que han sido menos abundantes los trabajos teóricos, aunque van siendo más frecuentes los empíricos, que han tratado de compatibilizar y/o complementar ambas perspectivas. En muchos casos los trabajos cualitativos cubren solamente los primeros pasos de la investigación social, como fases previas o exploratorias de la misma, para prolongarla después con una metodología más cuantitativa y extensiva como la inherente a las encuestas, al análisis de datos, multivariados, causales u otros. Este artículo presenta un ejemplo concreto del planteamiento inverso. Se parte de una gran encuesta con el objetivo de estructurar la realidad objeto de estudio, para posteriormente aplicar técnicas cualitativas, concretamente grupos de discusión.

The articulation between the quantitative and the qualitative: from the great surveys to the intensive data collection

Palabras clave: Análisis de datos, análisis de correspondencias, análisis de clasificación, metodología, consumo, técnicas cuantitativas y cualitativas

Clasificación AMS: 62D05, 62-07, 62H25, 62H30

* Vicent Borràs, Pedro López Roldán i Carlos Lozares Colina, Departament de Sociologia, Universitat Autònoma de Barcelona. Los tres autores son miembros del *Grup d'Estudis Sociològics sobre la Vida Quotidiana i el Treball* (QUIT) del Departament de Sociologia de la Universitat Autònoma de Barcelona.

– Recibido en agosto de 1999.
– Aceptado en noviembre de 1999.

1. INTRODUCCIÓN: EL PLANTEAMIENTO DE LA DICOTOMÍA

Las investigaciones de los sociólogos y sociólogas se caracterizan, entre otras cosas, por lo que ellos mismos denominan el pluralismo metodológico; es decir, por las diferentes orientaciones metodológicas y la correspondiente elección de métodos y técnicas provenientes de distintos horizontes epistemológicos. La división más socorrida y recurrente es la que establecen entre métodos cuantitativos y cualitativos. Esta diferente estrategia de investigación sirve, incluso, para clasificar a los profesionales de la sociología entre cuantitativistas y cualitativistas según la perspectiva que adopten.

Esta división, un tanto simplista, hunde sus raíces en la ya clásica separación entre los métodos de las ciencias de la naturaleza y los de las ciencias humanas. Tratando de esquematizar al máximo una polémica, larga e intensa en la historia de la sociología, se puede decir que la perspectiva cuantitativa sigue más de cerca las pautas habituales del llamado método científico, más propio de las ciencias de la naturaleza y/o experimentales: objetivación y delimitación del objeto de estudio, medición y formalización de conceptos, variables y datos, modelización de hipótesis y teorías, validación y fiabilidad de resultados por tests de ajuste o de bondad, etc. Por el contrario, la perspectiva cualitativa se basa más, como objeto propio de estudio, en el sentido o significado que para el actor o agente social, (y para el mismo investigador) tienen los fenómenos sociales así como en estudios sobre contenidos microsociológicos vinculados a la interacción social, a la intersubjetividad, sobre los grupos primarios, sobre el lenguaje, etc. La orientación cuantitativa utiliza una serie de técnicas de recogida de datos como la encuesta, de procedimientos de tipos experimental o cuasiexperimental al mismo tiempo que los formalismos de análisis algebraicos y/o estadísticos. La perspectiva cualitativa está más asociada a métodos y técnicas de recogida de información de base más etnográfica y/o de intervención o participativa como las entrevistas, la observación participante, las historias de vida, los grupos de discusión, la investigación-acción y sus análisis están más ligados a los de contenido, del discurso y/o hermenéuticos.

La polémica no se refiere sólo a los métodos y técnicas diferenciadas de investigación sino que tiene unas raíces más profundas y extensas de índole epistemológico y metodológico. La dicotomía atraviesa además otros órdenes de naturaleza conceptual y teórica: las distinciones entre macro y microsociología, estructura e interacción, objetividad y subjetividad. Cualquier intento de disolver o mediar en la dicotomía entre los distintos métodos y técnicas ha de pasar no sólo por una reflexión práctica sino también por la epistemológica y la teórica. En este sentido se han dado avances importantes a finales de los 80 y los 90 aunque se iniciaron ya en los 70. Por ejemplo, y refiriéndonos de manera sintética y esquemática a la distinción entre macro y microsociología, Ritzer (1993), da un panorama completo y fundado de los estudios en esta dirección. En particular en la década de los 80 se presentan, según Ritzer, posturas de acercamiento a partir de posiciones microteóricas o individualistas, como p.e. la teoría de la elección racional de Coleman; las cadenas rituales interaccionales de Collins y los análisis de

Knorr-Cetina y Cicourel. Estas posiciones de acercamiento mutuo se dan también desde la macrosociología como p.e. el enfoque multidimensional de Alexander partiendo de una base estructural-funcional; la entropía social de Bailey derivada de la teoría de los sistemas, la integración de Burt, a través de su teoría relacional de la acción y de las redes sociales. Estos teóricos comienzan a reconocer la importancia del individuo, o agente/actor activo, y de la interacción dentro de la estructura social. Otros teóricos de tendencia originariamente macro han incorporado algunos conceptos de la microsociología introduciendo así una alternativa más abierta y han comenzado a reflexionar con una visión más sintética intentando establecer puentes entre la macro y la microsociología p.e. Alexander, Collins, Bourdieu y Giddens. Estos tratan particularmente de implicar la acción del individuo en la estructura lo que origina la consideración de la acción como una relación entre un contexto formado y un contexto «formante».

No se trata aquí de enumerar ni relatar todos los aspectos y críticas mutuas entre ambas orientaciones de investigación; nuestro propósito es más sencillo: consiste más bien en buscar aquellos puntos de encuentro empíricos y pragmáticos que faciliten una posible complementariedad o conjunción entre la dinámica y el proceso de investigación cuantitativo y cualitativo presentando un ejemplo concreto de aplicación. De todas formas y antes de presentar la aplicación interesa hacer algunas reflexiones, aunque sean someras, sobre la posible confluencia, convergencia o complementariedad entre los métodos y técnicas cuantitativas y cualitativas.

2. ALGUNAS REFLEXIONES METODOLÓGICAS SOBRE LA APROXIMACIÓN, CONVERGENCIA O COMPLEMENTARIEDAD

2.1. La necesaria convergencia y aproximación

Las aproximaciones que se han realizado hasta ahora, como señalábamos al principio del artículo, entre las perspectivas aquí tratadas han sido más abundantes y dedicadas a la estrategia y a la práctica concreta de investigación que a una reflexión fecunda y profunda, de aire más metodológico, que tenga en cuenta todo el proceso de investigación y producción del conocimiento.

Kaplan (1964) indicó una serie de características o claves de distinción entre las dos perspectivas. Una de ellas es la que hace referencia a la dicotomía entre explicación y comprensión. Según él, la perspectiva cuantitativa hace hincapié en la objetivación, medición, explicación por causas y validación, esto es, en el «porqué» de los hechos sociales estudiados. Por el contrario, la cualitativa pone el acento en la captación comprensiva del sentido dado por los actores sociales y en la intersubjetividad de los fenómenos

sociales; se preocupa más del «cómo (y cuáles son los procedimientos por los que) suceden las cosas».

Para Kaplan uno de los objetivos y los retos de las ciencias sociales consiste en llegar a codificar el conocimiento y la intención personal además, evidentemente, de las acciones y/o interacciones manifiestas y explícitas. Ello es debido a que no podemos prescindir de los procesos intencionales y cognitivos ya que intervienen en la misma base de la interacción social. La idea de dejar de lado los elementos subjetivos o cognitivos haría incompleto cualquier objeto de estudio social. Como además es una exigencia misma de la ciencia la transmisión de los resultados y procedimientos de investigación se requiere que los contenidos estudiados se expliciten, se desvelen o se publiciten. Ello hace imperativo un nivel mínimo de covergencia metodológica[1].

De todas maneras la introducción de la informática y los nuevos desarrollos algebraicos y estadísticos (redes sociales, conjuntos borrosos y aproximados, inteligencia artificial distribuida –redes neuronales, etc.– teoría del caos) está precipitando y posibilitando la introducción en las ciencias sociales, (dentro de las cuales la sociología se encuentra en el más manifiesto de los «retrasos»), de procedimientos formales más flexibles y paralelos y más adaptados a la complejidad y contextualización de los fenómenos sociales y, por tanto, de los contenidos cualitativos. Somos de la opinión que la corriente que va en la dirección de esta mutación metodológica en sociología es o ha de ser imparable pues del éxito de su aplicabilidad y acoplamiento dependerá la subsistencia de la sociología en cuanto diferenciada de otros procedimientos exclusivamente interpretativos.

De todas formas, aunque todo ello suponga, desde nuestro punto de vista, un futuro prometedor para la sociología no termina de resolver el problema de la conjunción y/o complementariedad de las dos metodologías en todas sus fases, y no exclusivamente en la de análisis, y sobre todo en lo que se refiere a la calidad y contextualización de la información recogida, previa a cualquier tipo de análisis. Parece que se cumple una ley inversa entre las dos metodologías y las fases de la investigación en el sentido de que en la fase de recogida de información los métodos y técnicas propios de la metodología cualitativa son más próximos, complejos y válidos para dar cuenta del fenómeno social en su unidad y totalidad, mientras que los métodos cualitativos de análisis son menos fiables y válidos; por el contrario, la fase primera parece más endeble y sesgada en los métodos cuantitativos mientras que éstos son más consistentes, replicables y objetivados en la fase del análisis.

[1]Para más información ver las referencias de Layder (1993) Lozares, Martín y López (1998), Bericat (1998), Alvira (1983) Conde (1987), Ibañez (1979) Miles y Huberman (1984).

2.2. Hacia un bucle retroalimentado y en espiral

a) De lo cualitativo a lo cuantitativo.

Una de las formas más pobres de convergencia entre métodos cuantitativos y cualitativos es la que se ha seguido tradicionalmente de una forma pragmática por la cual la recogida de información, los métodos e interpretaciones de tipo cualitativo son sólo necesarios, pertinentes e importantes en las primeras fases de la investigación sociológica: es decir, en la fase de la búsqueda de la problemática y de los objetivos, en el bombardeo de ideas y proposiciones, en la contextualización de fenómenos, en el acercamiento a la práctica, comunicación y lenguaje de los investigados, en la inserción, inmersión o integración en su realidad cotidiana, en la búsqueda conceptual, tipológica y discursiva, etc., es decir, en lo que algunos llaman la fase exploratoria, inicial, piloto o de pretest de la investigación. El esquema de este proceso es el siguiente:

ANÁLISIS CUALITATIVO						ANÁLISIS CUANTITATIVO
Análisis cualitativo: exploratorio	→	Selección conceptual: variables e indicadores	→	Realización encuesta	→	Análisis cuantitativo: algebraico, estadístico

Este proceso representa, de manera manifiesta, una forma desequilibrada y simplificada de resolver la convergencia y complementariedad de la relación entre metodologías y supone además algunos problemas añadidos de orden metodológico. Uno de ellos consiste en la dificultad de aplicar correctamente la complementariedad, cuando el objeto de estudio es de naturaleza más compleja, contextualizada y con una carga más interactiva y/o subjetiva. En este caso puede ser difícil la reducción del objeto de estudio a ítems o índices, a su algebrización, a su delimitación sin romper la estructura, articulación y sentido del mismo. Esto es, la totalidad, unidad y vinculación contextual que suponen y configuran los fenómenos estudiados quedan sin significado por el efecto de la descomposición en sus elementos que conlleva la reducción y disección que exige el análisis cuantitativo.

b) De lo cuantitativo a lo cualitativo.

Para resolver este problema, Conde (1987), propone la aproximación al objeto de estudio a partir de los dos procedimientos consecutivos pero de forma invertida al precedente: primero, a través del análisis cuantitativo de correspondencias múltiples (ACM), y del análisis de conglomerados o de clasificación (AC), previo paso por las fases de preparación de variables, propuesta de hipótesis y encuesta, propias todas ellas del análisis cuantitativo; segundo por un procedimiento de metodología cualitativa como es el método de grupos de discusión. Su propuesta se basa en una analogía isomorfa de espacios topológicos.

El análisis de correspondencias posibilita mediciones y lecturas topológicas de los objetos, puesto que se basa en relaciones ordinales entre los objetos, mediciones expresadas en cantidades extensivas no métricas y formalizadas mediante lenguaje geométrico y no solo algebraico, estudiando las similitudes no entre magnitudes absolutas sino entre formas. Esto permite ver un sistema que articula y estructura al conjunto de objetos y fenómenos estudiados.

Por otro lado, el análisis cualitativo es formalizable desde el punto de vista topológico. Si tomamos la construcción de los grupos, concretamente de los grupos de discusión, se puede formalizar y construir un mapa de grupos ordenado según su posición en la estructura social más global. Así mismo a partir del análisis de los discursos, se pueden sintetizar los procesos y estructuras desarrollados a lo largo de la dinámica, pudiendo topologizarse en un mapa de discursos ordenados según las dimensiones que los articulan (Conde, 1987).

La propuesta que aquí presentamos abunda y prolonga esta dinámica, esto es, utilización de datos y análisis de tipo cuantitativo en una primera fase para pasar a la obtención de datos y análisis cualitativos en una segunda fase.

En una primera fase, por el tipo de medición y análisis que supone el ACM, permite representar la topología de los resultados mediante un lenguaje geométrico y no sólo algebraico. Posibilita estudiar las similitudes y diferencias entre las formas de los objetos y observar y analizar cómo se articula y estructura el sistema de datos del fenómeno estudiado.

En una segunda fase, una vez representados, identificados y caracterizados, en sí y dentro de la totalidad de la población, los grupos provenientes del análisis topológico y factorial del ACM, se extrae un conjunto de individuos en cada uno de los grupos que personifiquen y «representen» dichos grupos portando, en el grado más eminente posible, sus atributos. De alguna manera hemos construido, con una tal densidad informativa, unas nuevas entidades-identidades sociales de las que podemos suponer que son portadoras de un tipo de discurso (representaciones, interpretaciones, ideología, proyecto, etc.) extraído de (y sobre) la realidad social. Para descubrir e interpretar dicho posible y supuesto discurso interviene la fase cualitativa por medio del procedimiento metodológico y técnico del los grupos de discusión. La discusión y/o intercambio conversacional se lleva a cabo entre los miembros escogidos de dichos grupos. Una vez recogida la información, o sea la conversación entre ellos, sobre temáticas pertinentes y prefijadas, se procede a su análisis, normalmente por medio de análisis de contenido o del discurso o del hermenéutico u otros.

Esta propuesta lleva un camino opuesto al precedente en el sentido de que la fase más cuantitativa se aplica, en cierta manera, como «instrumento» para la segunda. En este esquema la metodología cuantitativa está situada en la primera fase mientras que la intervención e interpretación se lleva a cabo en la segunda, según el siguiente esquema.

ANÁLISIS CUANTITATIVO						ANÁLISIS CUALITATIVO
Análisis cuantitativo: grupos y/o tipos	→	Selección de individuos: muestreo cualitativo	→	Métodos de intervención: Grupos de discusión	→	Análisis cualitativo: análisis de contenido,

Con este tipo de planteamiento y proceso no se trata de utilizar los dos tipos de técnicas de recogida de información y análisis de manera independiente y yuxtapuesta para realizar más tarde una especie de isormorfismo topológico y proyectivo de acoplamiento. La propuesta pretende encontrar un hiato o vínculo metodológico y de contenido en la misma construcción del objeto entre dos niveles o planos del mismo que suponemos articulados. Por ello la complementariedad metodológica y técnica es más vinculante y adquiere más sentido que en el primer esquema.

c) El bucle retroalimentado y espiral.

Aunque la investigación presentada sigue el esquema precedente la propuesta más completa debería ser más exhaustiva y acumulativa y exigiría:

1) de un lado, siguiendo la línea de la investigación multiestratégica[2]
   a) descomponer el objeto de estudio en diferentes planos elaborando la estrategia metodológica y las hipótesis y contenido en y para cada uno de los planos (p.e. contexto, situación, interacción, el «en sí» de los actores, dimensiones temporales)
   b) elaborar, proponer e interpretar y/o validar hipótesis de vinculación entre los planos definidos y,
2) de otro lado, continuar de manera circular y consecutiva el intercambio entre métodos hasta que las exigencias de las hipótesis o la saturación de los resultados lo consuman pero, siempre, utilizando los resultados parciales del proceso precedente para profundizar en el consecuente a la manera de una espiral que se supera a sí misma.

ANÁLISIS CUALITATIVO 1		ANÁLISIS CUANTITATIVO 1		ANÁLISIS CUALITATIVO 2
Análisis cualitativo: Exploratorio	→	Análisis de Correspondencias: Tipos y grupos	→	Grupos de discusión

Este proceso tiene una triple función:

1. La vinculación de los diferentes planos del objeto de estudio: el que proviene de las características de los grupos o tipos sociales surgidos del estudio cuantitativo (o de

[2]Para ampliar este concepto se puede consultar Layder (1993) y Lozares, Martín y López (1998).

indicadores sociales) y el contenido discursivo de dichos grupos que proviene de los grupos de discusión.

2. El enriquecimiento sucesivo y mutuo, así como la acumulación informativa y validada, tanto de los componentes objetivables como subjetivos presentes en todo fenómeno social.

3. La mutua y recíproca validación interna y externa de los resultados obtenidos entre y en cada uno de los procedimientos en una especie de triangulación «sui generis».

3. APLICACIÓN A UNA INVESTIGACIÓN SOBRE EL CONSUMO

La utilización de ambos métodos y técnicas en una investigación sobre los hábitos de (y discursos sobre el) consumo nos ha parecido ejemplar para la propuesta que hacemos. La exclusiva aplicación de la encuesta y sus correspondientes métodos de análisis se hubiera limitado a un único plano del consumo, el que nos permite captar su vertiente más extensiva y distributiva de los productos: tasas de consumo, situaciones de hecho o comportamientos verbales mecánicamente socializados, etc. Como señala Ibáñez (1979) restringirse solamente al caparazón o superficie manifiesta de los hechos y comportamientos, aunque esté racionalizada y metrizada, sólo consigue la descripción de las situaciones que se observan sin aprehender y comprender los componentes discursivos e intencionales y, por tanto, dejando de lado aspectos decisivos de las tendencias de fondo y proyectivas de los fenómenos sociales.

La investigación que aquí resumimos muy someramente, ha llenado dos fases que presentamos seguidamente.

3.1. Estructuración del consumo y posiciones de clase

En una primera fase se plantea un doble objetivo. Primero, el de demostrar que los comportamientos del consumo presentan una estructura y una articulación interna y, segundo, mostrar que el consumo, en sus aspectos más atributivos, es el reflejo de las posiciones de clase. Los datos de los que partimos son los de una gran encuesta como la *Enquesta de la Regió Metropolitana de Barcelona 1990*[3].

Para conseguir este primer objetivo hemos utilizado el análisis de correspondencias múltiples. Esta técnica, al operar sobre las interacciones conjuntas de perfiles contex-

[3]Investigación realizada bajo la dirección general de Marina Subirats y la dirección metodológica de Carlos Lozares .

tuales globales e individuales, entre filas y columnas, responde de forma específica a los planteamientos teóricos y metodológicos propios de las relaciones entre atributos y sus relaciones.

Partimos, según criterios teóricos y operativos, de la división del conjunto de las variables del campo del consumo en áreas temáticas: la de los bienes duraderos, de la vivienda y hábitat, del parque móvil, del consumo cultural, de la movilidad de compra, de la distribución del gasto, del ocio, de las vacaciones y de las modalidades de compra. Con las variables que representan cada una de estas áreas, y en cada una de ellas, se busca un conjunto de dimensiones, sus factores más representativos, obteniendo así el contenido de la estructura correspondiente a cada área.

