Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 1998, volum 22, núm. 3
Segona època

Entitats patrocinadores:

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

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 1998, volum 22, núm. 3

SUMARI

PAPER ECOLOGIC

EDITORIAL

El darrer número del volum 22 inclou un total de set articles, amb la presència d'originals a les quatre seccions temàtiques de *Qüestiió*. D'aquesta manera, l'activitat editorial corresponent a l'any 1998 ha incorporat 23 articles, els quals suposen, juntament amb la informació de la resta de seccions de la revista, un volum de 574 pàgines. Aquestes xifres representen, de nou, un augment de la producció en relació al volum de l'any passat i, en general, dels promitjos enregistrats des del començament de la segona època de la revista (19 articles i 463 pàgines publicades per volum).

En segon lloc, cal esmentar que l'apartat de *Qüestiió* que es dedicava a la difusió de cursos, seminaris i congressos amplia a partir del present número la diversitat d'activitats referenciades, per la qual cosa la seva denominació («Ressenyes d'activitats institucionals») permet incloure les ressenyes d'altres tipus d'activitats que les institucions del nostre entorn organitzen en l'àmbit de l'estadística o la investigació operativa.

Finalment, i com és costum en l'editorial del darrer número de cada volum, *Qüestiió* fa balanç de l'activitat sobre la presentació i avaluació d'originals que han enregistrat les seves seccions. Atès que la darrera relació feia referència al període març-novembre 1997 (referenciat a l'editorial del darrer número del volum 21), la informació cobreix l'interval desembre 1997-novembre 1998.

- articles sotmesos: 25
- articles en procés d'avaluació: 29
- articles acceptats: 30 (26 publicats i 4 en espera de publicació)
- articles rebutjats: 7

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

Comentari de les seccions «Estadística», «Investigació Operativa» i «Biometria»

La secció «Estadística» conté tres originals. En el primer, «Almost unbiased ratio and product-type estimators in systematic sampling», de R. Singh i H.P. Singh, es defineix un estimador de la mitjana d'una població finita fent ús de la informació d'una altra variable auxiliar i es prova que és més eficient que l'estimador usual. Per la seva banda, l'article «Una distribución asintótica para un estimador natural del número de clusters de una población», de J.J. Prieto, proposa un estimador per al nombre de clusters estudiant la seva distribució assimptòtica i el seu biaix, que s'il·lustra

amb dades reals i simulades. El tercer article, «Algunas soluciones aproximadas para diseños split-plot con matrices de covarianza arbitrarias», de G. Vallejo i J.R. Escudero, és una contribució a l'Anàlisi de la Variància, sota el disseny «split-plot», quan el nombre de rèpliques és diferent, hi ha mesures repetides i la matriu de covariàncies dels errors és qualsevol.

L'únic article de la secció «Investigació Operativa», titulat «Tour eulerià sense girs en U en un graf orientat simple», de D. Soler, proporciona una condició suficient perquè un graf orientat sigui U-eulerià simple, verificable en temps polinomial i amb mínimes restriccions.

Finalment, l'article «The usefulness of discrimination based on distances on human evolution» inclòs a la secció «Biometria», de C. Arenas i D. Turbón, és una aplicació de l'anàlisi discriminant basat en distàncies entre observacions multivariants per a la identificació de cranis humans, que té avantatges sobre la discriminació clàssica, atès que permet el tractament de variables mixtes i dades faltants.

Carles M. Cuadras, editor executiu

Comentari de la secció «Estadística Oficial» i altres apartats

La secció «Estadística Oficial» acull dos articles ben diferenciats. En el primer cas, «La verificació aleatòria: una estratègia per millorar i avaluar la qualitat de l'entrada de dades», de J.A. Domènech, J.M. Losilla i M. Potell, l'anàlisi se centra en el procés inicial de la producció estadística, per al qual es proposa millorar l'eficàcia i la reducció dels costos de gravació amb la implementació del sistema DAT basat en la disminució d'errors d'operador no controlables de forma automàtica; un avantatge diferencial del mètode proposat rau en la generació d'indicadors d'aptituds, eficàcia i estimació dels percentatges d'errors efectius. En segon lloc, l'article de J.M. Mateo i J. Domingo-Ferrer, «A comparative study of microaggregation methods», aprofundeix en mètodes de microagregació per a la preservació del secret estadístic aplicats a registres individuals; la recent tesi doctoral del primer autor permet posar de manifest el notable avanç que ha experimentat l'enfocament basat en la microagregació orientada a les dades i la seva superioritat en l'àmbit de dades multivariants.

A continuació, la «Secció docent i problemes» prossegueix amb la presentació successiva de nous enunciats i la resolució dels problemes publicats en el número immediatament anterior.