Utilizando los ejes pertinentes en cada área se realiza, en cada una de ellas, un análisis tipológico para ver los grupos formados con relación a los factores precedentes que las estructuraban. Este procedimiento se lleva a cabo por el análisis de clasificación.[4]

Una vez tenemos todas las tipologías de los diferentes ámbitos pasamos a realizar el análisis final de todas ellas para ver la estructuración global, es decir, cuál es o cuáles son las diferentes dimensiones en las que se estructuran y articulan todos los ámbitos de consumo tratados. Como resultado esquematizamos solamente los resultados del primer eje que se recogen en el gráfico del anexo.

En el primer eje que explica el 62% de la varianza total tenemos una clara polaridad: en un extremo nos encontramos los que consumen de una forma precaria, tienen los niveles más bajos de consumo cultural, no tienen coche ni disfrutan de vacaciones, el nivel de equipamiento de los hogares es el básico. En el otro extremo tenemos a los que viven en hogares de lujo y bien equipados, practican un ocio rico, están bien motorizados y disfrutan de unas vacaciones más largas, viajando al extranjero. Vemos claramente en este eje que los diferentes ámbitos presentan una estructuración y una articulación propia que está por encima del propio ámbito, es decir, que los diferentes aspectos que componen el consumo pueden ser tratados de forma global dándose a conocer estructuras más globalizadoras. El fenómeno del consumo cruza todos los aspectos a él referidos, configurando realidades multidimensionales. Tener una vivienda amplia va unido a estar bien equipada, a poseer al menos un coche y a disfrutar de un ocio de élite. El no poseer cruza tanto los ámbitos culturales como los más vinculados a necesidades primarias.

En un segundo momento hay que ver qué vinculaciones con la posición de clase[5] tienen estos comportamientos del consumo, para ello hemos añadido al análisis la variable clase ocupacional. Tenemos que en un extremo, junto a niveles bajos de consumo

[4]Para un mejor seguimiento del proceso ver Borràs (1996).

[5]La variable de clase utilizada parte del estudio previo realizado por Subirats, Sánchez y Domínguez (1992)

cultural, ocio pobre y no disfrute de vacaciones se sitúan los inactivos, las amas de casa y los trabajadores agrarios, junto a éstos y en un nivel más elevado de consumo se encuentran los obreros no cualificados y los cualificados. En el otro extremo y junto a prácticas de consumo como ocio de élite, viviendas de lujo y bien equipadas, se sitúan los profesionales liberales, los directores y gerentes de empresas y los técnicos altos. En las posiciones intermedias se encuentran lo comerciantes, los autónomos, los empleados y los contramaestres y capataces.

Las hipótesis propuestas para esta primera fase quedan demostradas; por un lado, tenemos que las prácticas de consumo presentan una articulación y una estructura interna y, por otro lado, dichas prácticas van ligadas a la posición de clase. Si nos quedamos en este nivel de análisis nos habremos quedado en un primer nivel y no habremos profundizado ni intentado ver las estructuras de fondo. Preguntas como ¿Cuáles son las representaciones y las simbologías de clase respecto de aquello que consumimos? ¿El imaginario de clase contribuye a explicar el comportamiento respecto del consumo?. Para ello debemos de pasar a un segundo apartado que pueda ayudarnos a dar respuesta a estas cuestiones.

3.2. La segunda fase o los discursos sobre el consumo a partir de los grupos de discusión

Para esta segunda fase hemos seguido, como venimos señalando, el método de grupos de discusión. Los individuos que configuran cada grupo de discusión se extraen por una muestra cualitativa de los grupos obtenidos en los análisis por ACM y AC de la fase precedente. Por ello, se adquiere una plausible y cualitativa validez externa en el sentido de que las personas elegidas de cada grupo son representativas de la identidad del grupo ya que ha sido construido precisamente con un elevado grado de homogeneidad o similitud entre sus miembros. En nuestra investigación de aplicación hemos realizado solamente los grupos de discusión en dos de los grupos hallados precedentemente entre los cuatro de consumo/clase social: uno de clase trabajadora y otro de clase media, que ocupan una situación polar. Aunque cada grupo es homogéneo dada su pertenencia a una clase social y sus prácticas de consumo, sin embargo se ha introducido alguna variedad en su interior al considerar también en ellos las variables de sexo, edad y origen familiar (catalán de nacimiento o catalán nacido fuera de Catalunya).

Los resultados obtenidos en esta segunda fase cubren el objetivo de mostrar cómo opera la percepción y la representación social que los individuos poseen de sí mismos y de los grupos dominantes en el consumo. De tal forma que se configuran dos modelos de consumo diferenciados según el grupo social del que se hable.[6]

[6]Para una explicación de los resultados obtenidos en esta investigación se puede consultar Borràs (1998).

El modelo de consumo de las clases trabajadoras se basa en el concepto de necesidad como punto de partida y referente constante tanto por lo que se refiere a su realidad cotidiana, aquello posible y alcanzable, como para la realización de sus deseos e ilusiones. La austeridad, es decir, la autoadministración y el control riguroso en sus gastos es el aspecto que marcan la práctica en un intento, constante y permanente, de huir de la precariedad y la escasez.

El modelo de consumo de las clases medias, en cambio, está fundamentado en el deseo, en que no hay privación ni restricción; su punto de mira es opuesto al de las clases trabajadoras. No tiene como referente la escasez y la precariedad, sino que el referente se encuentra en todo lo que se puede conseguir, todo lo que se puede alcanzar, todos los bienes y servicios que se pueden consumir.

Para ambos grupos la familia sigue siendo un referente constante, aunque con más peso en las clases trabajadoras; la familia es una plataforma de solidaridad. Por ello, la referencia a sus miembros condiciona sus propósitos y deseos, al ser el ámbito primario por el que pasan sus consumos. Existen diferencias por género, esta referencia familiar está mas presente en las mujeres que en los hombres.

Aunque no entremos de lleno en ello esta investigación también ha mostrado la vinculación entre las vivencias y representaciones del trabajo (tanto el remunerado como el denominado doméstico familiar) y las prácticas de consumo. Para ambos grupos el trabajo productivo está presente en su universo de lo deseable, ya sea para prescindir de él, porque es visto como una carga, en el caso de las clases trabajadoras, o para convertirse en un medio de autorrealización como ocurre con las clases medias. Respecto al trabajo de la reproducción es vivido como una carga sobre todo para las mujeres de las clases medias; su ilusión pasa por poder contratarlo a terceros. Por tanto, el universo del consumo está mediatizado por la vivencia y representación del trabajo, sirviendo tanto en las identificaciones y percepciones que se tienen de los grupos dominantes como en las suyas propias. Definir cómo se consume y cómo les gustaría consumir implica definir qué relación tienen y les gustaría tener con el trabajo, tanto el referido a la producción como a la reproducción.

4. CONCLUSIONES

Las aproximaciones que se han realizado hasta ahora, entre las perspectivas cuantitativa y cualitativa han estado más abundantes y dedicadas a la estrategia y a la práctica concreta de investigación que a una reflexión fecunda y profunda, de aire más metodológico, que tenga en cuenta todo el proceso de investigación y producción del conocimiento. Por otro lado, parece que la fase de recogida de información de los métodos y técnicas propios de la metodología cualitativa son más próximos, complejos y válidos para dar cuenta del fenómeno social en su unidad y totalidad, mientras los métodos de

análisis son menos fiables y válidos; por el contrario, la fase primera parece más endeble y sesgada en los métodos cuantitativos mientras que éstos son más consistentes, replicables y objetivados en la fase del análisis.

La propuesta que nosotros presentamos pretende pues encontrar un vínculo metodológico y de contenido en la misma construcción del objeto entre dos niveles o planos del consumo que suponemos articulados. Por ello la complementariedad metodológica y técnica es más vinculante y adquiere más sentido que en los esquemas más tradicionales o clásicos

El ejemplo de aplicación ha tratado de mostrar cómo es posible invertir el proceso «más corriente» en la investigación social, que va de lo cualitativo como exploración a lo cuantitativo como análisis, en tanto que procedimiento predominante. Como decíamos, este proceso presentado puede continuar planteándonos nuevas preguntas y objetos de estudio a partir de los resultados cualitativos.

Lo importante de todo ello es que sobre un objeto de estudio construido a partir de indicadores/atributos de naturaleza cuantitativa y métrica, agrupados en el análisis, hemos superpuesto , con un contenido cualitativo, una identidad de sentido compuesta de deseos, imágenes, proyectos e intenciones. Así la realidad primaria se enriquece y densifica y quizás revele y muestre más los aspectos de los actores sociales que tienen los grupos de consumo que los simples agregados o conglomerados sociales.

BIBLIOGRAFÍA

Alonso, L.E., Callejo, J. (1992). *Consumo e individualismo metodológico: una perspectiva crítica*, Madrid, IV Congreso Español de Sociología.

Alvira Martín, F. (1983). «Perspectiva cualitativa-perspectiva cuantitativa en la metodología sociologica», *Revista Española de Investigaciones Sociológicas*, 22, 53-75.

Baudrillard, J. (1970). *La Société de Consommation.* Paris: Le Point de la Question.

Baudrillard, J. (1976). *La génesis ideológica de las necesidades.* Barcelona: Editorial Anagrama.

Baudrillard, J. (1987). *Crítica de la economía política del signo.* Madrid: Siglo XXI.

Bericat, E. (1998). *La integración de los métodos cuantitativo y cualitativo en la investigación social.* Barcelona: Ariel.

Borràs, V. (1995). «L'estrucutació interna del consum. La lògica del consum com a lògica de comunicacio». *Papers. Revista de Sociologia*, 47.

Borràs, V. (1996). «L'estrucutació del consum a través de l'anàlisi de correspondències». *Papers. Revista de Sociologia*, 48.

Borràs, V. (1998). *El consumo, un análisis sociológico.* Barcelona. Cedecs.

Bourdieu, P. (1974). «Les fractions de la classe dominanter et les modes d'appropriation de l'oeuvre d'art». *Information sur les Sciencies Sociales*, 13, 3, 7-31.

Bourdieu, P. (1978). «Sport and social class». *International Social Sciencie Council*, 17, 6, 819-840.

Bourdieu, P. (1988). *La distinción. Criterio y bases sociales del gusto.* Altea: Taurus Humanidades.

Castillo Castillo, J. (1987). *Sociedad de consumo a la española.* Madrid: Eudema Actualidad.

Conde, F. (1987). «Una propuesta de uso conjunto de las técnicas cuantitativas y cualitativas en la investigación social. El isomorfismo de las dimensiones topológicas de ambas técnicas», *Revista Española de investigaciones Sociológicas*, 39, 213-224.

Conde, F. (1990). «Un ensayo de articulación de las perspectivas cuantitativa y cualitativa en la investigación social», *Revista Española de Investigaciones Sociológicas*, 51, 91-117.

Cornejo, J.M. (1988). *Técnicas de investigación social: El análisis de correspondencias (teoría y práctica).* Barcelona: PPU.

Elias, N. (1982). *La sociedad cortesana.* México: Fondo de Cultura Económica.

Galbraith, J.K. (1992). *La Cultura de la Satisfacción.* Barcelona: Ariel.

Ibañez, J. (1979). *Más allá de la sociología. El grupo de discusión: Técnica y crítica.* Madrid: Siglo XXI.

Ibañez, J. (1987). «Una publicidad que se anuncia a sí misma». *Telos*, 8, 117-123.

Ibañez, J. (1994). *Por una sociología de la vida cotidiana.* Madrid: Siglo XXI.

Kaplan, A. (1964). *The conduct of inquiry.* San francisco: Chandler.

Katona, G. (1968). *La sociedad de consumo de masas.* Madrid: Rialp.

Layder, D. (1993). *New Strategies in Social Research.* Cambridge: Polity Press.

Leonini, L. (1990-91). «A che servono le cose? A proposito di due contributi sui consumi». *Quaderni di Sociologia*, 29, 168-178.

López, P. (1996). «La construcción de tipologías: metodología de análisis». *Papers. Revista de Sociologia*, 48, 9-29.

Lozares, C. (1990). «La tipología, más allá de la taxonomía. Conceptualización y cálculo». *Papers. Revista de Sociologia*, 34, 139-165.

Lozares, C. (1992). «La complementarietat quantitativa i qualitativa: un cas d'Analisi dels temporers de Ginebra». En: *Tècniques qualitatives en Ciències socials.* Barcelona: Societat Catalana de Sociologia, 27-36.

Lozares, C. (1992). «La complementarietat quantitativa i qualitativa: un cas d'anàlisi del medi popular a Suïssa». En: *Tècniques qualitatives en Ciències socials*. Barcelona: Societat Catalana de Sociologia, 17-25.

Lozares, C.; Martín, A. y López, P. (1998). «El tratamiento multiestratégico en la investigación sociológica». *Papers. Revista de Sociologia*, 55, 27-43.

Lozares, C.; Murman, A.; Pecorini, M. y de Rham, G. (Group GRISOC) (1988). *Portrait des salariés romands*. Lausanne: Editions d'En Bas.

Marafioti, R. (1988). *Los significantes del consumo. Semiología, medios masivos y publicidad*. Buenos Aires: Biblos.

McCracken, G. (1988). *Culture and consumption*. Bloomington: Indiana University Press.

Miguelez, F. y Torns, T. (1992). *Treball, condicions econòmiques i formes de consum*. Enquesta de la Regió Metropolitana de Barcelona 1990. Volum 2. Barcelona: Institut d'Estudis Metropolitans de Barcelona.

Miles, M.B. y Huberman, M. (1994). *Qualitative Data Analysis*. Thousand Oaks, London, New Delhi: Sage Publications.

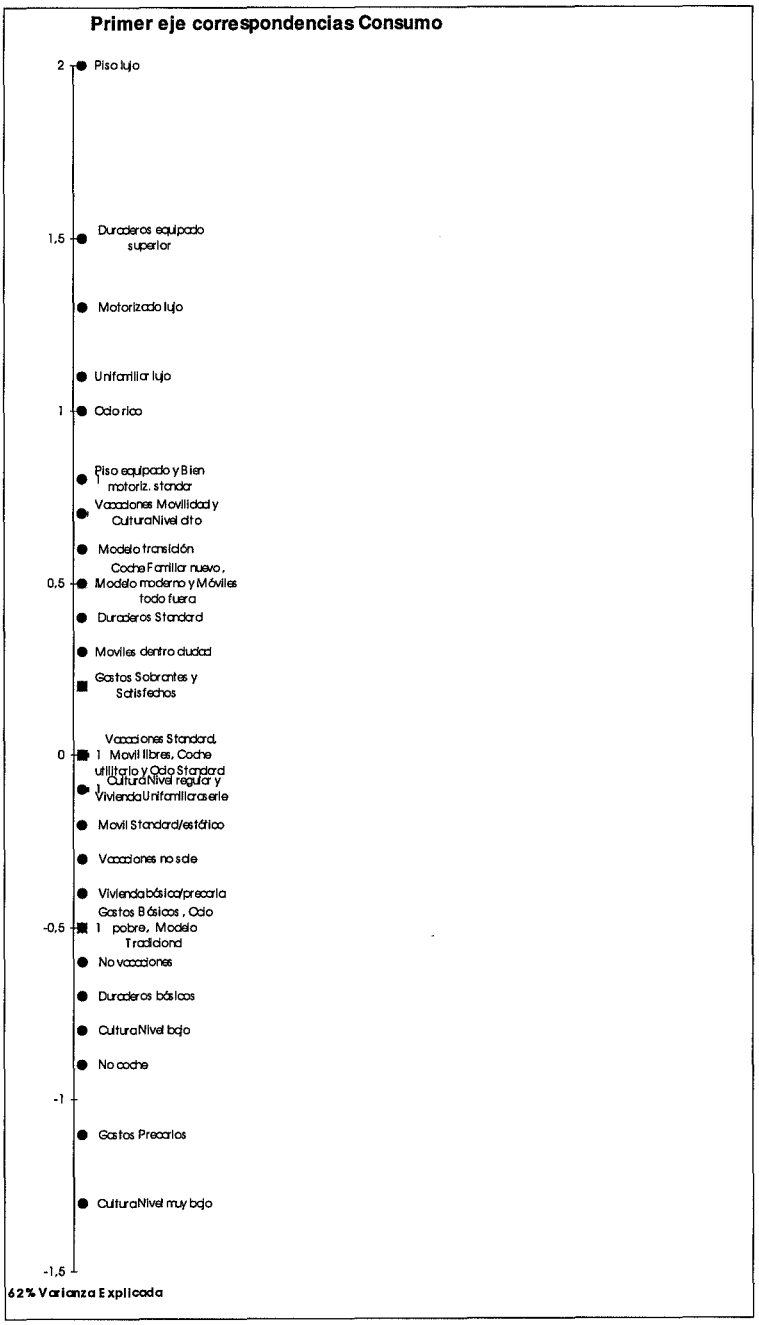
Ragin, Ch.C. (1994). *Construction Social Research*. Thousand Oaks: Pine Forge Press.

Ritzer, G. (1993). *Teoría sociológica contemporánea*. Madrid: Mac Graw-Hill.

Subirats, M., Sánchez, C. y Dominguez, M. (1992). *Grups i classes socials a la Regió Metropolitana de Barcelona*. Enquesta de la Regió Metropolitana de Barcelona 1990. Volum 5. Barcelona: Institut d'Estudis Metropolitans de Barcelona.

Veblen, T. (1944). *Teoría de la clase ociosa*. México: Fondo de Cultura Económica.

ANEXO


Primer eje correspondencias Consumo
2
1,5
1
0,5
0
-0,5
-1
-1,5
Piso lujo
Duraderos equipado superior
Motorizado lujo
Unifamiliar lujo
Ocio rico
Piso equipado y Bien motoriz. standar
Vacaciones Movilidad y CulturaNivel alto
Modelo transición
Coche Familiar nuevo, Modelo moderno y Móviles todo fuera
Duraderos Standard
Moviles dentro ciudad
Gastos Sobrantes y Satisfechos
Vacaciones Standard, Movil libres, Coche utilitario y Ocio Standard
CulturaNivel regular y ViviendaUnifamiliaraserie
Movil Standard/estático
Vacaciones no sde
Vivienda básica/precaria
Gastos Básicos , Ocio pobre, Modelo Tradicional
No vacaciones
Duraderos básicos
CulturaNivel bajo
No coche
Gastos Precarios
CulturaNivel muy bajo
62% Varianza Explicada

ENGLISH SUMMARY

THE ARTICULATION BETWEEN THE QUANTITATIVE AND THE QUALITATIVE: FROM THE GREAT SURVEYS TO THE INTENSIVE DATA COLLECTION

V. BORRÀS
P. LÓPEZ
C. LOZARES
Universitat Autònoma de Barcelona*

Most times, the qualitative works have been the beginnings of research, or even more, they have shaped the exploratory stages before the dimensionalisation and categorisation for the use of extensive techniques as surveys.

This article is an example of an inverse planning which starts from a large survey in order to structure the reality which is the target of the research and afterwards to apply qualitative techniques, concretely discussion groups. As a result, it is concerned with increasing the validity and representativeness of the discourses by individuals and groups which have been gained from the features obtained in the structural stage.

Keywords: Data analysis, factor analysis and principal components, cluster analysis, methodology, consumption, quantitative and qualitative techniques

AMS Classification: 62D05, 62-07, 62H25, 62H30

* Vicent Borràs, Pedro López Roldán i Carlos Lozares Colina, Departament de Sociologia, Universitat Autònoma de Barcelona. Los tres autores son miembros del *Grup d'Estudis Sociològics sobre la Vida Quotidiana i el Treball* (QUIT) del Departament de Sociologia de la Universitat Autònoma de Barcelona.

- Received August 1999.
- Accepted November 1999.

1. INTRODUCTION

The purpose is to find coincidental points which make easier the complementariness between the qualitative perspective and the quantitative perspective. This paper presents a specific example in which both techniques complement, and at the same time, it attempts to overcome the false dichotomy established.

2. THE NECESSITY OF A COMPLEMENTARINESS

One of the aims of social sciences is to codify the personal and unspoken knowledge. But the unspoken and experimental knowledge must become explicit, public and formalised so that it can be transmitted. Therefore, we would say that the knowledge must be explicable and comprehensive. Because of that, it is necessary to use a double approach based on both methodological perspectives.

3. THE APPROACH OF THE PERSPECTIVES

The proposal we make here starts from the use of data and analyses of a quantitative kind in a first stage to go to obtaining data and analyses of a qualitative kind in a second stage. Then, the scheme established for the research is the following one:

Quantitative Analysis → To select Members → Discussion Groups → Qualitative Analysis

4. APPLICATION TO THE CONSUMPTION ANALYSIS

Here we propose to start from the analysis of quantitative data obtained from a large survey like the Enquesta de la Regió Metropolina de Barcelona 1990, Survey from the Metropolitan Area of Barcelona 1990, and cover the triple purpose:

A. Our study target was double in the first stage of the research, on the one hand, proving that the consumption behaviours have an internal structure and articulation, and on the other hand that consumption is a reflection of class position.
B. To reach a second level in order to see which the imagery of class is, according to consumption, as well as which the representations and symbologies of class are, regarding the things consumed.
C. Discussion groups have been formed for this stage. So that they were representative and therefore, able to overcome the problems of external validity which this technique creates, we have used the results obtained in the previous stage.

QÜESTIIÓ, vol. 23, 3, p. 543-558, 1999

DISEÑO DE MUESTRAS EN ENCUESTAS DE POBLACIÓN Y HOGARES

J. PORRAS PUGA
Instituto Nacional de Estadística*

En este artículo se presenta el esquema general del diseño muestral que el Instituto Nacional de Estadística (INE) utiliza en las encuestas dirigidas a la población. Este diseño se conoce como Encuesta General de Población y fue implantado en el año 1971, aunque desde entonces ha sufrido ligeros cambios en cuanto a la periodicidad, criterios de estratificación, tamaño, etc. Se utiliza tanto para las encuestas de tipo continuo (encuestas coyunturales) como de tipo esporádico (encuestas estructurales).

Sampling design for Household Surveys

Palabras clave: Marco, estratificación, probabilidad proporcional, estimador de razón

Clasificación AMS: 62D05

* Instituto Nacional de Estadística (INE). Diseño de Muestras de Población y Hogares. Paseo de la Castellana, 183. 28071 Madrid. E-mail: juporras@ine.es

– Recibido en setiembre de 1999.
– Aceptado en noviembre de 1999.

1. INTRODUCCIÓN

La Encuesta de Población Activa (EPA) es una encuesta de tipo continuo y periodicidad trimestral cuyo objetivo es el conocimiento de las características de la población en relación con la actividad económica.

El INE viene realizando esta encuesta ininterrumpidamente desde 1964, habiendo sufrido desde entonces diversas modificaciones que han afectado tanto al cuestionario como al diseño de la muestra.

Desde 1971 el diseño de la EPA se enmarca en el de la Encuesta General de Población (EGP).

La EGP no es una encuesta en sí misma sino un diseño muestral válido para encuestas dirigidas a la población. El objetivo de la misma es mantener un diseño muestral actualizado que permita, en un momento dado, investigar las características de la población española en todos aquellos aspectos que interesen a la Administración del Estado.

El ámbito poblacional en la EGP es la población que reside en viviendas familiares principales. Estas son las utilizadas toda o la mayor parte del año como residencia habitual o permanente. Se excluye del ámbito poblacional la población residente en hogares colectivos.

El estudio básico que el INE realiza con este diseño, como se mencionó anteriormente, es la EPA. Otras encuestas realizadas en el marco de la EGP son, entre otras, las Encuestas de Presupuestos Familiares (1980-81 y 1990-91), la Encuesta de Fecundidad 1998 y la Encuesta sobre Discapacidades y Deficiencias (1986 y 1999).

En los apartados siguientes se presentan los aspectos más importantes del diseño de la encuesta.

2. MARCO

Se considera marco de una encuesta al conjunto de información que puede ser útil en cualquier momento del diseño de la misma.

En sentido restringido el marco está formado por la relación de unidades de muestreo. Las unidades de muestreo deben estar definidas de forma tal que su identificación sea inequívoca, no exista solapamiento entre ellas, a cada unidad se le pueda asignar una probabilidad de selección y que el conjunto de todas ellas coincida con la población que se pretende estudiar.

En sentido amplio, el marco está constituido además por toda la información complementaria, mapas, listas, comunicaciones, etc., que nos permita llegar a un mayor conocimiento de la población y que se pueda utilizar para la división de la población en estratos, selección de la muestra o formación de estimadores.

De acuerdo con el ámbito poblacional de la encuesta, para la selección de la muestra sería necesario disponer de un listado de todas las viviendas familiares existentes en el territorio nacional. Esto es difícil de conseguir ya que, por una parte, esta relación viene afectada por el paso del tiempo y, por otra, de errores tales como omisiones, duplicidades, etc.

Como consecuencia de lo anterior, se elige en lugar de viviendas áreas geográficas que son más estables en el tiempo y se muestrean dichas áreas.

Para definir el marco de la Encuesta es necesario partir de la división administrativa de España. Todo el Estado se encuentra dividido en 17 **Comunidades Autónomas** más Ceuta y Melilla y a su vez en 50 **provincias** de las cuales 47 son peninsulares y 3 insulares. Las provincias se encuentran divididas en **municipios** y éstos en distritos municipales. Hasta aquí tenemos la división administrativa oficial. Después el INE, juntamente con los Ayuntamientos, hace una nueva subdivisión de los distritos en **secciones censales**.

Las secciones se utilizan para todos los trabajos encomendados al INE en los que es necesario una división inframunicipal, entre otros para fines electorales como *secciones electorales*, lo cual exige de acuerdo con la Ley Electoral que cada sección incluya un máximo de 2.000 electores y un mínimo de 500. Por tanto, la sección censal puede considerarse como un área geográfica con límites perfectamente definidos, cuyo tamaño de población viene limitado por las condiciones antes expuestas.

El seccionado y su número varía considerablemente a lo largo del tiempo, por lo que con referencia 1 de enero de cada año y en cada Censo o Padrón se realiza una actualización del mismo. Por una parte, hay secciones que quedan despobladas y es necesario fusionarlas con otras y, por otra, también se produce el fenómeno contrario, es decir, las secciones crecen hasta superar los límites de población establecidos y es necesario dividirlas. Finalmente, para llegar a la vivienda familiar es posible confeccionar para cada sección censal una lista de viviendas familiares con sus direcciones postales, obtenidas del último Censo o Padrón.

Por tanto, el marco de la encuesta lo constituye:

- El marco de áreas, formado por las aproximadamente 32.000 secciones censales en que se encuentra dividido el territorio nacional.
- El marco de viviendas formado por la relación de viviendas familiares que se confecciona para cada sección censal seleccionada para la muestra.

3. UNIDADES DE MUESTREO Y CRITERIOS DE ESTRATIFICACIÓN

El tipo de muestreo utilizado es un **muestreo bietápico de conglomerados con submuestreo y estratificación de las unidades de primera etapa**. En cada provincia se diseña una muestra independiente.

Las unidades de primera etapa son las **secciones censales**. Con el objetivo de conseguir estimaciones fiables del cambio entre dos períodos de encuesta, la muestra de secciones permanece fija indefinidamente, con las excepciones siguientes:

a) Cuando los resultados obtenidos en los Censos arrojen variaciones sensibles en la estructura de la población que aconsejen una afijación distinta.

b) Se agoten los hogares consultables de la sección.

c) Cuando al actualizar las probabilidades de selección le corresponda salir de la muestra.

Estas unidades se estratifican atendiendo a un doble criterio:

Criterio geográfico (de estratificación)

Las secciones en cada provincia se agrupan según la importancia demográfica del municipio a que pertenecen.

Criterio socioeconómico (de subestratificación)

Dentro de cada estrato geográfico las secciones censales se agrupan en **subestratos**, atendiendo a la categoría socioeconómica de los hogares ubicados en la sección.

Para llegar a la formación de los estratos se consideran los siguientes tipos de municipios:

1. Municipios autorrepresentados: Son aquellos que dada su categoría dentro de la provincia deben tener siempre secciones en la muestra.

Son municipios autorrepresentados:

- La capital de la provincia.
- Municipios que tienen un número de habitantes tal que en la afijación proporcional dentro de la provincia le corresponden al menos 12 secciones en la muestra.
- Municipios que teniendo una situación demográfica destacada dentro de la provincia no hay otros similares con que agruparlos, aunque proporcionalmente le correspondan menos de 12 secciones en la muestra.

2. Municipios correpresentados: Son aquellos que dentro de la misma provincia forman parte de un grupo de municipios demográficamente similares y que son representados en común. De acuerdo con esta clasificación, en líneas generales, los estratos teóricos considerados responden a los siguientes conceptos:

Estrato 1: Municipio capital de provincia.

Estrato 2: Municipios autorrepresentados, importantes en relación con la capital.

Estrato 3: Otros municipios autorrepresentados, importantes en relación con la capital o municipios mayores de 100.000 habitantes.

Estrato 4: Municipios entre 50.000 y 100.000 habitantes.

Estrato 5: Municipios entre 20.000 y 50.000 habitantes.

Estrato 6: Municipios entre 10.000 y 20.000 habitantes.

Estrato 7: Municipios entre 5.000 y 10.000 habitantes.

Estrato 8: Municipios entre 2.000 y 5.000 habitantes.

Estrato 9: Municipios menores de 2.000 habitantes.

Hay que tener en cuenta que dada la diferente distribución de tamaños de los municipios entre las distintas provincias, no se ha podido realizar una estratificación uniforme para todas ellas. Por ejemplo, en la provincia de Lugo solamente hay 10 municipios con menos de 2.000 habitantes, por lo que se han agrupado los estratos teóricos 8 y 9 en el estrato 8 que contiene a los municipios de menos de 5.000 habitantes. Por el contrario, la provincia de Burgos tiene más de 350 municipios de menos de 2.000 habitantes incluidos en el estrato 9 y, sin embargo, tiene agrupados los estratos teóricos 7 y 8 en el estrato 7 al no haber apenas municipios entre 2.000 y 5.000 habitantes. No obstante, siempre que ha sido posible, se ha procurado realizar una estratificación uniforme para todas las provincias pertenecientes a una misma Comunidad Autónoma.

Para la formación de los **subestratos** se tiene en cuenta la categoría socioeconómica de los hogares ubicados en la sección. Las secciones cambian de subestrato debido a la variación de la estructura de la población, por lo que la subestratificación se revisa en cada Censo, utilizando la información que éste proporciona sobre las características que intervienen en la definición de categoría socioeconómica.

Esta información permite clasificar la población económicamente activa de la sección en 18 categorías que a su vez se agrupan en cuatro grupos homogéneos. El primero agrupa a la población cuya actividad principal es la agricultura; el segundo al conjunto de trabajadores por cuenta propia. El tercer grupo representa al conjunto de directivos

y profesionales por cuenta ajena y al personal administrativo y el cuarto grupo al resto de los trabajadores.

Existen dieciséis subestratos, quince de los cuales se obtienen en función de los porcentajes de población de los grupos anteriores y el decimosexto (subestrato cero) está formado por aquellas secciones con un elevado porcentaje de población inactiva.

La definición de los quince primeros subestratos se establece según:

1) Haya un claro predominio de uno de los cuatro grupos sobre los otros tres; 2) predominen dos sobre los otros dos; 3) predominen tres grupos, y 4) no hay un claro predominio de ninguno de los cuatro grupos. En alguno de los estratos pueden no existir varios de los subestratos.

El criterio matemático para considerar que un grupo es de predominio se establece según que el grupo considerado sea superior a los dos tercios del grupo predominante. Así, por ejemplo, supongamos que los porcentajes de los grupos de población económicamente activa en una sección son: Grupo 1=20, Grupo 2=40, Grupo 3=30 y Grupo 4=10. El grupo más importante es el 2. Se verifica además que: porcentaje grupo 3 $>$ 2/3 porcentaje grupo 2, lo que no sucede con el resto de los grupos. Por tanto el subestrato al que pertenece la sección será el 23, es decir, predominan estos dos sobre el resto.

Las **unidades de segunda etapa** están constituidas por las viviendas familiares principales (ocupadas permanentemente) y los alojamientos fijos (chabolas, cuevas, etc.). No se consideran las viviendas secundarias (ocupadas sólo una parte del año) y las disponibles para alquiler o venta, ya que no forman parte del ámbito poblacional definido anteriormente.

Dentro de las unidades de segunda etapa no se realiza submuestreo alguno, recogiéndose información de todas las personas que tengan su residencia habitual en las mismas.

4. TAMAÑO Y AFIJACIÓN DE LA MUESTRA

Para la determinación del tamaño de muestra se partió de una función de coste de tipo lineal y de la expresión del coeficiente de variación para una proporción en el muestreo de conglomerados con submuestreo.

Se empleó la siguiente función de coste:

$$Q = n\,Q_S + n\,m\,Q_V \quad \text{con} \quad Q_S = Q_F + d\,Q_D$$

donde:

Q = Presupuesto total para el pago a los entrevistadores
Q_S = Coste por unidad primaria (sección)
Q_V = Coste por unidad última (vivienda)
n = Número de secciones
m = Número de viviendas por sección
Q_F = Coste fijo por sección
Q_D = Coste diario del trabajo de campo
d = Número de días necesarios para el trabajo de campo

Todas las variables eran conocidas excepto n y m.

El coeficiente de variación para una proporción viene dado por:

$$C^2(\hat{P}) = \frac{V(\hat{P})}{\hat{P}^2} = \frac{1-\hat{P}}{\hat{P}} \cdot \frac{1+\delta(m-1)}{nm} = \frac{1-\hat{P}}{\hat{P}} F(\delta, m, n)$$

siendo:

$$F(\delta, m, n) = \frac{1+\delta(m-1)}{nm}$$

y δ el coeficiente de correlación intraclásica, que para el caso de la población activa se ha calculado y vale 0,05.

El mínimo de la expresión $C^2(\hat{P})$ respecto de las variables m y n se obtiene calculando el mínimo de la expresión $F(\delta, m, n)$ que es independiente de $\hat{P}$.

Para distintos valores de m compatibles con el trabajo de campo,

$$m = 4, 6, 8, 10, 11, 14, 17, 18, 19, \ldots, 91, 100$$

y los correspondientes valores de n dados por:

$$n = \frac{Q}{Q_S + m\,Q_V}$$

se obtienen distintos valores para $F(\delta, m, n)$.

El valor mínimo de $F(\delta, m, n)$ respecto de m y n correspondió a $m = 20$ y $n = 3.000$. En base a este resultado la muestra se fijó en un total de 3.060 secciones.

Posteriormente, con objeto de lograr una mayor representatividad en algunas Comunidades Autónomas y al mismo tiempo dar cumplimiento a las exigencias de la Unión Europea en cuanto al tamaño de la muestra en las Encuestas de Empleo, se ha ampliado la muestra en diversas ocasiones hasta alcanzar el tamaño actual de 3.484 secciones.

Para la afijación entre las provincias se tuvieron en cuenta los siguientes aspectos:

a) Disponer en cada provincia de un tamaño mínimo de muestra que permita dar estimaciones de la misma.

b) Los resultados nacionales deben tener la mayor fiabilidad posible.

Para compatibilizar estas condiciones se ha aceptado una **afijación de compromiso entre la uniforme y la proporcional**, a base de agrupar provincias de importancia demográfica similar y asignarles de 36 a 144 secciones (actualmente estos límites son de 39 a 156 secciones).

Dentro de cada provincia la afijación entre estratos es proporcional al tamaño de cada uno de ellos, si bien se han potenciado los estratos donde se encuentran los municipios de mayor tamaño, ya que se espera que la mayor parte de las características que se estudian estén correlacionadas con los niveles económico-social y cultural de los habitantes y es precisamente en estos estratos donde, en general, la dispersión debe ser mayor y donde el costo por entrevista es menor.

Dentro de los estratos, la afijación entre subestratos es estrictamente proporcional al tamaño (medido en número de viviendas familiares).

El tamaño de muestra final de unidades de segunda etapa depende de los objetivos de la encuesta, variando desde 16.000 viviendas en la Encuesta de Fecundidad a 75.000 investigadas en la Encuesta de Discapacidades.

5. SELECCIÓN DE LA MUESTRA

La selección de la muestra se realiza de tal forma que dentro de cada estrato cualquier vivienda familiar tenga la misma probabilidad de ser seleccionada, es decir, se tengan **muestras autoponderadas dentro de cada estrato**.

Para ello, las unidades de primera etapa (secciones censales) se seleccionan con probabilidad proporcional al número de viviendas familiares principales, según los datos del último Censo o Padrón. Dentro de cada sección seleccionada en primera etapa, se selecciona un número fijo de viviendas familiares, m, con igual probabilidad mediante la aplicación de un muestreo sistemático con arranque aleatorio.

Por tanto, la probabilidad de selección de la vivienda i, perteneciente a la sección j del estrato h, donde se han afijado K_h secciones sería

Siendo

$$P(v_{ijh}) = P(S_{jh}) \cdot P(v_{ijh}/S_{jh}) = K_h \cdot \frac{V_{jh}}{V_h} \cdot \frac{m}{V_{jh}} = K_h \cdot \frac{m}{V_h}$$

$P(S_{jh})$ = Probabilidad de selección de la sección j del estrato h
$P(v_{ijh}/S_{jh})$ = Probabilidad de selección de la vivienda i condicionada a la selección de la sección j.
v_{jh} = Total de viviendas de la sección j.
V_h = Total de viviendas del estrato h.
m = Número fijo de viviendas investigado en cada sección.

Como se ve, esta probabilidad no depende de i ni de j, es decir, la probabilidad de selección de una vivienda no depende de la sección a la que pertenece, sino solamente del estrato.

6. ESTIMADORES

Se utilizan **estimadores de razón** separados tomando como variable auxiliar las Proyecciones Demográficas de población elaboradas por el INE.

La expresión del estimador de una determinada característica X es la siguiente:

$$\hat{X} = \sum_h \sum_{i=1}^{n_h} \frac{P_h}{p_h} x_{hi}$$

extendiéndose el sumatorio en h a los estratos de una provincia, una comunidad autónoma o al total nacional, y donde:

P_h = Proyección de la población que reside en viviendas familiares, en el estrato h.
p_h = Número de personas que habitan en las viviendas de la muestra, en el estrato h, en el momento de la entrevista.
n_h = Número de viviendas en las secciones de la muestra en el estrato h.
X_{hi} = Valor de la característica investigada en la vivienda i-ésima, del estrato h.

A la expresión de este estimador se llega de la siguiente manera:

Al ser la muestra autoponderada en cada estrato, un estimador insesgado de la característica X se puede obtener mediante un estimador de expansión simple, que tiene la expresión:

$$\hat{X} = \sum_h \frac{V_h}{v_h} x_h$$

donde:

V_h = Total de viviendas en el estrato h.
v_h = Viviendas en la muestra en el estrato h.
x_h = Valor muestral de la característica investigada en el estrato h.

Este estimador tiene el inconveniente de que el valor de V_h sólo es conocido en el momento del Censo, por lo que sería necesario estimarlo para los períodos intercensales. Para evitar esto se recurre a un estimador de razón, utilizando como variable auxiliar las proyecciones demográficas de población estimadas por el INE para cada período de encuesta.

El estimador separado de razón tiene la expresión:

$$\hat{X}_r = \sum_h \frac{\hat{X}_h}{\hat{P}_h} P_h$$

siendo:

$\hat{P}_h$ = Población estimada en el estrato h a partir de la muestra

$$\hat{P}_h = \frac{V_h}{v_h} p_h$$

p_h = Población en las viviendas de la muestra en el estrato h.