Finalment, l'apartat dedicat a la ressenya d'activitats institucionals recull, en primer lloc, una presentació actualitzada de la Sociedad Española de Biometría, amb l'anunci

dels cursos que organitzará properament i de la convocatòria de la «VII Conferencia Española de Biometría» que se celebrarà l'any vinent (Palma de Mallorca, 10-12 de març de 1999). A continuació, figura una nova presentació del «Training for European Statisticians (TES) Institute» que *Qüestiió* ja redifon des del volum 21 (1997), juntament amb la ressenya dels cursos del «Core Programme 1998-99» que s'impartiran del gener al juny de 1999 i que s'adrecen especialment als estadístics professionals d'oficines d'estadística oficial en l'àmbit comunitari. Seguidament, es reprodueix l'anunci i el programa complet de les jornades internacionals «Generació d'informació estadística: qualitat i limitacions» (Barcelona, 30 de novembre-1 de desembre de 1998) que s'organitzen amb la col·laboració, entre d'altres institucions, de les tres entitats patrocinadores de *Qüestiió* que participen en la Xarxa Temàtica «Enquestes i Qualitat de la Informació Estadística».

Enric Ripoll, editor executiu

Estadística

QÜESTIIÓ, vol. 22, 3, p. 403-416, 1998

ALMOST UNBIASED RATIO AND PRODUCT-TYPE ESTIMATORS IN SYSTEMATIC SAMPLING

R. SINGH*
Prestige Institute of Management, Mandsaur

H.P. SINGH**
Vikram University

In this paper we have suggested almost unbiased ratio-type and product-type estimators for estimating the population mean $\overline{Y}$ of the study variate y using information on an auxiliary variate x in systematic sampling. The variance expressions of the suggested estimators have been obtained and compared with usual unbiased estimator $\overline{y}^*$, Swain's (1964) ratio estimator $\overline{y}_R^*$ and Shukla's product estimator $\overline{y}_P^*$. It has been shown that the proposed estimators are more efficient than usual unbiased estimator $\overline{y}^*$, ratio estimator $\overline{y}_R^*$ and product estimator $\overline{y}_P^*$. An empirical study is carried out to demonstrate the superiority of the constructed estimators over the estimators $\overline{y}^*$, $\overline{y}_R^*$ and $\overline{y}_P^*$.

Keywords: Almost unbiased ratio and product-type estimators, Auxiliary information, Bias, Variance, Systematic sampling.

AMS Classification: 62 DO5

* Prestige Institute of Management, Mandsaur, M.P., India.
** School of Studies in Statistics, Vikram University, Ujjain-456010, M.P., India.

- Received February 1997.
- Accepted March 1998.

1. INTRODUCTION

Systematic sampling has got the nice feature of selecting the whole sample with just one random start. Apart from its simplicity, which is of considerable importance, this procedure in many situations provides estimators more efficient than simple random sampling and/or stratified random sampling for certain types of population [Cochran (1946, 77), Gautschi (1957), Hajeck (1959)].

Suppose the population consists of N units $U = (U_1, U_2, \ldots, U_N)$ numbered from 1 to N in some order. Unless mention otherwise, we assume $N = nk$, where n and k are positive integers. Thus there will be k samples (clusters) each of size n. We select one sample at random out of k samples and observe the study variate y and auxiliary variate x for each and every unit selected in the sample. Let $(y_{ij}, x_{ij}; \quad i = 1,2,\ldots,k; \; j = 1,2,\ldots,n)$ denote the value of jth unit in the ith sample. The systematic sample means

$$\bar{y}^* = (1/n)\sum_{j=1}^{n} y_{ij}, \quad \bar{x}^* = (1/n)\sum_{j=1}^{n} x_{ij}, \quad (i = 1,2,\ldots,k)$$

are unbiased estimators of the population means $(\overline{Y}, \overline{X})$ of (y,x) respectively. It is assumed that the population mean $\overline{X}$ of the auxiliary variate x is known. Thus the classical ratio and product estimators for $\overline{Y}$ based on a systematic sample $(y_{ij}, x_{ij}; \; i = 1,2,\ldots,k; \quad j = 1,2,\ldots,n)$ of size n, are respectively defined by

$$\bar{y}_R^* = (\bar{y}^*/\bar{x}^*)\overline{X} \tag{1.1}$$

and

$$\bar{y}_P^* = \bar{y}^*(\bar{x}^*/\overline{X}) \tag{1.2}$$

which are respectively due to Swain (1964) and Shukla (1971). The biases and variances of $\bar{y}_R^*$ and $\bar{y}_P^*$ to the first degree of approximation are, respectively, given by

$$B(\bar{y}_R^*) = \frac{(N-1)}{nN}\overline{Y}\{1+(n-1)\rho_x\}(1-k\rho^*)C_x^2 \tag{1.3}$$