Sustituyendo $\hat{X}_h$ y $\hat{P}_h$ en la expresión de $\hat{X}_r$ se obtiene:

$$\hat{X}_r = \sum_h \frac{\frac{V_h}{v_h} x_h}{\frac{V_h}{v_h} p_h} P_h = \sum_h \frac{P_h}{p_h} x_h$$

que corresponde a la expresión inicial del estimador.

7. ACTUALIZACIONES EN EL MARCO DE LA ENCUESTA

Las continuas variaciones de población, bien en sus características, bien en su distribución espacial, exigen realizar actualizaciones en el marco que necesariamente repercuten en la estructura muestral.

En el marco de la EGP se consideran tres tipos de actualizaciones:

Actualización en el marco de viviendas con carácter restringido y exclusivo para las secciones de la muestra. Esta actualización, tiene por objeto incorporar las viviendas principales, *altas* de la sección, en el listado de viviendas de la misma.

Actualización en el marco de secciones, consecuencia de las modificaciones producidas por diversas incidencias como particiones, fusiones o variaciones de límites en las secciones seleccionadas. En cada uno de estos casos es necesario determinar la

probabilidad de selección de las nuevas secciones, así como el número de entrevistas a realizar en las mismas.

Actualización con carácter general relativa a todas las secciones y viviendas de la población, la cual se realiza cada cinco años coincidiendo con el *Censo de Población* o el *Padrón Municipal de Habitantes*.

7.1. Actualización en el marco de viviendas

Cada trimestre se actualiza el marco de viviendas en una sexta parte de las secciones seleccionadas para la muestra, con objeto de poder incorporar al marco aquellas viviendas, tanto de nueva construcción como las que se han transformado en viviendas familiares, las cuales no existían como tales cuando se realizó el Censo o Padrón. Estas viviendas se incorporan a la muestra con una probabilidad igual a la original de las viviendas de la sección.

Cada seis trimestres se produce, por tanto, una actualización completa del marco de viviendas en las secciones seleccionadas para la muestra.

Como consecuencia de esta actualización y con objeto de mantener la muestra autoponderada, el número de viviendas a seleccionar en cada sección varía según la expresión:

$$m' = m \cdot \frac{V_s'}{V_s}$$

De esta manera la muestra es autoponderada.

$$P(v_{ijh}) = P(S_{jh}) \cdot P(v_{ijh}/S_{jh}) = K_h \cdot \frac{V_s}{V_h} \cdot \frac{m\dfrac{V_s'}{V_s}}{V_s'} = K_h \cdot \frac{m}{V_h}$$

7.2. Actualización en el marco de secciones

Se consideran los siguientes casos:

1º) Partición de secciones

Es el caso de una sección S en la que el crecimiento del número de viviendas principales exige que se escinda en diversas partes $S_1, S_2, \ldots, S_K$, bien para formar nuevas secciones o para incorporarse a otras ya existentes.

Se plantea el problema de determinar las probabilidades de selección de las nuevas secciones para conocer cual es la que va a permanecer en la muestra, así como el número de viviendas a entrevistar en la misma para que la muestra sea autoponderada.

Se distinguen dos casos:

a) La sección S se fragmenta para formar dos o más secciones completas. En este caso se opera como sigue

Llamamos:

V_S = Número de viviendas de la sección S según el último Censo.
V'_S = Número de viviendas de la sección S después de actualizada.
V_{Sj} = Número de viviendas de la parte j de la sección S según datos del último Censo.
V'_{Sj} = Número de viviendas de la parte j de la sección S después de actualizada.

Se selecciona una de las nuevas secciones S_j con probabilidad proporcional a su tamaño actualizado V'_{Sj}/V_{Sj}.

El número de viviendas que deben ser objeto de entrevista es:

$$m' = m \cdot \frac{V'_S}{V_S}$$

las cuales son seleccionadas sistemáticamente.

De esta manera la muestra es autoponderada.

b) La sección S se fragmenta para anexionarse a una o más secciones existentes.

En este caso:

Se selecciona uno de los fragmentos con probabilidad proporcional a su tamaño según el último Censo, V_{Sj}/V_S, y la nueva sección S'_j a donde se haya incorporado dicha parte quedará automáticamente seleccionada.

El número de viviendas que han de ser entrevistadas viene dado por:

$$m' = m \cdot \frac{V'_{Sj}}{V_{Sj}}$$

siendo

$V'_{S'j}$ = Número de viviendas principales en la actualidad en la nueva sección S'_j.
$V_{S'j}$ = Número de viviendas principales que existían en el último Censo o Padrón dentro de los límites de la nueva sección S'_j.

2º) Fusión de secciones

Debido a que algunas secciones por los movimientos migratorios y naturales de la población van quedando vacías se procede a su fusión con otra u otras, de forma que en caso de ser seleccionada tengan unidades que investigar.

Si la sección S_j seleccionada para la muestra se fusiona con otra para formar la nueva sección S, ésta queda incorporada automáticamente a la muestra y el número de viviendas a entrevistar es:

$$m' = m \cdot \frac{V_S'}{V_S}$$

siendo

V_S' = Número de viviendas principales en la actualidad en la nueva sección S.
V_S = Número de viviendas principales, según último Censo o Padrón, dentro de los límites de la nueva sección S.

3º) Variación de límites

Éste es el caso de una sección que se forma con fragmentos de dos o más secciones por reajuste en sus límites.

Para el cálculo de la probabilidad de selección, este caso puede considerarse como un proceso en dos etapas: la primera de partición de cada sección y la segunda de fusión adecuada de las secciones resultantes de la partición.

7.3. Actualización de carácter general

Al obtenerse los resultados de un nuevo Censo o Padrón se procede a actualizar las probabilidades de selección de las secciones y a ajustar el número de entrevistas por sección.

Este procedimiento se realiza de tal forma que las probabilidades de selección de las secciones sean proporcionales al número de viviendas que en ese momento tenga cada una. En principio esto podría lograrse partiendo de cero y seleccionando una muestra nueva, pero ello provocaría una ruptura total con la muestra antigua, lo cual es arriesgado en el caso de encuestas continuas como es la EPA. Por ello se arbitra un procedimiento que sin distorsionar las probabilidades de selección que realmente corresponden a cada sección mantenga la muestra con las mínimas variaciones.

El procedimiento que se sigue es el siguiente:

Sea S una sección perteneciente al estrato h, seleccionada en un Censo o Padrón, C, con probabilidad

$$P_S = \frac{V_S^C}{V_h^C} = \frac{\text{Viviendas en } S \text{ según Censo } C}{\text{Viviendas en el estrato } h \text{ según Censo } C}$$

y supongamos que en el siguiente Censo o Padrón, C', le corresponde una probabilidad de selección dada por

$$P_{s'} = \frac{V_s^{C'}}{V_h^{C'}} = \frac{\text{Viviendas en } S \text{ según Censo } C'}{\text{Viviendas en el estrato } h \text{ según Censo } C'}$$

Se compara P_s con P_s' pudiendo ocurrir uno de los dos siguientes casos:

1) Si $P_s' > P_s$ la sección S permanece en la muestra con probabilidad P_s', ya que si fue seleccionada con una probabilidad P_s inferior a la que actualmente le corresponde, con mayor motivo hubiera salido seleccionada aplicándole su probabilidad actual P_s'.

2) Si $P_s' < P_s$ la sección permanece en la muestra con probabilidad P_s'/P_s y sale de la muestra con probabilidad $1 - P_s'/P_s$.

Este criterio motivará la salida de la muestra de un cierto número de secciones. Estas serán sustituidas por otras secciones del mismo estrato pero seleccionadas de **entre las que no perteneciendo a la muestra hayan aumentado de probabilidad**.

Con este criterio se mantiene el esquema de que la probabilidad que tiene una sección de pertenecer a la muestra es la que realmente le corresponde, es decir, proporcional al número de viviendas actuales.

8. COMENTARIO FINAL

Con este tipo de diseño muestral y las correspondientes actualizaciones, el INE mantiene un marco actualizado sobre el que realiza todas las investigaciones de tipo social y económico dirigidas a la población.

El desarrollo de nuevas técnicas de estimación, así como la disponibilidad del futuro Padrón Continuo permitirá introducir mejoras en este diseño.

ENGLISH SUMMARY

SAMPLING DESIGN FOR HOUSEHOLD SURVEYS

J. PORRAS PUGA
Instituto Nacional de Estadística*

This paper present the general sampling design used by the Instituto Nacional de Estadística (INE) for household surveys. This design is named Encuesta General de Población *and it was implemented in 1971. Since then has had several methodological changes, related to periodicity, criteria of strata, etc. It is used in continuous and non continuous surveys.*

Keywords: Frame, stratification, proportional probability, ratio estimator

AMS Classification: 62D05

* Instituto Nacional de Estadística (INE). Diseño de Muestras de Población y Hogares. Paseo de la Castellana, 183. 28071 Madrid. E-mail: juporras@ine.es

- Received September 1999.
- Accepted November 1999.

1. INTRODUCTION

The Spanish Labour Force Survey is a continuous survey that has been conducted by the INE since 1964. Since 1971 this survey uses the general design of the **Encuesta General de Población** (EGP). The EGP is an updated sampling design used by the INE in Household Surveys, that are useful for the Government Policy. It includes the population living in private dwellings. The institutions (hotels, hospital,...) are excluded from the sample.

2. DESIGN OF THE E G P

A two-stage sampling is used with stratification of the first stage units. The first stage units are the enumeration areas. These are stratified, within each province, using the population size of the municipality. Within each strata, they are substratified according to the socio-economic characteristic of the population in the enumeration areas.

The second stage units are the private dwellings.

The sample size was calculated using the criterion of minimum variance for the estimator of a proportion with fixed budget and a lineal cost function.

The number of primary units was established in 3.060.

Nowadays the sample size is 3.484 enumeration areas.

The units are selected in such a way to obtain self -weighted samples within each stratum. The first stage units are selected with proportional probability to the size and second stage units are selected with equal probability.

The design use **Ratio Estimator**, and the auxiliary variable is the Population Projection.

The frame is updated in two ways:

- Updating of the dwellings in the sampling primary units. Each of these units is updated every six quarter.
- General updating of all the primary units with the information obtained from the Census or from the Population Register.

Biometria

QÜESTIIÓ, vol. 23, 3, p. 561-571, 1999

GENERALIZATION OF THE KAPPA COEFFICIENT FOR ORDINAL CATEGORICAL DATA, MULTIPLE OBSERVERS AND INCOMPLETE DESIGNS

V. ABRAIRA*
A. PÉREZ DE VARGAS**
Hospital Ramón y Cajal

This paper presents a generalization of the kappa coefficient for multiple observers and incomplete designs. This generalization involves ordinal categorical data and includes weights which permit pondering the severity of disagreement. A generalization for incomplete designs of the kappa coefficient based on explicit definitions of agreement is also proposed. Both generalizations are illustrated with data from a medical diagnosis pilot study.

Keywords: Agreement, kappa, incomplete designs

AMS Classification: 62P10, Q2B15

* Unit of Clinical Biostatistics. Hospital 'Ramón y Cajal'. Crta. Colmenar km 9,1, planta -2 D. 28034 Madrid (Spain). Tel. +34 91 3368103. Fax. +34 91 3369016. E-mail: victor.abraira@hrc.es.

** Department of Biomathematics. University Complutense of Madrid. Facultad de Biología. Ciudad Universitaria. 28040 Madrid (Spain). Tel. +34 91 3945078. Fax. +34 91 3945051.
E-mail: alpedeva@eucmax.sim.ucm.es.
Address for correspondence: Unidad de Bioestadística Clínica. Hospital 'Ramón y Cajal'. Crta. Colmenar km 9,1, planta -2 D. 28034 Madrid (Spain).

- Received May 1998.
- Accepted July 1999.

1. INTRODUCTION

An important feature of any measurement or classification device is the reproducibility or reliability, which in classification is also referred to as concordance or agreement. From the seminal paper by Cohen [1], introducing the kappa coefficient (κ) to assess concordance between two observers using binary classifications, a great effort has been made to extend this index to more general conditions. Thus, Cohen [2] generalized kappa to weighted kappa in order to encompass ordinal variables incorporating an a priori assignment of weights to each of the cells of the $k \times k$ table of joint nominal scale; Landis and Koch [3] proposed an approach by expressing the quantities which reflect the extent to which the observers agree among themselves as functions of observed proportions obtained from underlying multidimensional contingency tables, using the GSK method [4]; Davies and Fleiss [5] proposed a generalization for multiple observers by the average of pairwise agreement. Although some limitations of kappa index are known such as that its value depends on the balance and symmetry of marginal totals of the table [6, 7] and some alternative methods of evaluating agreement among observers have been proposed [8, 9, 10, 11], the kappa index is still a very frequently used statistic in clinical epidemiology literature (e.g. Elmore *et al.* [12], Jelles *et al.* [13], Pérez *et al.* [14].

This paper generalizes Schouten's [15] and Gross's [16] proposal for multiple observers and incomplete design, as to encompass ordinal variables with the inclusion of weights to enable pondering the severity of disagreement among different categories. Another generalization for incomplete designs is also proposed, based on the explicit definitions of agreement by Landis and Koch [17]. This generalization is approached in a simpler way than the very general method of Koch *et al.* [18].

Both generalizations were motivated by the study shown in section 5. We tried to assess the concordance among several physicians evaluating the current health status of people affected by the Toxic Oil Syndrome. In this study there were some ordinal multicategorical variables as «peripheral neuropathy» and «sclerodermiform changes of the skin», and in order to avoid seeing each patient too many times at short intervals times for the same sign, an incomplete design should be used.

2. GENERALIZATION OF κ INDEX

The κ index, proposed by Cohen, is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is the proportion of observed agreement and P_e the proportion of expected agreement in the hypothesis of independence between observers. When there are only two observers, the definition of agreement is obvious. However, when there are more than two observers, agreement can be defined in diverse ways [19]. In this paper, we will restrict ourselves to pairwise agreement [5] (section 2.1) and majority agreement [17] (section 2.2).

2.1. Pairwise agreement

A set of N subjects is classified in K ordinal categories by a set G of $J > 1$ observers, with an incomplete design, that is to say, each subject i is only classified by a subset G_i of $J_i \leqslant J$ observers. Let X_{ik} be the number of observers classifying the ith subject into the kth category and w_{lm} the weight corresponding to the agreement-disagreement between categories l and m, obviously with the conditions:

$$w_{mm} = 1; \quad 0 \leqslant w_{lm} < 1 \quad \forall\, l \neq m; \quad w_{lm} = w_{ml}$$

For the ith subject, the number of weighted agreements is:

$$NA_i = \frac{1}{2}\sum_{k=1}^{K} w_{kk} X_{ik}(X_{ik}-1) + \sum_{l=1}^{K}\sum_{m>l}^{K} w_{lm} X_{il} X_{im}$$

and as the number of possible pairs of classifications for each subject i is $\dfrac{J_i(J_i-1)}{2}$, the proportion of weighted agreements for the ith subject is:

$$(2) \quad \frac{\sum_{k=1}^{K} w_{kk} X_{ik}(X_{ik}-1) + 2\sum_{l=1}^{K}\sum_{m>1}^{K} w_{lm} X_{il} X_{im}}{J_i(J_i-1)} = \frac{\sum_{m=1}^{K}\sum_{l=1}^{K} w_{lm} X_{il} X_{im} - J_i}{J_i(J_i-1)}$$

because:

$$\sum_{k=1}^{K} w_{kk} X_{ik} = \sum_{k=1}^{K} X_{ik} = J_i$$

Then, the average proportion of observed agreements for all subjects is the sum of (2) for all subjects divided by the number of subject, so:

$$(3) \quad P_o = \frac{1}{N_c}\sum_{i=1}^{N_c} \frac{\sum_{m=1}^{K}\sum_{l=1}^{K} w_{lm} X_{il} X_{im} - J_i}{J_i(J_i-1)}$$

where N_c is the number of subjects classified by more than one observer.

Let $P_j(k)$ $(j = 1, \ldots, J; k = 1, \ldots, K)$ represent the proportion of times which the jth observer classifies into the kth category. Then, the proportion of expected agreements for the ith subject in the hypothesis of independence between the pair l and m of observers is:

$$\sum_{u=1}^{K} \sum_{k=1}^{K} w_{uk} P_l(u) P_m(k)$$

We note that with incomplete designs the expected agreement is different for each subject because each one is classified by a different subset of observers. Then, the average expected proportion of pairwise agreement in the hypothesis of independence for the ith subject is:

$$\frac{2}{J_i(J_i-1)} \sum_{l=1}^{J_i} \sum_{m>l}^{J_i} \sum_{u=l}^{K} \sum_{k=l}^{K} w_{uk} P_l(u) P_m(k)$$

where, obviously, the sums for m and l are restricted to set G_i of observers which have classified the ith subject. Then, the average expected proportion of pairwise agreement for all the subjects is:

$$P_e = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{2}{J_i(J_i-1)} \sum_{l=1}^{J_i} \sum_{m>l}^{J_i} \sum_{u=l}^{K} \sum_{k=l}^{K} w_{uk} P_l(u) P_m(k) \tag{4}$$

The κ index is calculated with (1), using (3) and (4). If weights are not included, that is to say $w_{mm} = 1$; $w_{lm} = 0$ $\forall l \neq m$, the expressions (3) and (4) are reduced to expressions given by Schouten [15] and Gross [16].

2.2. Majority agreement

In a study with multiple observers, agreement among observers can be also defined as majority or consensus: there is agreement at an observation if a majority of observers agree; e.g., if there are seven observers, it is possible define agreement when at least five of them agree. Obviously, it is advisable [17] to have a clear majority, e.g., 7-0, 6-1 splits, rather than «tie-breaking» majorities, e.g., 4-3 splits. It is possible to define the following indicator variables z_p, one for each agreement definition [17]:

$$z_{0i} = \begin{cases} 1 & \text{if all observers agree, for the subject } i \\ 0 & \text{otherwise} \end{cases}$$

$$z_{1i} = \begin{cases} 1 & \text{if at least } J-1 \text{ observers agree, for the subject } i \\ 0 & \text{otherwise} \end{cases}$$

$$z_{pi} = \begin{cases} 1 & \text{if at least } J-p \text{ observers agree, for the subject } i \\ 0 & \text{otherwise} \end{cases}$$

for calculating the proportion of observed agreement by means of them as follows:

$$P_{o(p)} = \frac{\sum_{i=1}^{N_c} z_{pi}}{N_c} \tag{5}$$

where N_c is the number of subjects observed whom it is able to observe the defined agreement; that is to say, the number of subjects observed by, at least, $J-p$ observers. In the hypothesis of independence, the proportion of expected agreement for each subject is:

$$\sum_{V \in V_{K,J_i,p}} P_1(V) \cdots P_{J_i}(V)$$

where $V_{K,J_i,p}$ represents the set of permutations with repetition of K elements taken J_i at a time, with at least $J_i - p$ of them remaining equals and $P_j(k)$ $(j = 1, \ldots, J; k = 1, \ldots, K)$, as in section 2.1, the proportion of times which the jth observer classifies into the kth category. The average proportion of expected agreement is its sum for all subjects divided by the number of subjects observed in whom it is possible to observe the defined agreement; that is to say

$$P_{e(p)} = \frac{1}{N_c} \sum_{i=1}^{N_c} \sum_{V \in V_{K,J_i,p}} P_1(V) \cdots P_{J_i}(V) \tag{6}$$

The κ index is calculated with (1), using (5) and (6). The complete design can be considered as a particular case in which $J_i = J \; \forall \; i$ and so $N_c = N$, in this case, (5) and (6) are reduced to formulas given by Landis and Koch [17].

3. INFERENCES ABOUT κ

The kappa statistic is an estimator of the κ parameter for subjects and observers population. In previous formulas, κ is implicitly defined as a function of the probabilities that

are estimated by the proportions P_o and $P_j(k)$. To make inferences about κ we need to compute its standard error. A very general method for this is the jackknife technique [20]. Parr and Tolley [21] have shown that for all real functions (such as kappa) of multinomial proportions, with continuous first and second partial derivatives, in large samples, the jackknife estimator approximately follows a normal distribution and its variance is estimated by the variance of pseudo-value.

Then, a confidence interval for κ is:

$$J(\kappa) - t_{\alpha/2,(N-1)}\frac{S_j}{\sqrt{N}} \leqslant \kappa \leqslant J(\kappa) + t_{\alpha/2,(N-1)}\frac{S_j}{\sqrt{N}}$$

where $J(\kappa)$ is the jackknife estimate and S_j the pseudo-value standard deviation.

4. SOFTWARE

A computer program was written in FORTRAN 77 and runs in PC's under DOS. The program calculates proposed kappa indexes, their jackknife estimates and their standard error, also estimated by same method. It is included in the statistical package PRESTA [22] (PRESTA is a statistical package in Spanish, available on the Internet URL http://www.hrc.es/bioest.html).

5. EXAMPLE

It is a pilot study previous to another study which was made to assess the current health status of people affected by the disease which came to be known as toxic oil syndrome (TOS). The TOS was developed in people who consumed adulterated rapeseed oil sold as cooking oil and it affected more than 20.000 people. A TOS description can be seen in Nadal and Tarkowski [23]. The study to assess the current health status [24], was conducted with all of the 4.015 affected registered in the seven TOS Follow-up Centers of Madrid. Clinical histories and patients' physical examinations were used as data sources. Physical examinations were made by nine different physicians from the Follow-up Centers. The pilot study, shown here, was conducted to assess reliability of the variables potentially most affected by observer subjectivity. The categorical variables included in the study were: peripheral neuropathy, classified in three levels: «no neuropathy», «doubtful neuropathy» and «certain neuropathy»; severity of sclerodermiform changes of the skin, classified in four levels: «no sclerodermia», «fair sclerodermia», «moderate sclerodermia» and «atrophic skin»; and joint contractures classified as «yes» and «no».

5.1. Study Design

Patients: A non random sample of 10 patients affected by TOS chosen to cover all range of clinical degrees of the disease.

Observers: A random sample of 6 physicians chosen from the nine whom later did the current health status study.

Procedure: Before the study, the six physicians participated in a 5-hour workshop, where they were trained in the protocol of variables collection. The workshop included a physical examination of several TOS patients, different from those who would later participate in the study. In order to avoid each patient's being seen too many times for the same sign at short time intervals, a balanced incomplete block design (Fleiss [25]) was selected. In this design, each patient is examined by 3 physicians, each physician examines 5 patients, and all possible pairs of physicians examine the same 2 patients. The examination designation scheme is laid out in Table 1. The efficiency factor [25] for estimating the coefficient of reliability of this design is 0.8, which seem like a reasonable compromise. The order of examinations in each patient was randomly determined using a permutations table. Patients were informed in writing of the purposes of the study and gave their written consent to participate in the study. In order to guarantee the patients' confidentiality no identification data was saved in computer files.

5.2. Results

Proportions of observed and expected agreement, kappa index, its jackknife estimate and its standard error for all variables are shown in tables 2, 3 and 4. Weighted kappa not was used in joint contractures variable because it has only two categories, squared error weights were used for the other variables [26]. The indexes found indicate a fair to moderate agreement according the benchmark of Landis and Koch [3], which obliged us to repeat observer training before conducting the current health status study. Although the sample size is small, big differences between sample estimation and jackknife estimation of kappa are not observed; which leads us to have confidence in the jackknife estimation of standard error. The differences between pairwise and weighted pairwise indexes, in tables 3 and 4, illustrate that the greatest disagreement occurs between contiguous categories; the differences between pairwise and majority kappa suggest that at least one observer classifies differently from the others. Marginal frequencies of peripheral neuropathy are shown in table 5, where it is seen that «physician 2» is clearly different, as he classified a proportion of 0.6 into the «doubtful» category and 0 in «certain». If analysis is repeated without this observer, all indexes increase considerably (0.7439 with SE=0.1727 for pairwise agreement; 0.8888 with SE=0.0832 for weighted pairwise agreement and 0.7379 with SE=0.2844 for majority agreement).

6. CONCLUSIONS

In the assessment of reliability among multiple observers, unbalanced designs often appear, either by design as in the presented example, or due to missing data. In this paper, we have proposed a simple modification of previous kappa indexes to include unbalanced designs in weighted kappa for ordinal variables and kappa for majority. We have also illustrated their use with real data and, in the example, we have shown how differences among several indexes (pairwise, weighted pairwise and majority) permit identification of the sources of disagreement, which is the main aim of this kind of studies.

Table 1. Balanced incomplete block design used in the TOS study

Patient	*Physi. 1*	*Physi. 2*	*Physi. 3*	*Physi. 4*	*Physi. 5*	*Physi. 6*
1	x			x		x
2			x	x	x	
3			x	x		x
4	x		x		x	
5	x				x	x
6	x	x	x			
7		x	x			x
8		x			x	x
9	x	x		x		
10		x		x	x	

Table 2. Joint contractures (2 categories)

Agreement	P_o	P_e	κ	$J(\kappa)$	$SE(\kappa)$
pairwise	0.6667	0.4827	0.3557	0.3827	0.2267
majority of 3	0.5000	0,2240	0.3557	0.3827	0.2267

P_o: proportion of observed agreement
P_e: proportion of expected agreement
κ: kappa index
$J(\kappa)$: jackknife estimate of kappa index
$SE(\kappa)$: jackknife standard error

Table 3. Peripheral neuropathy (3 categories)

Agreement	P_o	P_e	κ	$J(\kappa)$	$SE(\kappa)$
pairwise	0.6667	0.3387	0.4960	0.4995	0.1387
*pairwise we**	0.8667	0.6607	0.6071	0.6095	0.1738
majority of 3	0.5000	0.1176	0.4334	0.4373	0.1622

$*$ Weighted kappa with quadratic weights
P_o: proportion of observed agreement
P_e: proportion of expected agreement
κ: kappa index
$J(\kappa)$: jackknife estimate of kappa index
$SE(\kappa)$: jackknife standard error

Table 4. Sclerodermiform changes of the skin (4 categories)

Agreement	P_o	P_e	κ	$J(\kappa)$	$SE(\kappa)$
pairwise	0.6667	0.2507	0.5552	0.5757	0.1343
*pairwise we**	0.9407	0.6868	0.8108	0.8401	0.1062
majority of 3	0.5000	0.0656	0.4649	0.4825	0.1679

$*$ Weighted kappa with quadratic weights
P_o: proportion of observed agreement
P_e: proportion of expected agreement
κ: kappa index
$J(\kappa)$: jackknife estimate of kappa index
$SE(\kappa)$: jackknife standard error

Table 5. Marginal frequencies of peripheral neuropathy

	No	*Doubtful*	*Certain*
Physician 1	0.400	0.200	0.400
Physician 2	0.400	0.600	0.000
Physician 3	0.600	0.200	0.200
Physician 4	0.400	0.200	0.400
Physician 5	0.400	0.200	0.400
Physician 6	0.400	0.400	0.200
Mean	0.433	0.300	0.267

ACKNOWLEDGEMENT

This work was supported in part by FIS grant 96/0421. The authors would like to thank Kathleen Seley for her help in correcting this manuscript.

REFERENCES

[1] Cohen J. (1960). «A coefficient of agreement for nominal scales», *Educational and Psychological Measurement*, 20, 37-46.

[2] Cohen J. (1968). «Weighted kappa: Nominal scale agreement with provision for scaled disagreement or parcial credit», *Psychological Bulletin*, 70, 213-220.

[3] Landis J.R. and Koch G.G. (1977). «The measurement of observer agreement for categorical data», *Biometrics*, 33, 159-174.

[4] Grizzle J.E., Starmer C.F. and Koch G.G. (1969). «Analysis of categorical data by linear models», *Biometrics*, 25, 489-504.

[5] Davies M. and Fleiss J.L. (1982). «Measuring agreement for multinomial data», *Biometrics*, 38, 1047-1051.

[6] Feinstein A.R. and Cicchetti D.V. (1990). «High agreement but low kappa: I. The problems of two paradoxes», *Journal of Clinical Epidemiology*, 43, 543-549.

[7] Guggenmoos-Holzmann I. (1993). «How reliable are chance-corrected measures of agreement?», *Statistics in Medicine*, 12, 2191-2205.

[8] Cicchetti D.V. and Feinstein A.R. (1990). «High agreement but low kappa: II. Resolving the paradoxes», *Journal of Clinical Epidemiology*, 43, 551-558.

[9] Rosner B. (1982). «Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes», *Biometrics*, 38, 105-114.

[10] Donner A. and Donald A. (1988). «The statistical analysis of multiple binary measurements», *Journal of Clinical Epidemiology*, 41, 899-905.

[11] Graham P. and Jackson R. (1993). «The analysis of ordinal agreement data: beyond weighted kappa», *Journal of Clinical Epidemiology*, 46, 1055-1062.

[12] Elmore J.G., Wells C.K., Lee C.H., Howard D.H. and Feinstein A.R. (1994). «Variability in radiologist's interpretations of mammograms», *New England Journal of Medicine*, 331, 1493-1499.

[13] Jelles F., Van Bennekom C.A.M., Lankhorst G.F., Sibbel C.J.P. and Bouter L.M. (1995). «Inter- and intra-rater agreement of the rehabilitation activities profile», *Journal of Clinical Epidemiology*, 48, 407-416.

[14] Pérez B., Abraira V., Núñez M., Boixeda P., Pérez Corral F. and Ledo A. (1997). «Evaluation of agreement among dermatologists in the assessment of the color of

Port Wine Stains and their clearance after treatment with the Flaslamp-Pumped Dye Laser», *Dermatology*, 194, 127-130.

[15] Schouten H.J.A. (1986). «Nominal scale agreement among observers», *Psychometrika*, 51, 453-466.

[16] Gross S.T. (1986). «The kappa coefficient of agreement for multiple observers when the number of subjects is small», *Biometrics*, 42, 883-893.

[17] Landis J.R. and Koch G.G. (1977). «An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers», *Biometrics*, 33, 363-374.

[18] Koch G.G., Imrey P.B. and Reinfurt D.W. (1972). «Linear model analysis of categorical data with incomplete response vectors», *Biometrics*, 28, 663-692.

[19] Abraira V. (1997). *Precisión de las clasificaciones clínicas*. Doctoral Thesis. Universidad Computense de Madrid.

[20] Efron B. and Gong G. (1983). «A leisurely look at the bootstrap, the jackknife and cross-validation», *The American Statistician*, 37, 36-48.

[21] Parr W.C. and Tolley H.D. (1982). «Jackknifing in categorical data analysis», *The Australian Journal of Statistics*, 24, 67-79.

[22] Abraira V. and Zaplana J. (1984). «PRESTA, un paquete de procesamientos estadísticos», *Proceeding de la Conferencia Iberoamericana de Bioingeniería*. 100, Gijón.

[23] Nadal J. and Tarkowski S. (1992). «Toxic oil syndrome. Current knowledge and future perspectives. World Health Organization». *Regional Publications European Series*. Nº. 42, Copenhagen.

[24] Gómez de la Cámara A., Posada M., Abaitua I., Barainca M.T., Abraira V., Diez M. and Terracini B. (1998). «Health status measurements in Toxic Oil Syndrome», *Journal of Clinical Epidemiology*, 51, 867-873.

[25] Fleiss J.L. (1986). *The design and analysis of clinical experiments*. John Wiley & Sons, New York.

[26] Fleiss J.L. and Cohen J. (1973). «The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability», *Educational and Psychological Measurement*, 33, 613-619.

Secció Docent i Problemes

SECCIÓ DOCENT I PROBLEMES

La «Secció docent i problemes» té l'objectiu de publicar articles de caire docent, difícilment publicables en revistes de recerca. A cada número de *Qüestiió* s'inclouen d'un a tres problemes i les solucions es donen en el número següent.

Els lectors poden 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.

SOLUCIONS ALS PROBLEMES PROPOSATS AL VOLUM 23 N. 2

PROBLEMA N. 76

1. De $\Pi_n = \dfrac{\lambda_0\,\lambda_1\,\cdots\,\lambda_{n-1}}{\mu_1\,\mu_2\,\cdots\,\mu_n}\,\Pi_0$, al ser $\lambda_n = \lambda,\ \mu_n = \mu,\ \forall n\ \Rightarrow$

$$\Pi_n = \Psi^n\,\Pi_0,\ \psi = \frac{\lambda}{\mu}$$

$$1 = \sum_{n=0}^{\infty} \Pi_n = {}^{1}\frac{1}{1-\psi}\,\Pi_0\ \Rightarrow\ \Pi_0 = 1-\psi\ \Rightarrow\ \boxed{\Pi_n = \psi^n(1-\psi)\ \forall n}$$

2. La función generatriz de probabilidades resulta:

$$G(s) = \sum_{n_0}^{\infty} s^n\,\psi^n(1-\psi) = (1-\psi)\sum_{n=0}^{\infty}(s\,\psi)^n = (1-\psi)(1-s\,\psi)^{-1}$$

El número medio de clientes en el sistema será por tanto:

$$L = G'(s)\Big|_{s=1} = -\psi(1-\psi)(1-s\,\psi)^{-2}\Big|_{s=1} = \frac{\psi}{1-\psi}$$

El número medio de *estaciones desocupadas* será: $\emptyset = 1 \times \Pi_0 = 1-\psi$

$$L = L_q + s - \emptyset\ \Rightarrow$$

El número medio de clientes esperando: $L_q = \dfrac{\psi}{1-\psi} - 1 + (1-\psi) = \dfrac{\psi^2}{1-\psi}$

3. $P(X \leqslant n) = \displaystyle\sum_{i=0}^{n} \psi^n(1-\psi) = (1-\psi)\,\frac{1-\psi^{n+1}}{1-\psi} = 1-\psi^{n+1}$

La probabilidad de espera: $P(X > s=1) = 1-\Pi_0 = 1-(1-\psi) = \psi$.

El tiempo medio de espera: $T_q = \dfrac{L_q}{\lambda} = \dfrac{\psi}{\mu-\lambda}$

[1] La existencia de régimen permanente supone que $\psi < 1$.

4. Como $\psi < 1 \Rightarrow \Pi_0$ es la mayor probabilidad $\equiv$ $n = 0$ es el estado más probable.

5. $\Pi_n'(\psi) = -\psi^n + (1-\psi)n\,\psi^{n-1}$;
   $\Pi'(\psi) = 0 \Rightarrow \psi = (1-\psi)n \Rightarrow 1/n = 1/\psi - 1 \Rightarrow$

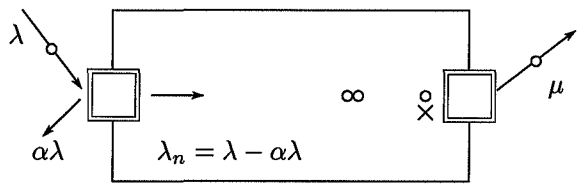
$$1/n + 1 = \frac{1+n}{n} = 1/\psi \Rightarrow \boxed{\psi = \frac{n}{n+1}} \qquad \begin{array}{l} n = 1,\ \psi = 1/2 = 3/6 \\ n = 2,\ \psi = 2/3 = 4/6 > 3/6 \end{array}$$

$$\max_{\psi} \Pi_n(\psi) = \left(\frac{n}{n+1}\right)^n \overbrace{\left(1 - \frac{n}{n+1}\right)}^{1/n+1} = \frac{n^n}{(n+1)^{n+1}} = \begin{cases} 1/4 & n = 1 \\ 4/9 & n = 2 \end{cases}$$

R. Alonso

PROBLEMA N. 77

a) Las tasas de *entrada* son ahora $\lambda_0 = \lambda$; $\lambda_n = \lambda(1-\alpha), n \geqslant 1$.



1. $\Pi_1 = \dfrac{\lambda}{\mu}\Pi_0 = \psi\,\Pi_0$, $\Pi_2 = \dfrac{\lambda\lambda(1-\alpha)}{\mu\mu}\Pi_0 = \psi^2(1-\alpha)\,\Pi_0, \ldots, \Pi_n =$
$= \psi(1-\alpha)\,\Pi_{n-1} = \psi^n(1-\alpha)^{n-1}\,\Pi_0$ **Nota**: $\psi(1-\alpha) < 1$.

$$1 = \Pi_0 + \sum_{n=1}^{\infty}\Pi_n = \Pi_0 + \frac{\psi\,\Pi_0}{1-\psi(1-\alpha)} = \frac{1+\alpha\,\psi}{1-\psi(1-\alpha)}\Pi_0 \Rightarrow$$

$$\Pi_0 = \frac{1-(1-\alpha)\psi}{1+\alpha\,\psi} \Rightarrow$$

$$\boxed{\Pi_n = \psi^n(1-\alpha)^{n-1}\,\Pi_0 = \psi^n(1-\alpha)^{n-1}\,\frac{1-(1-\alpha)\psi}{1+\alpha\,\psi}\;\forall\, n > 0} \underset{\alpha\,=\,0}{=} \psi^n(1-\psi)$$

2. $G(s) = \Pi_0 + \sum_{n=1}^{\infty} s^n\left(\psi^n(1-\alpha)^{n-1}\right)\Pi_0 = \Pi_0\left(1 + s\psi\sum_{n=1}^{\infty}\left(s\psi(1-\alpha)\right)^{n-1}\right) =$
$\Pi_0\left(1 + s\psi\dfrac{1}{1-s\psi(1-\alpha)}\right)$

$$G'(s) = \Pi_0\left(\psi\frac{1-s\psi(1-\alpha)+s(\psi(1-\alpha))}{(1-s\psi(1-\alpha))^2}\right) \Rightarrow$$

$$G'(1) = \frac{1(1-\alpha)\psi}{1+\alpha\psi}\left(\psi\frac{1}{(1-\psi(1-\alpha))^2}\right) \Rightarrow$$

$$L = \frac{\psi}{(1+\alpha\psi)(1-\psi(1-\alpha))} = \begin{cases} \dfrac{\psi}{1-\psi} & \alpha = 0 \\ \dfrac{\psi}{1+\psi} & \alpha = 1 \end{cases}$$

$$\emptyset = \Pi_0$$

$$L_q = L - s + \emptyset = L - (1 - \Pi_0) = \frac{\psi}{(1+\alpha\psi)(1-\psi(1-\alpha))} - \frac{\psi}{1+\alpha\psi} =$$

$$\frac{\psi}{1+\alpha\psi}\left(\frac{\psi(1-\alpha)}{(1-\psi(1-\alpha))}\right) \underset{\alpha=0}{=} \frac{\psi^2}{1-\psi}$$

3. $P(X > s = 1) = 1 - \Pi_0 = 1 - \dfrac{1-(1-\alpha)\psi}{1+\alpha\psi} = \dfrac{\psi}{1+\alpha\psi} \underset{\alpha=0}{=} \psi$

$$\overline{\lambda} = \lambda\,\Pi_0 + \sum_{n=1}^{\infty} \lambda(1-\alpha)\Pi_n = \lambda\left(\Pi_0 + (1-\alpha)\sum_{n=1}^{\infty}\Pi_n\right) =$$

$$= \lambda\,(\Pi_0 + (1-\alpha)(1-\Pi_0)) = \lambda\,(1-\alpha(1-\Pi_0)) =$$

$$= \lambda\left(1 - \alpha\frac{\psi}{1+\alpha\,\psi}\right) = \frac{\lambda}{1+\alpha\,\psi}$$

$$T_q = \frac{L_q}{\overline{\lambda}} = \frac{1}{\lambda}\frac{\psi^2(1-\alpha)}{(1-\psi(1-\alpha))} \underset{\alpha=0}{=} \frac{\lambda}{\mu-\lambda}$$

b. Si $\alpha = 1$ el rechazo a entrar cuando la unidad está ocupada es total:

$$\lambda_n = \begin{cases} \lambda & n = 0 \\ 0 & n \geqslant 1 \end{cases} \Rightarrow \Pi_1 = \frac{\lambda}{\mu}\Pi_0,\ \Pi_n = \frac{0}{\mu}\Pi_0 = 0\ \forall\, n > 1 \Rightarrow$$