$$B(\bar{y}_P^*) = \frac{(N-1)}{nN}\overline{Y}\{1+(n-1)\rho_x\}k\rho^* C_x^2 \tag{1.4}$$

$$\mathrm{Var}(\bar{y}_R^*) = \frac{(N-1)}{nN}\overline{Y}^2\{1+(n-1)\rho_x\}\left[\rho^{*2}C_y^2 + (1-2K\rho^*)C_x^2\right] \tag{1.5}$$

$$\mathrm{Var}(\bar{y}_P^*) = \frac{(N-1)}{nN}\overline{Y}^2\{1+(n-1)\rho_x\}\left[\rho^{*2}C_y^2 + (1-2K\rho^*)C_x^2\right], \tag{1.6}$$

and the variance of usual unbiased estimator $\bar{y}^*$ is given by

$$\mathrm{Var}(\bar{y}^*) = \frac{(N-1)}{nN}\{1+(n-1)\rho_y\}S_y^2, \tag{1.7}$$

where

$$\rho_x = \frac{E(x_{ij} - \overline{X})(x_{ij} - \overline{X})}{E(x_{ij} - \overline{X})^2}, \qquad \rho_y = \frac{E(y_{ij} - \overline{Y})(y_{ij} - \overline{Y})}{E(y_{ij} - \overline{Y})^2}$$

are intraclass correlation between a pair of units within the systematic sample for the auxiliary variate x and study variate y respectively;

$$\rho = \frac{E(y_{ij} - \overline{Y})(x_{ij} - \overline{X})}{\sqrt{\{E(y_{ij} - \overline{Y})^2 E(x_{ij} - \overline{X})^2\}}}$$

is the correlation coefficient between y and x;

$$\rho^* = [\{1 + (n-1)\rho_y\} / \{1 + (n-1)\rho_x\}], \quad K = \rho(C_y/C_x)$$

and (C_y, C_x) are the coefficients of variation of the variates (y,x) respectively.

It is obvious from (1.3) and (1.4) that the estimator $\overline{y}_R^*$ and $\overline{y}_P^*$ are biased. In some situations, bias is disadvantegeous. To keep this in view Kushwaha and Singh (1989) suggested a class of almost unbiased ratio and product-type estimators for population mean $\overline{Y}$ using Jack-knife technique introduced by Quenouille (1956). However, it has been observed that their estimators attained the minimum variance equals to the approximate variance of usual linear regression estimator in systematic sampling, for the well known optimal choice $K^* = \rho^* K$. Later Banarasi et al (1993) suggested a ratio, product and difference estimator in systematic sampling. Banarasi *et al* (1993) estimator attains the minimum variance equals to the approximate variance of usual linear regression estimator for the known value of K. However the optimum estimator in the class of estimators suggested by Banarasi *et al* (1993) is biased. In this paper an effort has been made to propose almost unbiased ratio and product-type estimators which depend only on the well known optimum choice $K^* = \rho^* K$. The value of K can be made known quite accurately either due to past experience or by pilot sample surveys. Various authors including Murthy (1967, p. 325), Reddy (1978) and Srivenkataramana and Tracy (1983) have advocated about the assessment of the value of K^*.

2. THE CLASS OF RATIO-TYPE ESTIMATORS

Let the correlation between y and x be positive. Suppose $d_1 = \overline{y}^*$, $d_2 = \overline{y}^*(\overline{X}/\overline{x})$ and $d_3 = \overline{y}^*(\overline{X}/\overline{x}^*)$ such that $d_i \in D \quad (i = 1,2,3)$, where D denotes the set of all possible ratio-type estimators for estimating population mean $\overline{Y}$. By definition, the class D will consist of all d_r of the form

$$d_r = \sum_{i=1}^{3} w_i d_i \in D \tag{2.1}$$

where

$$\sum_{i=1}^{3} w_i = 1 \quad \text{for} \quad w_i \in R \tag{2.2}$$

and w_i $(i = 1,2,3)$ denotes the constants used for reducing the bias in the class of estimators and R stands for the set of real numbers. Such procedure of estimation of mean has also been discussed by Singh and Singh (1993).

3. VARIANCE

To obtain the variance of d_r to the first degree of approximation, we write

$$e_0 = (\bar{y}^* - \bar{Y})/\bar{Y}, \quad e_1 = (\bar{x}^* - \bar{X})/\bar{X}$$

such that

$$\mathrm{E}(e_0) = \mathrm{E}(e_1) = 0$$

and

$$\begin{aligned}
\mathrm{E}(e_0^2) &= \frac{(N-1)}{nN}\{1+(n-1)\rho_y\}C_y^2 \\
\mathrm{E}(e_1^2) &= \frac{(N-1)}{nN}\{1+(n-1)\rho_x\}C_x^2 \\
\mathrm{E}(e_0 e_1) &= \frac{(N-1)}{nN}\{1+(n-1)\rho_y\}^{1/2}\{1+(n-1)\rho_x\}^{1/2}\rho C_y C_x.
\end{aligned}$$