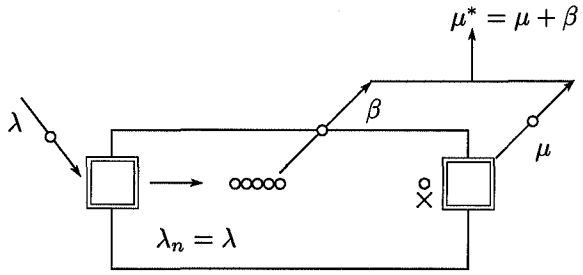
$$\Psi\Pi_0 + \Pi_0 = 1 \Rightarrow \boxed{\Pi_0 = \frac{1}{1+\psi},\ \Pi_1 = \frac{\psi}{1+\psi}} \qquad \boxed{\emptyset = 1\,\Pi_0 = \frac{1}{1+\psi}}$$

$$\boxed{L = 0\,\Pi_0 + 1\,\Pi_1 = \Pi_1 = \frac{\psi}{1+\psi}} = L_q + s - \emptyset$$

$$0 \quad 1 \quad \frac{1}{1+\psi}$$

R. Alonso

PROBLEMA N. 78



1. $\Pi_n = \dfrac{\lambda^n}{\mu(\mu+\beta)^{n-1}}\Pi_0 \;\forall\, n > 1$

$$\left(1 + \sum_{n=1}^{\infty} \frac{\lambda^n}{\mu(\mu+\beta)^{n-1}}\right)\Pi_0 = 1 \Rightarrow \Pi_0 = \frac{\mu(\mu+\beta-\lambda)}{\mu^2 + (\lambda+\mu)\beta}$$

$$\Pi_0^{-1} = 1 + \frac{\lambda}{\mu}\frac{1}{1 - \dfrac{\lambda}{\mu+\beta}} = 1 + \frac{\lambda}{\mu}\frac{\mu+\beta}{\mu+\beta-\lambda} = \frac{\mu^2 + (\lambda+\mu)\beta}{\mu(\mu+\beta-\lambda)}$$

$$G(s) = \Pi_0 + \sum_{n=1}^{\infty} s^n \frac{\lambda^n}{\mu(\mu+\beta)^{n-1}}\Pi_0 = \Pi_0\left\{1 + s\psi\sum_{n=1}^{\infty}\left(\frac{s\lambda}{\mu+\beta}\right)^{n-1}\right\} =$$

$$= \Pi_0\left\{1 + s\psi\frac{1}{1 - \dfrac{s\lambda}{\mu+\beta}}\right\} = \Pi_0\left\{1 + \psi(\mu+\beta)\frac{s}{\mu+\beta-s\lambda}\right\} \Rightarrow$$

$$G'(s) = \Pi_0\left(\psi(\mu+\beta)\frac{(\mu+\beta-s\lambda) - s(-\lambda)}{(\mu+\beta-s\lambda)^2}\right)$$

2. $L = G'(1) = \dfrac{\mu(\mu+\beta-\lambda)}{\mu^2+(\lambda+\mu)\beta}\left(\frac{\lambda}{\mu}/(\mu+\beta)\frac{\mu+\beta}{(\mu+\beta-\lambda)^2}\right) =$

$$= \frac{\lambda}{\mu^2+(\lambda+\mu)\beta}\frac{(\mu+\beta)^2}{\mu+\beta-\lambda} \underset{\beta=0}{=} \frac{\psi}{1-\psi}$$

R. Alonso

PROBLEMA N. 79

a) $\Pi_n = \dfrac{\lambda(\lambda(1-\alpha))^{n-1}}{\mu(\mu+\alpha)^{n-1}}\Pi_0 \qquad \left(1+\sum_{n=1}^{\infty}\frac{\lambda}{\mu}\frac{(\lambda(1-\alpha)^{n-1}}{(\mu+\beta)^{n-1}}\right)\Pi_0 = 1$

$$\Pi_0^{-1} = 1+\frac{\lambda}{\mu}\frac{1}{1-\dfrac{\lambda(1-\alpha)}{\mu+\beta}} = 1+\frac{\lambda}{\mu}\frac{\mu+\beta}{\mu+\beta-\lambda(1-\alpha)} =$$

$$= \frac{\mu^2+(\lambda+\mu)\beta+\alpha\lambda\mu}{\mu(\mu+\beta-(1-\alpha)\lambda)}$$

b) $G(s) = \Pi_0 + \sum_{n=1}^{\infty} s^n\left(\psi\left(\dfrac{\lambda(1-\alpha)}{\mu+\beta}\right)^{n-1}\right)\Pi_0 =$

$$= \Pi_0\left(1+s\psi\sum_{n=1}^{\infty}\left(s\frac{\lambda(1-\alpha)}{\mu+\beta}\right)^{n-1}\right) =$$

$$= \Pi_0\left(1+s\psi\frac{1}{1-s\dfrac{(1-\alpha)\lambda}{\mu+\beta}}\right) = \Pi_0\left(1+s\psi\frac{\mu+\beta}{\mu+\beta-s\lambda(1-\alpha)}\right)$$

$$G'(s) = \Pi_0\left(\psi(\mu+\beta)\frac{\mu+\beta-s\lambda(1-\alpha)+s\lambda(1-\alpha)}{(\mu+\beta-s\lambda(1-\alpha))^2}\right)$$

$$L = G'(1) = \frac{\mu(\mu+\beta-(1-\alpha)\lambda)}{\mu^2+(\lambda+\mu)\beta+\alpha\lambda\mu}\left(\frac{\lambda}{\mu}(\mu+\beta)\frac{\mu+\beta}{(\mu+\beta-\lambda(1-\alpha))^2}\right) =$$

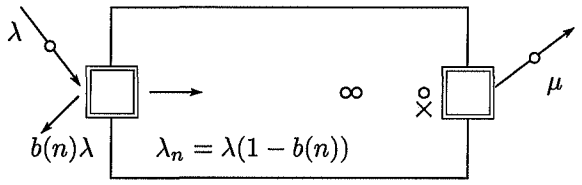
$$= \frac{\lambda}{\mu^2+(\lambda+\mu)\beta+\alpha\lambda\mu}\left(\frac{(\mu+\beta)^2}{(\mu+\beta-\lambda(1-\alpha))}\right) =$$

$$= \begin{cases} \dfrac{\lambda}{\mu^2+(\lambda+\mu)\beta}\left(\dfrac{(\mu-\beta)^2}{\mu+\beta-\lambda}\right), & \alpha = 0 \\ \dfrac{\psi}{(1+\alpha\psi)(1-\psi(1-\alpha))}, & \beta = 0 \end{cases}$$

R. Alonso

PROBLEMA N. 80

Las tasas de *entrada* son ahora: $\lambda_0 = \lambda$; $\lambda_n = \lambda(1-\alpha).n \geqslant 1$.



a) $\lambda_n = \left(1 - \dfrac{n}{1+n}\right)\lambda = \dfrac{1}{1+n}\lambda \Rightarrow \Pi_n = \dfrac{\frac{1}{1}\lambda\frac{1}{2}\lambda\cdots\frac{1}{n}\lambda}{\mu\mu\cdots\mu}\Pi_0 =$

$$= \frac{\psi^n}{n!}\Pi_0 \Rightarrow \Pi_0 = e^{-\psi} \Rightarrow \Pi_n = \frac{\psi^n}{n!}e^{-\psi} \equiv X \in \mathcal{P}(\psi)$$

b) $L = E(X) = \psi \quad L_q = L - s + \emptyset = L - (1 - \Pi_0) = \psi - 1 + e^{-\psi}$

R. Alonso

PROBLEMA N. 81

a) $\Pi_1 = 1\,\Pi_0$, $\Pi_2 = 1(\Pi_1 = \Pi_0)$, $\ldots$, $\Pi_n = \Pi_0$, $\forall\, n \leqslant K$;

$$1 = \sum_{n=0}^{K} \Pi_0 \;\Rightarrow\; \boxed{\Pi_n = \frac{1}{1+K}, \quad n \leqslant K}$$

b) $$L = \sum_{n=0}^{K} n\,\Pi_0 = \frac{1}{1+K}\frac{K(K+1)}{2} = \frac{K}{2}$$

c) $$\emptyset = 1 \times \Pi_0 = \frac{1}{1+K}$$

d) $$L_q = L - s + \emptyset = \frac{K}{2} - 1 + \frac{1}{1+K} = \frac{K(1+K) - 2(1+K) + 2}{2(1+K)} \overset{2}{=} \frac{K(K-1)}{2(1+K)}$$

R. Alonso

---

[2] $L_q = \sum_{n=2}^{K} (n-1)\Pi_0 = \frac{1}{1+K}\sum_{n=2}^{K} (n-1) = \frac{1}{1+K}\sum_{n=1}^{K-1} n = \frac{1}{1+K}\frac{(K-1)K}{2}$.

PROBLEMES PROPOSATS

PROBLEMA N. 82

Let $x_1, \ldots, x_n$ independently distributed with $x_i \sim N_p(\mu_i, \Omega)$, where Ω is nonsingular. It is assumed that $n > p$ and that the matriz $M' = (\mu_1, \ldots, \mu_n)$ is of rank 1. Define

$$S = \sum_{i=1}^{n} x_i x_i'$$

Find $E(S^{-1})$ as a second order approximation to the exact solution of Steerneman (1997, 1999), given by

$$E(S^{-1}) = \frac{1}{n-(p+1)} \Omega^{-1}.$$

(The author invite readers to propose a solution).

References

Steerneman, A.G.M. (1997). «Problem 331», *Statistica. Neerlandika*, 51, 381.

Steerneman, A.G.M. (1999). «Solution 331», *Statistica. Neerlandika*, 53, 252-254.

Heinz Neudecker
Cesaro Schagen
heinz@fee.uva.nl

PROBLEMA N. 83

El estimador de regresión lineal multivariante de la media de una población finita de tamaño N, $\overline{Y} = (1/N)\sum_{i=1}^{N} y_i$, usando el principio de mínimos cuadrados en base a n observaciones de la variable de interés y, siendo x_j $(j = 1, 2, \ldots, k)$ la j-ésima variable auxiliar, es:

$$\widehat{\overline{Y}} = \overline{y} + \sum_{j=1}^{k} b_j \left(\overline{X}_j - \overline{x}_j\right),$$

donde $\overline{y}$ es la media muestral de la variable de interés, b_j el coeficiente de regresión j-ésimo, y $\overline{x}_j$ y $\overline{X}_j$ son la media muestral y poblacional respectivamente para la j-ésima variable auxiliar. Demostrar que si los coeficientes de regresión están acotados, el estimador $\widehat{\overline{Y}}$ es consistente.

M. Ruíz Espejo
UNED

Housila P. Singh
Vikram University

Novetats de Software

DEMONSTRATION OF THE POWER OF STABLE: DEVELOPMENT OF STATISTICAL APPLICATIONS USING A NEW VISUAL PROGRAMMING ENVIRONMENT

A. PRAT, G. PRATS, I. SOLÉ and J.M. CATOT
Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

D. FARRAS
Cebal Entec

The European Commission funded the project STABLE in 1997. The objective of STABLE was to construct statistical software using a visual environment called IRIS Explorer. The STABLE system is an integration of an existing application building system, IRIS Explorer, and an existing widely used statistical software system, GENSTAT. This system will join two basic characteristics that will make it flexible and competitive: on one hand, the easy interaction that provides a visual programming environment with the user along with useful visualisation facilities; on the other hand the ability to produce tailored end-user applications.

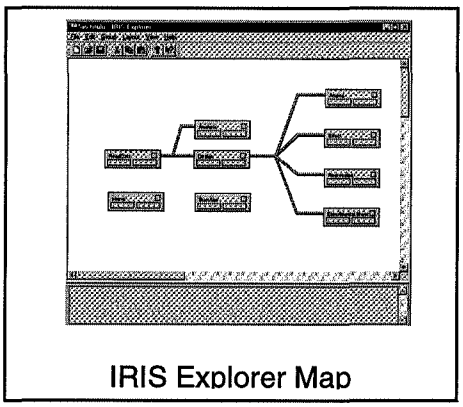
In order to prove the potential of STABLE in building tailored end-user software, as well as to provide feedback to the STABLE developers, three very diverse and challenging industrial applications were chosen. From these cases several different demonstrators were build and used, so the system has been studied under several situations. The main characteristics of STABLE are described by Ford et al. (1998).

Challenges of STABLE project

There is a need in the industry for tailored system, the reasons are the following ones:

- The need to increase the application of statistical methods in industry and service organisations is growing up, as well as the need of applications that would be easy to programme and use.
- The need to have detailed statistical analysis performed quickly by end-user in Industry without much programming.
- Adapt the interface to the appropriate terminology of the end-user.

- Produce applications tailored to the existing needs and that can be easily modified to handle any further requirements in the future without knowledge of the programming languages.



IRIS Explorer Map

It is known that from the origins the hardware has been described under the module concept. This means that although the basic part of the hardware is a transistor, in order to understand the system the level of abstraction arrives util big blocks or modules that contains millions of transistors. This makes possible to divide the functionality of the system in parts or modules. The last tendencies in programming languages are visual software, and lately doing a simile with the hardware, module programming is growing up. IRIS Explorer [NAG (1995)] is using this new generation of programming language, which is currently aimed at computational physicists, chemists and engineers. The IRIS Explorer data types are therefore designed to hold the data structures used by these workers. However, IRIS Explorer was never intended to be a closed system, so the possibility to create new types to handle new data structures is allowed. At present there is an ongoing project in order to integrate the functionality of some statistical package into IRIS Explorer which will enlarge the library of statistical modules that already exist. An expert user should be able to develop himself his own statistical modules.

The Consortium selected three end-user applications:

- GESA (Spain), with a system for forecasting electricity demand in the Balearic Islands.
- CEBAL ENTEC (Spain), with a system for speed up the design process of manufacturing pressure resistant containers for aluminium.
- LIMAGRAIN (France), with a system for the analysis of field trials.

A can designed by Cebal

In this report we will describe the Cebal Entec prototype, this is a System for Experimental Design in order to improve the process of aluminium impact extrusion. Cebal Entec is a French multinational company that belongs to the Pechiney Group. It is known as one of the most important manufacturers of pressure resistant containers

from aluminium by impact extrusion. This technique is being used by Cebal Entec on the plant that it has in Badalona.

The Specifications of Cebal Entec

From the common study realised by UPC (Polytechnical University of Catalonia) and Cebal Entec it was agreed that the application required should be able to solve designs of factorial experiments as well as fractional factorial with factors at two levels [Prat et al. (1997)]. Replicates and blocks should be allowed. From that point an application has been build using the available statistical modules.

The demonstrator will consist of different subsets of modules, the main groups to consider are those ones:

- *Section of the design*: initial menu is displayed, a design can be defined or imported from a file, this part takes care of all initial variables.
- *Entry of experimental data*: allows the user to enter the experimental data values collected and place them along with the initial design, all in the same structure.
- *First results and transformations*: shows all relevant information about the initial data (exploratory data analysis) and it has the option to perform different data transformations.
- *Selection of a model*: based in previous information and graphics, allows the user to select the right model to be displayed in the Anova table, deciding which variables to include and which ones not to. The system is dynamic so the user can either analyse and modify the conditions at the same time.
- *Final graphics and results*: from the previous analyses, the user should arrive to a conclusion. This can be evaluated in depth with the final information displayed by the programme. Sometimes the conclusion could be that more experimentation should be done.

How do we create a Demonstrator?

The first step in the general process for the development of the demonstrators is the production of the specifications of the system. This was done following the structure provided in an internal STABLE document [the Guidelines showed in Prat and Catot (1997)] and obtained after an analysis of the user needs and the study of the original possibilities of GENSTAT [NAG (1999)] and NAG library [NAG (1995)].

These Guidelines recommend starting the specifications of each demonstrator with a general description of the major functions and components of the demonstrator. The description of the problem to be solved and the needs of the organisation must be clear, as well as the formulation of the conceptual model of the new solution. Finally a detai-

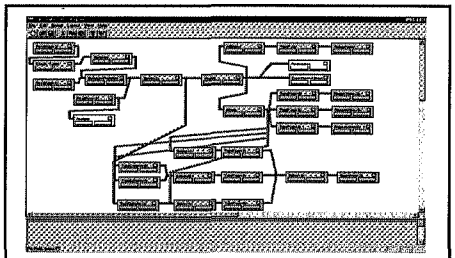
led description of inputs, outputs, operations, data transformations and algorithms used has to be done.

The second step in the development is the construction of the demonstrator, from the specifications produced in the previous step, creating a map using the modules produced by RES (IACR-Rothamsted Experimental Station) and NAG (The Numerical Algorithm Group Ltd.), and designing the end-user interface for each one of the modules. The power of the system allows the end-user to modify in an easy way different aspects of the application, so that it can be personalised to his specific needs.

The last step of the process is the validation of the demonstrators developed, proving the interest of the industrial organisations for their use.

How do we build an application?

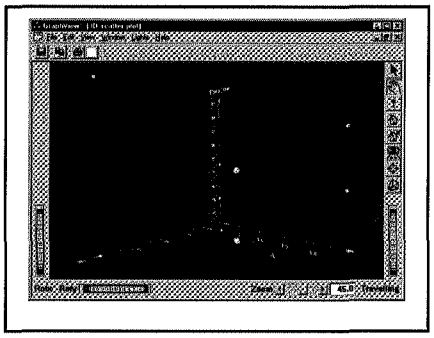
The initial process of building an application is not much different than other programming languages. The programmer must have a clear idea of what does he wants the application to do, which will be the parameters that the final user will have to fix and which will be the ones that will be fixed by the programmer. All this information must be extracted from the previous analysis of the Demonstrator building. On the other side a clear idea of the visual interface must be defined. Although this interface is dynamic and can be rebuild it is very important from the beginning to have clear ideas. So first thing to do once the programme has been designed in mind is to select the modules from the library in order to obtain a map. At this point the programmer would have all the information available and it is his personal decision, to decide which of this information is relevant to the final user, and which must be hidden. That is a basic part of the process, because depending on how much information the final user will see, and how simple it is presented on the interface, the application will be much or less easy to use. The power of the software used in the STABLE [Ford et al. (1998)] applications allows the programmer to change the visual interface adding news capabilities.


Map formed by several modules

Another strong point of the IRIS Explorer system [Payne and Harding (1997) and Craig (1997)] is that it has a very intuitive way to present the information. A very complex application that can be build by 30 or 40 modules and that will have more that 60 connection lines, can be displayed to the final user in just 5 modules with no more that 10 connection lines. This is possible because the software has some options in order to make groups from the current modules changing them into new modules that contain all the information of the group, so that the user can understand what is basically doing

each part of the programme. For example there is a zone that treats the input data and transforms it into the type of variable used by the IRIS Explorer. Other part evaluates the data and does the calculations, another evaluates the graphical output display and so on. So each part at the same time is formed by several modules but displaying the information that way the user can easily understand what is each module doing, and to which sector does it belong.

Building the final user interface



Once the map is designed and also all modules are placed and connected, the software is ready to be run. No interface is needed to be build to run the programme in order to check its capabilities. Once the result is satisfactory enough the programmer can build an interface in order to facilitate the interaction between the machine and the user. This interface can be modified at any time. The user can make its own groups of modules in order to make the application more understandable to the user, and after an interface can be designed from each group of modules.

One of the more powerful characteristics of the IRIS Explorer system is its potential graphical display. All kind of graphics are allowed (even three- dimension plots) with the possibility to rotate them, zoom effects, etc. This becomes very helpful when the user must decide between some aspects of the model displayed by the programme. As much information is displayed in a graphical way the user could decide with more security.

Conclusions

The conclusions extracted about the performance of STABLE* are very positive. The initial expectations were completely satisfied with the final result. As it has been explained before a demonstrator was designed in order to solve the design of experiment of Cebal Entec. The application was really easy to programme because of the capabili-

* STABLE is partially funded by EU ESPRIT 4 Project 22832 and by the CICYT TIC97-1446-CE, with the Institut d'Estadística de Catalunya (IDESCAT) acting as a member of its Board of Advisors.

ties of the system. One of the advantages of using the IRIS Explorer software was that apart from the statistical modules, it has a complete library with a lot of modules related to different subjects, so this makes a lot more powerful the system. Professional graphical display was done with almost non expertise in previous programming. The system is very intuitive. Another important aspect of the software was to verify that the interface is easily removable, having the possibility to modify or remove it in order to adapt to the end user requirements.

Software and hardware requirements. Licences

IRIS Explorer software runs under Windows NT. In order to run the system at a rational speed it must be installed under Pentium processors at a recommended speed of 200 MHz.

A licence for the use of STABLE system, which includes IRIS EXPLORER and the reenginered modules of GENSTAT and NAG library, can be obtained contacting NAG-The Numerical Algorithms Group Ltd at Wilkinson House, Jordan Hill Road, Oxford, OX2 8DR, England [Phone +44 (0) 1865 511245; fax +44 (0) 1865 310139].

References

Craig, P. (1997). *Implementing a statistical data type in IRIS Explorer.* The Numerical Algorithms Group Ltd.

Ford, B. et al. (1998). «STABLE: A programming visual environment for the development of statistical applications», in *Proceedings of New Tools and Technologies in Statistics 1998, Sorrento.* Contributed Paper

NAG (1995). *IRIS Explorer User's Guide.* The Numerical Algorithms Group Ltd.

NAG (1999). *GENSTAT 5.* The Numerical Algorithms Group Ltd. (see website in: www.nag.co.uk/stats)

Payne, R.W. and Harding, S.A. (1997). *Data Structures for STABLE.* Rothamsted Experimental Station.

Prat, A., X. Tort-Martorell, P. Grima y L. Pozueta (1997). *Métodos estadísticos, control y mejora de la calidad.* First Edition, Edicions UPC.

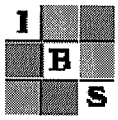
Prat, A. and Catot, J.M. (1997). «Guidelines for the Specifications of a Demonstrator. STABLE Project». *Internal document.* Statistics Department and Operational Research, UPC. June 1997.

Ressenyes d'activitats institucionals

Sociedad Española de Biometría

Región Española
de la Sociedad Internacional de Biometría

Spanish Region
of the International Biometric Society



http://www.iata.csic.es/ibsresp

La **Sociedad Española de Biometría/Región Española de la Sociedad Internacional de Biometría** (abreviadamente **SEB** o ***REsp***) tiene como objetivos promover, impulsar y difundir el desarrollo y la aplicación de los métodos matemáticos y estadísticos a la biología, medicina, psicología, farmacología, agricultura y otras ciencias afines (ciencias relacionadas con los seres vivos). Cualquier profesional o alumno de estas disciplinas puede ser miembro de la SEB.

Consejo Directivo

Presidenta:	Guadalupe Gómez i Melis (Biología)
Vicepresidente:	Emilio A. Carbonell Guevara (Agronomía)
Secretario y Tesorero:	Fernando López Santoveña (Agronomía)
Vocal en calidad de Miembro del Consejo de la IBS:	Carles M. Cuadras Avellana (Biología)
Vocales:	María Jesús Bayarri García (Medicina)
	Juan Luis Chorro Gascó (Psicología)
	Rosa Estarelles Rodríguez (Psicología)
	José Luis González Andújar (Agronomía)
	Martín Ríos Alcolea (Medicina)
	Alex Sánchez Pla (Biología)
Corresponsal de la REsp en el «Biometric Bulletin» de la IBS:	María Luz Calle Rosingana

La SEB promueve directamente o participa en la promoción de cursos monográficos sobre distintas técnicas de Análisis Estadístico.

Durante 1998 se celebró un curso sobre «Regresión Logística» y otro sobre «Modelos Mixtos con S-Plus», en colaboración con la Universidad Politécnica de Valencia.

En 1999 se ha celebrado en Barcelona un curso sobre Análisis de Supervivencia con S-Plus, en colaboración con la Fundació Politècnica de Catalunya, y ha tenido lugar en Palma de Mallorca la VII Conferencia Española de Biometría.

Próximamente se anunciarán nuevos cursos y actividades.

The TES Institute
Training of European Statisticians

GENERAL INTRODUCTION

Seeing the need of harmonising statistics at the European level, Eurostat decided in the early nineties to set up the TES Project. This programme was in charge of offering truly European vocational training and staff development opportunities at post-graduate level through annual training programmes for target groups ranging from young statisticians to executives of National Statistical Institutes.

In November 1996. the TES Project became the TES Institute, a non-profit association created by ten Member States of the European union and the four Member States of the European Free Trade Association. At present it counts among its members the representatives of seventeen European National Statistical Institute and of the Centre Universitaire de Luxembourg.

The training programmes offered by the TES Institute provide both theoretical and practical background but the courses have all a very strong applied character.

These programmes also offer participants the opportunity to meet colleagues form all over Europe and other countries since the TES Institute has extended its activities to the Central European, Mediterranean Basin and TACIS countries.

The above characteristics represent the basic conditions to acquire sharper competence in their work environment and highlight the European dimension of their activity.

After ten years of existence, the programme became entire part of the statistical world. For the time being, around 500 participants coming from more than 30 countries are trained every academic year. Such an interest is mainly due to the large number of courses on offer. Indeed, the TES portfolio comprises more than 80 courses of short duration all at post-graduate level.

After a few years of co-operation with the Central European countries, the TES Institute has recently extended the co-operation to the MEDSTAT and TACIS region. Such an internationalisation is the direct result of the growing importance of training as a part of the current technological and intellectual development. Therefore, as far as statistics and economics are concerned, it is of the utmost importance to extend the best national practices to an international level.

It is obvious that the TES programmes should be considered as a complement and not a substitute to the training provided at national level.

In brief, one may say that by offering training opportunities which are complementary to the ones provided at national level, the TES Institute is offering a new approach of the subsidiarity concept.

COURSES ON OFFER

The 1999-2000 Programme offering 28 courses has started from September 1999 and will last until June 2000.

Due to a delay in making PHARE funds available for CEC countries, it has been decided to postpone the following courses which are of particular interest for this region:

- The Revised System of Accounts (ESA95) - Sector Accounts
- The Revised System of Accounts (ESA95) - Quarterly Accounts

The new dates for the above courses can be found in the table below.

On the other hand the course on *Sampling Techniques and Practice* in French has been cancelled seeing the very low interest among the Member States.

The table below indicates the courses to be held from January to June 2000.

Course Title	Course Leader	Location	Dates
The Revised European System of Accounts (ESA 95) - Goods and Services	Konijn	Luxembourg	10-12 Jan-00
Basic Principles of Publication and Dissemination of Statistical Products	Swires-Hennessy	London	17-21 Jan-00
Living Conditions, Social Indicators	Social Reporting Everaers	Lisbon	24-27 Jan-00
Introduction to the Analysis of Multivariate Discrete Data	Israels	Voorburg	07-11 Feb-00
Symbolic Data Analysis	Diday	Paris	07-11 Feb-00
Nomenclatures, Classifications and Their Harmonisation	Langkjaer	Luxembourg	14-17 Feb-00
Workshop on the Internet	Feldbaek	Neuchâtel	17-18 Feb-00
The European Statistical System	Prieto	Luxembourg	06-08 Mar-00
Confidentiality and Protection of Privacy	Nanopoulos	Luxembourg	13-15 Mar-00
Adding Value through Strategic Management	Scrivener	London	20-24 Mar-00
Dealing with non-Response	Lynn	London	3-7 Apr-00
Seasonal Adjustment Methods	Maravall	Luxembourg	10-14 Apr-00
The Revised European System of Accounts (ESA 95) - Financial Accounts	Coin	Luxembourg	26-28 Apr-00
Notions of Sampling and Survey for Managers	Droesbeke	Rome	08-10 May-00
The Revised European System of Accounts (ESA 95) - Sector Accounts	Newson	Luxembourg	15-17 May 00
Demographic Data and Their Analysis	Andersen	Copenhagen	22-26 May-00
The Revised European System of Accounts (ESA 95) - Quarterly Accounts	Mazzi	Luxembourg	22-25 May 00
Comparative Analysis of Statistical Packages and Data Bases for Statistics	Cole & Campbell	Manchester	05-07 Jun-00
Marketing and Sales of Statistical Products and Services	Carlsen	Copenhagen	19-21 Jun-00
National Accounts Statistics in Practice	Lequiller	Paris	19-30 Jun-00
Sampling Techniques and Practice	Smith	Southampton	19-30 Jun-00

CONSULTING

Apart from the above mentioned training activities, the TES Institute becomes more and more involved in consulting activities. The main objective of setting-up of a training centre for statisticians in any specific country will be reached through the following actions:

- Consulting for curriculum development for their future training programme.
- Training of future trainers of the programme.

PUBLICATIONS

The TES Institute has started the production of TES Manuals on subjects covered by the vocational training programme.

The first manual available at the TES Institute presents *The Role of Statistics in a Democracy.*

The second manual to be published in January will cover the *Indices for bilateral and multilateral Comparison of Prices, Quantities and Values.*

Further manuals will cover topics of *Sampling Techniques, Seasonal Adjustment Methods* and *Social Statistics.*

OTHER ACTIVITIES

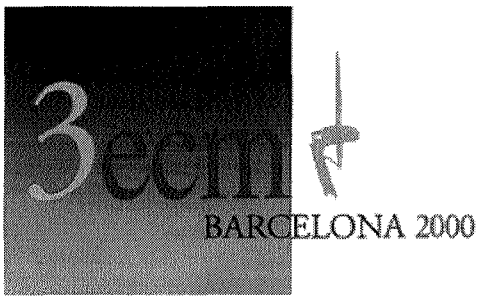
The TES Institute has been associated with the Maastricht School of Management as training and advice provider in the framework of their MBA in Decision Support Systems.

You can directly contact the TES Institute for any further information on this MBA.

GENERAL INFORMATION

For further details on any of the above mentioned topics, please contact Ms Valérie Vandewalle co-ordinator for *Evaluation and Information* matters:

by phone: (352) 29.85.85.34
fax: (352) 29.85.29
or e-mail: vvandewalle@tes-institute.lu



Tercer Congrés Europeu de Matemàtiques

Barcelona, del 10 al 14 de juliol del 2000

Primer Anunci

El Comitè Organitzador es complau a anunciar que el **Tercer Congrés Europeu de Matemàtiques (3ecm)** tindrà lloc a Barcelona del 10 al 14 de juliol de l'any 2000. L'organitza la Societat Catalana de Matemàtiques (SCM), sota els auspicis de la Societat Matemàtica Europea (EMS).

Conferenciants plenaris

- **Robbert Dijkgraaf** (Universitat d'Amsterdam, Holanda)
- **Hans Föllmer** (Universitat Humboldt de Berlín, Alemanya)
- **Hendrik W. Lenstra, Jr.** (Universitat de Califòrnia a Berkeley, Estats Units, i Universitat de Leiden, Holanda)
- **Yuri I. Manin** (Institut Max Planck de Matemàtiques, Bonn, Alemanya)
- **Yves Meyer** (Escola Normal Superior de Cachan, França)
- Carles Simó (Universitat de Barcelona)
- **Marie-France Vignéras** (Universitat de París 7, França)
- **Oleg Viro** (Universitat d'Uppsala, Suècia, i POMI de Sant Petersburg, Rússia)
- **Andrew J. Wiles** (Universitat de Princeton, Estats Units)

Programa científic

El programa del congrés inclourà nou conferències plenàries, trenta conferències invitades en sessions paral·leles, conferències impartides pels guardonats amb els premis de l'EMS, minisimposis, taules rodones i sessions de pòsters. Igual com es va fer en els congressos europeus anteriors, s'atorgarà un cert nombre de premis a investigadors/res joves en matemàtiques, de menys de trenta-dos anys d'edat. Els minisimposis són una de les novetats del **3ecm**; el Comitè Científic escollirà una llista de temes actuals i interdisciplinaris. El programa complet de conferències, minisimposis i taules rodones s'especificarà en el segon anunci. Tots els participants podran presentar comunicacions en forma de pòsters. També està previst que s'organitzin demostracions de programari matemàtic, vídeo i material multimèdia.

Amb finalitats organitzatives, s'ha establert la següent llista numerada de temes científics:

1. Lògica i fonaments
2. Àlgebra. Teoria de nombres
3. Geometria algebraica i analítica
4. Geometria diferencial
5. Topologia
6. Matemàtica discreta i informàtica
7. Modelització i simulació
8. Equacions diferencials ordinàries i sistemes dinàmics
9. Equacions en derivades parcials
10. Anàlisi funcional
11. Anàlisi complexa
12. Probabilitat i estadística
13. Anàlisi real
14. Física matemàtica

Comitès

- El Comitè Científic és presidit per **Sir Michael Atiyah** (Universitat d'Edimburg).
- El Comitè de Premis és presidit per **Jacques-Louis Lions** (Col·legi de França).
- El Comitè de Taules Rodones és presidit per **Miguel de Guzmán** (Universitat Complutense de Madrid).
- El Comitè Organitzador és presidit per **Sebastià Xambó Descamps** (Universitat Politècnica de Catalunya).

Presentació de pòsters

Tots els participants inscrits al **3ecm** podran presentar treballs en forma de pòsters. El Comitè Organitzador decidirà quins pòsters s'accepten a partir de resums que haurà d'haver rebut abans de l'1 d'abril del 2000. Us demanem que envieu el vostre resum preferiblement fent servir el programa que hi haurà a la pàgina web **http://www.iec.es/3ecm/posters.htm**. També es pot enviar per correu electrònic a **posters.3ecm@upc.es**, posant com a *subject* només el número de la secció escaient (vegeu la llista de temes indicada més amunt). Si no el podeu enviar electrònicament, utilitzeu l'adreça següent: *Pòsters 3ecm (Prof. Josep M. Font), Facultat de Matemàtiques, Universitat de Barcelona, Gran Via 585, 08007 Barcelona.*

Presentacions de programari matemàtic

Durant el congrés tindrà lloc una sessió de programari matemàtic, en la qual es podran presentar programes relacionats amb tots els camps de les matemàtiques i aplicables a objectius diversos. El Comitè Organitzador avaluarà les propostes i en seleccionarà un cert nombre, aplicant criteris d'originalitat matemàtica, novetat i possibilitats d'aplicació, i tenint en compte l'equilibri temàtic de la sessió.

Les propostes han d'arribar als organitzadors abans de l'1 de febrer del 2000. Es poden enviar electrònicament, fent servir el full que hi ha a la web **http://www.iec.es/3ecm/mathsoft.htm**, o bé a l'adreça **mathsoft.3ecm@upc.es**, posant com a *subject* només la paraula *mathsoft*. L'adreça següent també es pot utilitzar per enviar material complementari: *Mathsoft 3ecm (Prof. Santiago Zarzuela), Facultat de Matemàtiques, Universitat de Barcelona, Gran Via 585, 08007 Barcelona.*

Presentacions de vídeo i multimèdia

Durant el **3ecm** hi haurà un seguit d'activitats complementàries i actes culturals. Una d'aquestes activitats serà la producció d'un DVD amb vídeos i material multimèdia amb contingut matemàtic. Aquest DVD s'exhibirà en sessions públiques i també serà accessible en diversos llocs de la seu del **3ecm**. Es podran enviar contribucions per a aquest DVD des de totes les àrees de les matemàtiques.

Els treballs hauran d'arribar als organitzadors abans de l'1 de febrer del 2000. S'han d'enviar per correu ordinari a: *Video 3ecm (Prof. Santiago Zarzuela), Facultat de Matemàtiques, Universitat de Barcelona, Gran Via 585, 08007 Barcelona.* Trobareu més informació a la web **http://www.iec.es/3ecm/video.htm**. Podeu utilitzar l'adreça **video.3ecm@upc.es** per contactar amb els organitzadors d'aquesta activitat.

Activitats satèl·lit

Els congressos i les altres activitats de la llista següent han estat acceptats com a satèl·lits del **3ecm** pel Comitè Executiu abans del mes de febrer de 1999. Us encoratgem que feu la llista més llarga. Les propostes s'han de fer arribar al president del Comitè Organitzador abans de l'1 de febrer del 2000, per correu electrònic a **3ecm@iec.es** o bé per carta a la SCM.

- **Summer School on Interactions between Algebraic Topology and Invariant Theory. Ioannina, Grècia, del 26 de juny a l'1 de juliol del 2000.** *Contacteu amb:* Nondas Kechagias (Universitat de Ioannina), **nkechag@cc.uoi.gr.**
- **Functional Analysis Valencia 2000, an International Functional Analysis Meeting on the Occasion of the 70th Birthday of Professor Manuel Valdivia. València, del 3 al 7 de juliol del 2000.** *Contacteu amb:* José Bonet (Universitat de València), **vlc2000@mat.upv.es**, o bé Klaus D. Bierstedt (Universitat de Paderborn), **vlc2000@uni-paderborn.de.**
- **6th International Conference on Harmonic Analysis and Partial Differential Equations. El Escorial, Madrid, del 3 al 7 de juliol del 2000.** *Contacteu amb:* Eugenio Hernández (Universitat Autònoma de Madrid), **eugenio.hernandez@uam.es.**
- **Alhambra 2000, a Joint Mathematical European-Arabic Conference. Granada, del 3 al 7 de juliol del 2000.** *Contacteu amb:* Ceferino Ruiz (Universitat de Granada), **alhambra2000@ugr.es.**
- **First Euro-Mediterranean Topology Meeting. Bellaterra, del 4 al 7 de juliol del 2000.** *Contacteu amb:* Carlos Broto (Universitat Autònoma de Barcelona), **broto@mat.uab.es.**
- **cem 2000, Congrés d'Educació Matemàtica, I Jornades d'Educació Matemàtica a Catalunya. Mataró, del 3 al 5 o del 5 al 7 de juliol del 2000.** *Contacteu amb:* Xavier Vilella (FEEMCAT), **xvilella@pie.xtec.es.**
- **Distributions with Given Marginals and Statistical Modelling. Barcelona, del 17 al 19 de juliol del 2000.** *Contacteu amb:* Carles M. Cuadras (Universitat de Barcelona), **carlesm@porthos.bio.ub.es.**

Preinscripcions

Si desitgeu rebre el segon anunci i més informació per correu electrònic sobre el **3ecm**, us podeu preinscriure a través de la web **http://www.iec.es/3ecm** (si encara no ho heu fet). La preinscripció no costa diners ni us obliga a res. Per tal d'esdevenir participants del **3ecm**, caldrà que formalitzeu la inscripció quan s'obri el termini per fer-ho i pagueu la quota corresponent. També us podeu preinscriure per correu electrònic a l'adreça **3ecm@iec.es**, o bé enviant una carta a la SCM. Cal que indiqueu el vostre nom, la vostra institució, l'adreça postal completa, l'adreça de correu electrònic (si en teniu) i els camps científics que us interessin (en podeu escollir un o més d'un de la llista indicada més amunt).