Expressing (2.1) in terms of e's with (2.2), we have

$$d_r = \bar{Y} + \bar{Y}\left[e_0 - (w_2 + 2w_3)e_1 + O(e^2)\right]. \tag{3.1}$$

Let us choose

$$w_2 + 2w_3 = w \quad \text{(say, another constant)} \tag{3.2}$$

Then, to the first degree of approximation, the variance of d_r is given by

$$\mathrm{Var}(d_r) = \frac{(N-1)}{nN}\bar{Y}^2\{1+(n-1)\rho_x\}\left[\rho^{*2}C_y^2 + w(w-2\rho^* K)C_x^2\right] \tag{3.3}$$

which is minimized for

$$w = \rho^* K = K^* \quad \text{(say)} \tag{3.4}$$

Substitution of (3.4) in (3.3) yields the minimum variance of d_r as

$$\text{(3.5)} \qquad \text{min. Var}(d_r) = \frac{(N-1)}{nN}\{1+(n-1)\rho_y\}(1-\rho^2)S_y^2$$

where $S_y^2 = \dfrac{N}{(N-1)}\text{E}(y_{ij}-\overline{Y})^2$.

We have thus proved the following theorem.

Theorem 3.1. *Up to terms of order* n^{-1},

$$Var(d_r) \geq \frac{(N-1=}{nN}\{1+(n-1)\rho_y\}(1-\rho^2)S_y^2$$

with equality holding if

$$w = \rho^* K = K^*.$$

4. BIAS REDUCTION OF ORDER $O(n^{-1})$

Equation (3.5) shows that the class of estimators attain the minimum variance equal to that of the usual linear regression estimator $\bar{y}_{1r}$ in systematic sampling, defined as

$$\text{(4.1)} \qquad \bar{y}_{1r} = \bar{y}^* + \hat{\beta}^*(\overline{X} - \bar{x}^*)$$

where $\hat{\beta}^*$ is the systematic sample regression coefficient of y on x.

From (3.2) and (3.4), we have

$$\text{(4.2)} \qquad w_2 + 2w_3 = \rho^* K$$

We note from (2.2) and (4.2) that there are three unknown quantities to be determined from only two equations. It is, therefore, not possible to obtain unique values for the constants w_i's, $(i = 1,2,3)$ to be used for bias reduction. To get the unique values for these constants w_i's, $(i = 1,2,3)$, we shall impose the additional linear restriction as

$$\text{(4.3)} \qquad \sum_{i=1}^{3} B(d_i) = 0$$

where $B(d_i)$ stands for the bias in the $i-$th $(i = 1,2,3)$ estimator of the population mean $\overline{Y}$.

Equations (2.2), (4.2) and (4.3) may be expressed as

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ B(d_1) & B(d_2) & B(d_3) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \rho^* K \\ 0 \end{bmatrix} \tag{4.4}$$

It is well known in systematic sampling that

$$B(d_1) = B(\bar{y}^*) = 0 \tag{4.5}$$

and to the first degree of approximation

$$B(d_3) = \frac{(N-1)}{nN} \bar{Y} \{1 + (n-1)\rho_x\} (3 - 2\rho^* K). \tag{4.6}$$

The bias of d_2 is given at (1.3). Thus using (1.3), (4.5) and (4.6) in (4.4), we get the values of w_1, w_2 and w_3 as

$$\left.\begin{aligned} w_1 &= (1 - k\rho^*)^2 \\ w_2 &= (3 - 2k\rho^*) K\rho^* \\ w_3 &= (K\rho^* - 1) K\rho^* \end{aligned}\right\} \tag{4.7}$$

Substitution of (4.7) in (2.1) yields an almost unbiased ratio-type estimator for $\bar{Y}$ as

(4.8)
$$d_{ru} = [(1 - K\rho^*)^2 \bar{y}^* + (3 - 2K\rho^*) K\rho^* \bar{y}^* (\bar{X}/\bar{x}^*) - (1 - K\rho^*) K\rho^* \bar{y}^* (\bar{X}/\bar{x}^*)^2]$$

with the variance

$$\mathrm{Var}(d_{ru}) = \frac{(N-1)}{Nn} \{1 + (n-1)\rho_y\} (1 - \rho^2) S_y^2. \tag{4.9}$$

We have thus proved the following theorem.

Theorem 4.1. *The estimator d_{ru} at (4.8) is an «optimum almost unbiased ratio-type» in the class of ratio-type estimators d_r at (2.1), with the variance given at (4.9).*

It is to be noted that the estimator d_{ru} at (4.8) is unbiased upto terms of order n^{-1}. Same process may be repeated by considering $B(d_i)$, $(i = 1,2,3)$ to terms of order $O(n^{-2})$, to get the unbiased ratio-type estimator to terms of order $O(n^{-2})$ and so on.

In many situations of practical importance $\rho_y \simeq \rho_x$ and known, for instance, see Murthy (1967) and Srivenkataramana and Tracy (1983). Thus $\rho^* \to 1$ and $K \to K$ and hence the values of w_i's, $(i = 1,2,3)$ are given by

$$\left.\begin{aligned} w_1 &= (1-K)^2 \\ w_2 &= (3-2K)K \\ w_3 &= K(K-1) \end{aligned}\right\} \tag{4.10}$$

Putting (4.10) in (2.1) we get the almost unbiased ratio-type estimator for $\overline{Y}$ as

$$d_{ru}^* = \left[(1-K)^2\bar{y}^* + K(3-2K)\bar{y}^*(\overline{X}/\bar{x}^*) + K(K-1)\bar{y}^*(\overline{X}/\bar{x}^*)^2\right] \tag{4.11}$$

with the variance

$$\mathrm{Var}(d_{ru}^*) = \frac{(N-1)}{Nn}\{1+(n-1)\rho_y\}(1-\rho^2)S_y^2. \tag{4.12}$$

It is further remarked that the estimator d_{ru}^* in (4.11) can also be obtained from the estimator d_{ru} with $\rho^* \simeq 1$. The value of K can easily be guessed quite accurately either through pilot survey or experienced gathered in due course of time, for instance, see Murthy (1967, p. 325) and Reddy (1978).

From (1.5), (1.7) and (4.9) we have

$$\mathrm{Var}(\bar{y}^*) - \mathrm{Var}(d_{ru}) = \frac{(N-1)}{nN}\{1+(n-1)\rho_y\}\rho^2 S_y^2 \;>\; 0 \tag{4.13}$$

and

$$\mathrm{Var}(\bar{y}_R^*) - \mathrm{Var}(d_{ru}) = \frac{(N-1)}{nN}\overline{Y}^2\{1+(n-1)\rho_x\}\,C_x^2(1-K\rho^*)^2. \tag{4.14}$$

$$> 0 \quad \text{unless } K\rho^* = 1.$$

We have thus established the following theorems.

Theorem 4.2. *The inequality*

$$Var(d_{ru}) \;<\; Var(\bar{y}^*)$$

always holds good.

Theorem 4.3. *The inequality*

$$Var(d_{ru}) < Var(\bar{y}_R^*)$$

is always true except when $K\rho^* = 1$.

Remark 4.1. It is customary in systematic sampling to arrange the population units such that ρ_y, ρ_x are small, preferably negative, since positive ρ_y, ρ_x inflate sampling variance. Thus it may be possible to assess ρ_y, ρ_x quite accurately for the arrangement used. This together with assessed $K = \rho(C_y/C_x)$ leads to the optimal choice of K^*. □

5. A CLASS OF PRODUCT-TYPE ESTIMATORS

Suppose $d_1^* = \bar{y}^*$, $d_2^* = \bar{y}^*(\bar{x}^*/\overline{X})$ and $d_3^* = \bar{y}^*(\bar{x}^*/\overline{X})^2$ such that $d_i^* \in D^*$ for $i = 1,2,3$, where D^* denotes the sets of all possible product-type estimators for estimating the population mean $\overline{Y}$. By definition, the set D^* will consist of all d_p^* of the form

$$d_p^* = \sum_{i=1}^{3} w_i^* d_i^* \in D^* \tag{5.1}$$

for

$$\sum_{i=1}^{3} w_i^* = 1 \quad \text{and} \quad d_i^* \in R \tag{5.2}$$

where w_i's, $(i = 1,2,3)$ denote the constants used for bias reduction.

As in section 3, the following theorem can easily be proved.

Theorem 5.1. *Up to terms of order* n^{-1},

$$Var(d_p^*) \geqslant \frac{(N-1)}{nN}\{1 + (n-1)\rho_y\}(1-\rho^2)S_y^2$$

with equality holding if

$$w = -\rho^* K = -K^*,$$

where $w^* = (w_2^* + 2w_3^*)$.