Patrocinadors (relació actualitzada a febrer de 1999)

Generalitat de Catalunya, Comissionat per a Universitats i Recerca
Generalitat de Catalunya, Departament d'Ensenyament
Ministerio de Educación y Cultura, S.E.U.I.D.
Fundació Catalana per a la Recerca
Ajuntament de Barcelona
Institut d'Estudis Catalans
Universitat de Barcelona
Universitat Autònoma de Barcelona
Universitat Politècnica de Catalunya
Institut d'Estadística de Catalunya
International Mathematical Union
Real Sociedad Matemática Española
Sociedad Española de Matemática Aplicada
Fundación Retevisión
Fundació «la Caixa», Museu de la Ciència
Borsa de Barcelona
Port de Barcelona
Fundació Caixa Catalunya
Fundació Banc Sabadell
Fundació Caixa de Sabadell
Logic Control
Springer-Verlag

Adreces de contacte

Correu electrònic: **3ecm@iec.es**

Web: **http://www.iec.es/3ecm/** o també **http://www.si.upc.es/3ecm/**

Correu ordinari: Societat Catalana de Matemàtiques
Institut d'Estudis Catalans
Carrer del Carme, 47
08001 Barcelona

Telèfon: +34 93 270 16 20

Fax: +34 93 270 11 80

INTERNATIONAL WORKSHOP ON STATISTICAL MODELLING

Bilbao, Spain: Monday 17 to Friday 21 July, 2000

15th IWSM

New Trends in Statistical Modelling

First Announcement and Call for Papers

The International Workshop on Statistical Modelling concentrates on the various aspects of statistical modelling, including theoretical developments, applications and computational methods. Papers motivated by real practical problems are desirable, but theoretical contributions addressing problems of practical importance or related to software developments are also welcome.

The scientific programme is characterized by having invited lectures & tutorials, contributed papers, posters and software demonstrations. Contributed papers should be suitable for a 20 to 30 minutes oral presentation (including discussion) and focus on motivation, statement of key results and conclusions, and emphasize examples, wherever possible.

Invited speakers to date:

Christopher Bishop (Cambridge, UK), Joel L. Horowitz (Iowa City, Iowa, USA), Johannes Ledolter (Vienna, Austria), Winfried Stute (Giessen, Germany), Mark Steel (Edinburgh, UK), James Zidek (Vancouver, British Columbia, Canada), Dale L. Zimmerman (Iowa City, Iowa, USA).

A Tutorial on Goodness of Fit Tests for Regression Models will be given by W. González-Manteiga (Santiago de Compostela, Spain).

Students:

Professors should encourage students to attend the workshop. The programme is designed to allow for discussions and interchange between junior and senior scientists. A special session will be devoted for students contributions, an award for the best presentation will be given.

Scientific programme committee:

Ludwig Fahrmeir (Munich, Germany), Eva Ferreira (Bilbao, Spain, Co-chair), John Hinde (Exeter, U.K., Secretary), Michel Mouchart (Louvain-La-Neuve, Belgium), Vicente Núñez-Antón (Bilbao, Spain, Chair), Jean D. Opsomer (Ames, Iowa, U.S.A.), Juan Romo (Madrid, Spain), Esther Ruíz-Ortega (Madrid, Spain), Bill Venables (Australia).

Local organizing committee:

Ma. Victoria Esteban-González, Eva Ferreira, Petr Mariel, Vicente Núñez-Antón, Jesús Orbe-Lizundia, Susan Orbe-Mandaluniz, Marta Regúlez-Castillo, Juan M. Rodríguez-Póo, Gonzalo Rubio-Irigoyen, Fernando Tusell-Palmer.

Further information:

Details about registration for the workshop, instructions for authors and further information is available from the workshop homepage

http://iwsm.bs.ehu.es

Deadlines:

Jan 31: Submission of abstracts
Mar 13: Notification of acceptance
Apr 17: Submission of final manuscripts.

For additional information please contact:

Vicente Núñez-Antón
Departamento de Econometría y Estadística
Facultad de Ciencias Económicas y Empresariales
Universidad del País Vasco
Avda. Lehendakari Aguirre, 83
48015 Bilbao, Spain
Phone: +34 94 601 37 49
Fax: +34 94 601 37 54
E-mail: vn@alcib.bs.ehu.es

Informació per als autors i lectors

NORMES PER A LA PRESENTACIÓ D'ARTICLES A *QÜESTIIÓ*

La revista accepta, per a la seva publicació, articles originals no sotmesos a consideració en cap altra revista dins els àmbits de l'Estadística, la Investigació Operativa, l'Estadística Oficial i la Biometria. Els articles poden ser teòrics o aplicats, incloent aspectes computacionals i/o de caire docent, i poden presentar-se en anglès, francès, català o qualsevol altra llengua oficial a l'Estat espanyol.

Tots els originals destinats a les esmentades seccions temàtiques de *Qüestiió*, incloent-hi els articles per a la «Secció docent i problemes», seran sotmesos sistemàticament a una avaluació prèvia a càrrec d'especialistes independents i/o membres del Consell Editorial, llevat dels articles convidats per la revista i les reimpressions d'articles. El resultat de l'avaluació serà comunicat a l'autor principal als efectes d'eventuals correccions formals o dels seus continguts.

Per a totes les trameses d'originals, la revista emetrà un acusament de recepció la data del qual figurarà com a «data de rebuda» en la publicació de l'article. Per la seva banda, la «data d'acceptació» de l'article serà la data de recepció de la versió definitiva.

Per a la presentació d'articles, l'autor trametrà a la Secretaria de *Qüestiió* (Institut d'Estadística de Catalunya) dues còpies del treball mecanografiat en DIN A4, a una sola cara, a doble espai i amb marges amplis. Cada article ha d'incloure el títol, el nom de l'autor o autors, la seva afiliació i l'adreça completa, així com un resum de 75-100 paraules al principi de l'article, seguit de les principals paraules clau (en l'idioma original) i la seva adscripció a la classificació AMS. Abans de sotmetre els articles a la revista, s'aconsella els autors que revisin la correcció lingüística de textos d'acord amb l'idioma original i les eventuals traduccions a l'anglès.

Les referències bibliogràfiques es faran indicant el cognom de l'autor seguit de l'any de la publicació entre parèntesi [i.e.: Mahalanobis (1936), Rao (1982b)] i seran llistades alfabèticament al final de l'article; les referències múltiples d'un mateix autor s'ordenaran cronològicament. Les notes explicatives es numeraran correlativament i han d'aparèixer al peu de la pàgina corresponent. Les taules i figures també es numeraran correlativament en el text i seran reproduïdes directament dels originals tramesos en cas que no sigui possible la seva autoedició.

Una vegada avaluat satisfactòriament l'article cal que, a més de la versió impresa, l'autor el trameti en disquet de 3.5 polsades i en format MS-DOS, on han de constar de forma clara els noms dels autors i el títol de l'article. Aquesta versió final s'ha de trametre preferiblement en el processador de textos LaTeX2ε [subsidiàriament, es poden trametre els textos i les taules en Word Perfect —versió 6.0A o anterior— o ASCII]; en el cas de figures, diagrames o gràfics es recomanen els formats adients per als programes-editors PS, EPS o PCX. Els autors han de garantir la correspondència exacta entre la versió impresa i la còpia electrònica. D'altra banda, si l'article no està escrit en llengua anglesa s'haurà d'adjuntar la traducció del títol original, de l'abstract i de les paraules clau, així com un ampli resum en anglès (amb una extensió d'entre 2 i 5 pàgines i amb la mateixa estructura de l'article original).

La Secretaria de *Qüestiió* posa a disposició dels autors que ho sol·licitin plantilles en format LaTeX2ε de *Qüestiió* per a la seva edició i les referències adients de la classificació AMS.

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE SUBMISSION OF ARTICLES FOR *QÜESTIIÓ*

The journal well comes submission of articles and contributions that are not being considered for publication in any other journal in the fields of Statistics, Operational Research, Official Statistics or Biometrics. Articles may be theoretical or applied, including teaching aspects and applications, and will be accepted in English, French, Catalan or any of the other official languages in Spain.

All originals assigned to the thematic sections of *Qüestiió*, including articles for the «Teaching section and problems» will be systematically reviewed by independent referees and/or members of the Editorial Board, who will send a report to the main author of the article in order to correct, if necessary, any formal or content aspects. The articles invited by the journal and articles reprinted will be excluded from this evaluation process.

For all submissions, the journal will issue a receipt corresponding to the submission date, which will appear as «date received» in the final publication of the article. The «acceptance date» of the article, which will appear in its final publication, will be the date of sending the final version to the journal.

For the presentation of original articles, the author should send, to the Secretary of *Qüestiió* (Institut d'Estadística de Catalunya), two copies of the paper typed on A4 sheets, one side of the paper only, double spaced and with wide margins. Each article should include the title, the name of the author or authors, their affiliation, full address and also an abstract of the paper (75-100 words) at the beginning of the article, followed by the main keywords (in the original language) and its assignation in the AMS classification. Before submitting their papers, authors are advised to seek assistance in the writing of their articles for the correct use of English and/or of original language.

Bibliographical references should state the author's name followed by the year of publication in brackets [e.g.: Mahalanobis (1936), Rao (1982b)] and they should be listed at the end of the article in alphabetical order; multiple references to the same author should be given in chronological order. Footnotes should be numbered in the article and appear at the foot of the corresponding page. Figures and tables are to be numbered in consecutive order in the text using Arabic numerals and will be directly reproduced from the originals submitted if it is not impossible to print them electronically.

Once the evaluation has been passed, the author is required to provide the article on a diskette (a 3.5-inch disk in MS-DOS format) together with its paper copy; it must be a new diskette and must bear very clearly the names of the authors and the title of the article. This final version should be processed by LaTeX 2_ε, preferably, or, failing that, by Word Perfect (6.0A or earlier) or ASCII for text and tables; for figures, diagrams or graphs, the appropriate formats of PS, EPS or PCX software tools are strongly recommended. Authors must ensure that the version of the electronic copy is exactly the same as the paper copy which accompanies it. Furthermore, if the article is not written in English, the translation of its original title, short abstract and keywords should be enclosed, as well as a full summary of the article in English (that is, 2-5 pages with the same structure as the original).

The Secretary of *Qüestiió* can send, by request of the authors, the LaTeX 2_ε style of *Qüestiió* for manuscript preparation and the appropriate AMS classification references.

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS INSTITUCIONALS A *QÜESTIIÓ*

Qüestiió convida les entitats patrocinadores, les institucions col·laboradores, els organismes públics i privats, i tota la comunitat científica vinculada a l'estadística o la investigació operativa, a la publicació d'anuncis institucionals sobre cursos, seminaris, congressos i activitats similars que, preferentment, tinguin lloc en el nostre país. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les entitats interessades, de manera que *Qüestiió* no fa una cerca sistemàtica d'esdeveniments d'aquesta naturalesa, ni té cap ànim d'exhaustivitat en les ressenyes d'activitats finalment publicades.

Una vegada aprovada la inclusió dels anuncis sol·licitats es procedirà a la seva publicació, i es reproduirà directament dels originals tramesos amb les mides adequades i la màxima qualitat tipogràfica possible; en aquest cas, *Qüestiió* no procedeix a cap mena de procés d'autoedició de la versió impresa que l'anunciant hagi tramès. Si els originals es trameten en els mateixos termes electrònics exigits per als articles (vegeu «Normes per a la presentació d'articles a *Qüestiió*»), la revista procedirà a la seva autoedició. Si es desitja una qualitat superior a la reproducció simple o l'autoedició, o bé la seva publicació en color, els sol·licitants hauran de posar-se en contacte amb la Secretaria de *Qüestiió* per tal de trametre els fotolits dels textos originals corresponents.

La disposició dels textos i les figures adjuntes dels anuncis han de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de *Qüestiió* es reserva la decisió final pel que fa a la seva publicació.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la inserció en els números/volums de *Qüestiió* que prèviament s'hagi establert. La revista no es fa responsable dels retards, per part de l'anunciant, que impedeixin la publicació de l'anunci en els termes previstos.

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR INSTITUTIONAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió invites all sponsor entities, collaborating institutions, other public and private bodies and the entire scientific community related to Statistics or Operations Research to submit institutional advertisements on courses, seminars, congress and similar activities that will be held, preferably in our country. These will be accepted in English, French, Catalan or any of the other official languages in Spain. The initiative should always come from the entities interested in advertising them so that *Qüestiió*'s aim is not to do a systematic search of these events and therefore does not publish a comprehensive list of such activities.

Once their insertion is approved the advertisements will be reproduced from the most accurate photocopy of the originals sent by the advertiser to *Qüestiió* in paper copy, with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editing process to the printed version that the advertiser has sent. If the original advertisements are sent in the same electronic format requested by the articles (please see «Guidelines for the submission of articles for *Qüestiió*») the journal will print it directly from the file. If a better quality than the simple reproduction or automatic printing or a colour version of the adverts is desired, the authors should contact the Secretary of *Qüestiió* in order to negotiate this.

The typesetting of texts and figures in the advertisement should have maximum intelligibility and clearness, neither compressing the information too much nor using formats or letter fonts that are too small. Furthermore, the information has to be reliable, without errors and respectful of the people and institutions. The management of *Qüestiió* has the right to a final decision concerning the insertion of the advertisement.

Advertisers commit themselves to give the text/materials on request in order to insert them in the issues of *Qüestiió* that have been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published on the agreed terms.

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS PRIVATS O AMB FINALITAT COMERCIAL A *QÜESTIIÓ*

Qüestiió accepta la publicació d'anuncis privats o amb finalitat comercial sobre productes, serveis o altres eines promocionals a l'entorn de l'estadística o la investigació operativa. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les organitzacions que hi estiguin interessades, de manera que *Qüestiió* no fa una cerca sistemàtica de novetats o productes d'aquesta naturalesa ni té cap ànim d'exhaustivitat en els anuncis finalment publicats.

Els anuncis en **blanc i negre** s'elaboren a partir de la fotocòpia més acurada possible dels originals que trameti l'anunciant en versió impresa, amb les mides adequades i la màxima qualitat tipogràfica. Per tant, en aquest cas la revista no efectua cap procés d'edició ulterior respecte de la versió impresa que l'anunciant hagi tramès. Alternativament, si els anuncis originals es trameten en els mateixos termes formals exigits per als articles (vegeu «Normes per a la presentació d'articles a *Qüestiió*»), la revista procedirà a la seva autoedició. Igualment, si es desitja una qualitat superior a la reproducció simple, els sol·licitants hauran de trametre els fotolits dels originals corresponents o encarregar-los a *Qüestiió,* que els facturarà separadament.

Els anuncis **en color** requereixen els fotolits dels textos originals, que poden ser subministrats directament per l'anunciant o bé encarregats per la revista a compte de l'anunciant; en el segon cas, l'anunciant ha de trametre a la revista els originals impresos en color amb la màxima qualitat, per tal de filmar-los amb les millors garanties i condicions. El cost dels fotolits realitzats per *Qüestiió* serà sempre a càrrec de l'anunciant, a qui se li repercutirà l'import i l'IVA d'aquests, juntament amb les tarifes que corresponen a la modalitat d'anunci per la qual hagi optat.

La disposició dels textos i figures adjuntes dels anuncis ha de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de *Qüestiió* es reserva la decisió final de la seva inclusió.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la seva inserció en el(s) número(s)/volum(s) de *Qüestiió* que prèviament s'hagi establert. La revista no es fa responsable dels retards per part de l'anunciant que impedeixin la publicació de l'anunci en els termes previstos.

Imports:

1 pàgina en color (un número aïllat):	125.000 PTA + IVA
1 pàgina en color (tres números consecutius):	200.000 PTA + IVA
1 pàgina en blanc i negre (un número aïllat):	30.000 PTA + IVA
1 pàgina en blanc i negre (tres números consecutius):	50.000 PTA + IVA
1/2 pàgina en blanc i negre (un número aïllat):	20.000 PTA + IVA
1/2 pàgina en blanc i negre (tres números consecutius):	35.000 PTA + IVA

Mides opcionals dels anuncis:

1 pàgina sencera (espai intern):	19.0 cm. × 12.3 cm.
1 pàgina sencera (espai extern):	23.8 cm. × 17.0 cm.
1/2 pàgina (espai intern):	9.5 cm. × 12.3 cm.
1/2 pàgina (espai extern):	11.9 cm. × 17.0 cm.

Forma de Pagament:

- Transferència bancària al compte: 2013-0100-53-0200698577
- Xec bancari nominatiu a l'Institut d'Estadística de Catalunya
- Pagament amb targeta de crèdit

El pagament serà per l'import total de la factura corresponent, on hi figurarà el cost dels fotolits en el cas que l'edició de l'anunci hagi estat a càrrec de l'Institut. En el cas que l'anunciant necessiti una factura proforma, només cal que ho faci saber amb l'antelació suficient.

Correspondència:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE PRIVATE OR COMMERCIAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió accepts for their publication both private and commercial advertisements on products, services or other promotional tools related to statistics or operational research and will be accepted in English, French, Catalan or any of the official languages in Spain. The initiatives should always come from entities interested in advertising them so that *Qüestiió*'s aim is not to do a systematic search of news and therefore does not publish a comprehensive list of such private or profit activities.

The **black and white** advertisements are made out from the most accurate photocopy of the originals sent by the advertiser to *Qüestiió* in paper copy with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editorial process to the printed version that the advertiser has sent. Alternatively, if the original advertisements are sent in the same formal terms required by the articles (please see «Guidelines for the submission of articles for *Qüestiió*»), the journal will proceed to its autoedition. In the same way, if a better quality than the simple reproduction is wanted, the authors should send the photolits of the corresponding original texts or, on the other hand, order to *Qüestiió* their fulfilment, which will be invoiced separately from the rates charged as advertisements.

The advertisements **in colour** need the photolits of the original texts, which can be provided directly by the advertiser or requested by *Qüestiió* to the advertiser charge; in the second case, the advertiser must sent to the journal the originals printed in colour with the best possible quality, so that they can be filmed at the best conditions and guarantees. The cost of the photolits made by *Qüestiió* will always be charged to the advertiser together with the VAT derived from it, plus the prices corresponding to the type of the advertisement that has been chosen.

The set up of texts and figures of the advertisement should provide the maximum intelligibility and clearness, neither squeezing together the information nor using set ups or letter types that are too small. On the other hand the publicity has to be reliable, without fraud and respectful to the persons and institutions. The direction of *Qüestiió* has the right of the last decision concerning the insertion of the advertisement.

The advertiser committs himself to give the texts/materials on request, in order to insert them in the issue(s) of *Qüestiió* that had been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published in the agreed terms.

Rates:

1 colour page (only one issue):	125.000 PTA + VAT
1 colour page (three consecutive issues):	200.000 PTA + VAT
1 black and white page (only one issue):	30.000 PTA + VAT
1 black and white page (three consecutive issues):	50.000 PTA + VAT
1/2 black and white page (only one issue):	20.000 PTA + VAT
1/2 black and white page (three consecutive issues):	35.000 PTA + VAT

Advertisement sizes (optional):

1 full page (internal space):	19.0 cm. × 12.3 cm.
1 full page (external space):	23.8 cm. × 17.0 cm.
1/2 page (internal space):	9.5 cm. × 12.3 cm.
1/2 page (external space):	11.9 cm. × 17.0 cm.

Payment:

- A bank transfer to account number: 2013-0100-53-0200698577
- A bank cheque to Institut d'Estadística de Catalunya
- Charge on a credit card

The payment should be for the amount shown at the invoice, where it will be shouwn the total cost of the photolits, in case that *Qüestiió* would be in charge of the filmation of the advertisement. If advertiser need a pro-forma invoice, he should let us know some time in advance so that *Qüestiió* could send it to the proper address.

Mail address:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

Butlleta de subscripció a la revista ***Qüestiió***

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ Ciutat _____

Tel. _____ Fax _____ NIF _____

Data _____

Signatura

Desitjo subscriure'm a ***Qüestiió*** per a l'any 1999.
El preu de la subscripció és de 3.000 PTA (IVA inclòs).

Forma de pagament

☐ Transferència al compte 2013-0100-53-0200698577

☐ Domiciliació bancària al compte número

☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Xec nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retorneu aquesta butlleta (o una fotocòpia) a:

Qüestiió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Exemplar per a l'entitat bancària

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de la revista ***Qüestiió***

El sotasignat _____

autoritza el Banc/Caixa _____

Adreça _____

Codi postal _____ Ciutat _____

a abonar les subscripcions a la revista ***Qüestiió*** amb càrrec al seu compte

número ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Data _____

Signatura

Qüestiió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

620