Proceeding exactly in the same way as in section 2,3 and 4, we get the values of w_i's, $(i=1,2,3)$ as

$$\left.\begin{aligned} w_1^* &= [1+\rho^* K(1+\rho^* K)] \\ w_2^* &= -\rho^* K(1+2\rho^* K) \\ w_3^* &= \rho^{*2} K^2. \end{aligned}\right\} \tag{5.3}$$

Use of these w_i's, $(i=1,2,3)$ removes the bias of d_p^* upto terms of order n^{-1} at (5.1). Thus the substitution of w_i's, $(i=1,2,3)$ in (5.1) yields an almost unbiased product-type estimator

(5.4)
$$d_{pu}^* = \left[\{1+\rho^* K(1+\rho^* K)\}\bar{y}^* - \rho^* K(1+2K\rho^*)\bar{y}^*(\bar{x}^*/\overline{X}) + \rho^{*2} K^2 \bar{y}^*(\bar{x}^*/\overline{X})^2\right]$$

with the variance

$$Var(d_{pu}^*) = \frac{(N-1)}{Nn}\{1+(n-1)\rho_y\}(1-\rho^2)S_y^2. \tag{5.5}$$

In case $\rho_y \simeq \rho_x \Rightarrow \rho^* = 1$, the expressions in (5.3) reduce to:

$$\left.\begin{aligned} w_1^* &= [1+K(1+K)] \\ w_2^* &= -K(1+2K) \\ w_3^* &= K^2 \end{aligned}\right\} \tag{5.6}$$

and hence the estimator d_{pu}^* at (5.4) takes the form

$$d_{pu}^{**} = \left[\{1+K(1+K)\}\bar{y}^* - K(1+2K)\bar{y}^*(\bar{x}^*/\overline{X}) + K^2\bar{y}^*(\bar{x}^*/\overline{X})^2\right] \tag{5.7}$$

with the variance given by

$$Var(d_{pu}^{**}) = \frac{(N-1)}{Nn}\{1+(n-1)\rho_y\}(1-\rho^2)S_y^2. \tag{5.8}$$

We have thus proved the following theorem.

Theorem 5.2. *The estimator d_{pu}^* at (5.4) is an «optimum almost unbiased product-type» in the class of product-type estimators d_p^* at (5.1), with the variance given at (5.5).*

From (1.6), (1.7) and (5.5) we have

$$(5.9) \qquad Var(\bar{y}^*) - Var(d^*_{pu}) = \frac{(N-1)}{Nn}\{1+(n-1)\rho_y\}\,\rho^2 S_y^2 > 0,$$

and

$$(5.10) \qquad Var(\bar{y}^*_p) - Var(d^*_{pu}) = \frac{(N-1)}{Nn}\{1+(n-1)\rho_x\}\,(1+K\rho^*)^2$$

$$> O \quad \text{provided } K\rho^* \neq -1.$$

We have thus established the following theorems.

Theorem 5.3. *The inequality*

$$Var(d^*_{pu}) < Var(\bar{y}^*)$$

is always true.

Theorem 5.4. *The inequality*

$$Var(d^*_{pu}) < Var(\bar{y}^*_p)$$

always holds good except when $K\rho^* = -1$.

6. EMPIRICAL STUDY

In this section, the relative efficiencies of almost unbiased ratio-type estimator d and product-type estimator d^*_{pu} have been evaluated with the help of live data of Population I and Population II respectively.

POPULATION I

To see the effect of different sample sizes on the approximate relative variances of $\bar{y}^*$, $\bar{y}^*_R$ and d_{ru}, the data on volume of timber of 176 forest strips given in Murthy [1967, p. 131-132] have been considered. The value of intraclass correlation coefficient $\rho_y \simeq \rho_x = \rho_w$ (say) have been given by Murthy [1967, p. 149] and Kushwaha and Singh (1989) for different systematic samples of sizes 4, 8, 16 and 22 strips by enumerating all possible systematic samples after arranging the data in ascending order of strip length. For the systematic sampling to be efficient, the units within the same systematic sample should be as heterogenous as possible with respect to

the characteristic under consideration. As the volume (y) of timber is expected to be related to the strip length (x), the arrangement according to this is likely to be approximately similar to the arrangement according to the volume of timber, the study variable y.

The intraclass correlations ρ_w were computed by the formula

$$\rho_w = 1 - \frac{n}{(n-1)}(\rho_w^2/\rho^2), \tag{6.1}$$

where ρ^2 and ρ_w^2 are the population and within sample variances respectively, given as

$$\sigma^2 = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{X})^2,$$

$$\sigma_w^2 = \frac{1}{k}\sum_{i=1}^{k}\sigma_{wi}^2 = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{x}_i)^2.$$

The intraclass correlation ρ_w generally varies with the sample size and arrangement of units in the population. Summarized data are as follows:

$$N = 176, \quad Y = 282.6136, \quad X = 6.9943, \quad C = 0.3036,$$

$$C = 0.1791, \quad S = 309.8317, \quad r = 0.6722, \quad K = 0.8752.$$

The values of constants w_i's ($i = 1,2,3$), the relative variances of $\overline{y}^*$, $\overline{y}_R^*$ and d_{ru}, and the percent relative efficiencies (PRE's) of $\overline{y}_R^*$ and d_{ru} with respect to $\overline{y}^*$ have been computed for different values of n and displayed in Table 6.1.

Table 6.1. *Showing the relative variances and PRE's of different estimators of* $\overline{Y}$

	Sample size n			
	4	8	16	22
ρ_w	−0.1510	−0.1106	−0.0522	−0.0435
w_1	0.0156	0.0156	0.0156	0.0156
w_2	1.0937	1.0937	1.0937	1.0937
w_3	−0.1092	−0.1092	−0.1092	−0.1092
$RV(\overline{y}^*)$	$4.1281 \cdot 10^{-2}$	$0.8521 \cdot 10^{-2}$	$0.4094 \cdot 10^{-2}$	$0.1187 \cdot 10^{-2}$
$RV(\overline{y}_R^*)$	$2.3007 \cdot 10^{-2}$	$0.4749 \cdot 10^{-2}$	$0.2282 \cdot 10^{-2}$	$0.0662 \cdot 10^{-2}$
$Rv(d_{ru})$	$2.2628 \cdot 10^{-2}$	$0.4671 \cdot 10^{-2}$	$0.2244 \cdot 10^{-2}$	$0.0651 \cdot 10^{-2}$
$RE(\overline{y}_R^*, \overline{y}^*)$	179.43	179.43	179.43	179.43
$RE(d_{ru}, \overline{y}^*)$	182.45	182.45	182.45	182.45

Table 6.1 exhibits that the performance of the proposed almost unbiased ratio-type estimator d_{ru} is better than the usual unbiased estimator $\bar{y}^*$ and Swain's (1964) ratio estimator $\bar{y}_R^*$.

POPULATION II

To see the effect of different sample sizes on the approximate relative variances of $\bar{y}^*$, $\bar{y}_P^*$ and d_{pu}^*, the live data of population of size $N = 16$, on per capita consumption (y) and deflated prices (x) of veal reported in Maddala [1977, p. 98] have been considered. The value of intraclass correlation coefficient $\rho_y \simeq \rho_x = \rho_w$ (say) have been computed for different systematic samples of sizes 2, 4 and 8 by enumerating all possible systematic samples after arranging the data in ascending order of consumption. Summarized data are as follows:

$$N = 16, \quad Y = 7.6375, \quad X = 75.4313, \quad C = 0.0519,$$
$$C = 0.0097, \quad S = -8.8319, \quad r = -0.6823, \quad K = -1.5805.$$

The values of constants w_i's, $(i = 1,2,3)$, the relative variances of $\bar{y}^*$, $\bar{y}_P^*$ and d_{pu}^*, and the percent relative efficiencies (PRE's) of $\bar{y}_P^*$ and d_{pu}^* with respect to $\bar{y}^*$ have been computed for various values of n and shown in Table 6.2.

Table 6.2. *Showing the relative variances and percent relative efficiencies (PRE's) of different estimators of* $\bar{Y}$

	Sample size n		
	2	4	8
ρ_w	−0.7220	−0.2744	−0.0932
w_1^*	1.9175	1.9175	1.9175
w_2^*	−3.4155	−3.4155	−3.4155
w_3^*	2.4980	2.4980	2.4980
$RV(\bar{y}^*)$	$6.7632 \cdot 10^{-3}$	$4.3012 \cdot 10^{-3}$	$2.1141 \cdot 10^{-3}$
$RV(\bar{y}_P^*)$	$4.0316 \cdot 10^{-3}$	$2.5640 \cdot 10^{-3}$	$1.2602 \cdot 10^{-3}$
$RV(d_{pu}^*)$	$3.6147 \cdot 10^{-3}$	$2.2989 \cdot 10^{-3}$	$1.1299 \cdot 10^{-3}$
$RE(\bar{y}_P^*, \bar{y}^*)$	167.76	167.76	167.76
$RE(d_{pu}^*, \bar{y}^*)$	187.10	187.10	187.10

Table 6.2 clearly indicates that the proposed almost unbiased product-type estimator d_{pu}^* is more efficient than usual unbiased estimator $\bar{y}^*$ and Shukla's (1971) product estimator $\bar{y}_P^*$.

7. CONCLUSIONS

It is observed from (1.3) and (1.4) that both the estimators $\bar{y}_R^*$ and $\bar{y}_P^*$ suggested by Swain (1964) and Shukla (1971) respectively, are biased, which is a drawback in some practical situations while the suggested estimators d_{ru} and d_{pu}^* are almost unbiased. We note from (4.13) and (5.9) that the estimators d_{ru} and d_{pu}^* are always better than systematic sample mean estimator $\bar{y}^*$. It is further, observed from (4.14) and (5.10) that the suggested estimator d_{ru} (d_{pu}^*) is always better than $\bar{y}_R^*$ $(\bar{y}_P^*)$ except when $K\rho^* = 1(K\rho^* = -1)$, the case where both the estimators $d_{ru}(d_{pu}^*)$ and $\bar{y}_R^*$ $(\bar{y}_P^*)$ are equally efficient [see Tables 6.1 and 6.2].

Thus we conclude that the suggested estimator d_{ru} (d_{pu}^*) is superior to $\bar{y}_R^*$ $(\bar{y}_P^*)$ according to both the criterion unbiasedness as well as variance.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their useful suggestions on a previous draft of the paper.

REFERENCES

[1] **Banarasi, Kushwaha, S.N.S.** and **Kushwaha, K.S.** (1993). «A class of ratio, product and difference (RPD) estimators in systematic sampling». *Microelectron. Reliab.*, **33, 4**, 455–457.

[2] **Cochran, W.G.** (1946). «Relative efficiency of systematic and stratified random samples for a certain class of population». *Ann. Math. Statist.*, **17**, 164–177.

[3] **Cochran, W.G.** (1977). *Sampling Techniques*, 3rd edition, New York: Wiley.

[4] **Gautschi, W.** (1957). «Some remarks on systematic sampling». *Ann. Math. Statist.*, **28**, 385–394.

[5] **Hajeck, J.** (1959). «Optimum strategy and other problems in probability sampling». *Casopis pro Pestovani Matematiky*, **84**, 387–423.

[6] **Kushwaha, K.S.** and **Singh, H.P.** (1989). «Class of almost unbiased ratio and product estimators in systematic sampling». *Jour. Ind. Soc. Ag. Statistics*, **41, 2**, 193–205.

[7] **Maddala, G.S.** (1977). *Econometrics*. Mc Graw Hills Pub. Co. New York.

[8] **Murthy, M.N.** (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

[9] **Quenouille, M.H.** (1956). «Notes on bias in estimation». *Biometrika*, **43**, 353–360.

[10] **Reddy, V.N.** (1978). «A study on the use of prior knowledge on certain population parameters in estimation». *Sankhyā*, **C, 40**, 29–37.

[11] **Shukla, N.D.** (1971). *Systematic sampling and product method of estimation.* Proceeding of all India Seminar on Demography and Statistics. B.H.U., Varanasi, India.

[12] **Singh, S.** and **Singh, R.** (1993). «A new method: almost separation of bias precipitates in sample surveys». *Jour. Ind. Statist. Assoc.*, **31**, 99–105.

[13] **Srivenkataramana, T.** and **Tracy, D.S.** (1983). «Interchangeability of the ratio and product methods in sample surveys». *Commun. Statist. Theor. Math.*, **12 (18)**, 2143–2150.

[14] **Swain, A.K.P.C.** (1964). «The use of systematic sampling in ratio estimate». *Jour. Ind. Stat. Assoc.*, **2 (213)**, 160–164.

QÜESTIIÓ, vol. 22, 3, p. 417-441, 1998

UNA DISTRIBUCIÓN ASINTÓTICA PARA UN ESTIMADOR NATURAL DEL NÚMERO DE CLUSTERS EN UNA POBLACIÓN

J.J. PRIETO MARTÍNEZ
Universidad Carlos III de Madrid*

Un estimador natural, $\hat{K}$, es propuesto para estimar el número de clusters, K, existentes en una población heterogénea. Una ley límite normal es rigurosamente probada para dicho estimador. La demostración utiliza un método de Holst (1979). Un ejemplo para un conjunto de datos reales y un estudio realizado por simulación es presentado para el estimador propuesto.

A little law for a natural estimator of number of clusters in a population.

Palabras clave: Clusters, población heterogénea, ley límite normal.

Clasificación AMS: 1162G05

*Universidad Carlos III de Madrid. Departamento de Estadística y Econometría. c/Madrid, 126. 28903 Getafe (Madrid).

-Recibido en octubre de 1997.
-Aceptado en abril de 1998.

1. INTRODUCCIÓN

Sea una población constituida por un número desconocido K de clusters. Existe una gran cantidad de trabajos en la literatura estadística sobre los métodos de estimación del número de clusters, pero la mayoría han sido desarrollados en torno a la idea de que las probabilidades de observación de los diferentes clusters son iguales. Ver, por ejemplo, Lewontin y Prout (1956), Darroch (1958), Harris (1968), Johnson y Kotz (1977), Marchand y Schroeck (1982) Darroch y Ratcliff (1980), Holst (1981) y Esty (1985).

Existe un concepto que está muy ligado con el de números de clusters de una población, que es el cubrimiento muestral. Se define como la suma de las probabilidades de los clusters observados en una muestra. En el caso de clusters igualmente probables, el cubrimiento viene dado por el número de clusters observados en una muestra, D, dividido por el número de clusters que constituyen la población, K. Darroch y Ratcliff (1980) utilizaron exactamente la idea del cubrimiento muestral para estimar K.

Ahora bien, considerar la hipótesis de que las probabilidades de los distintos clusters son iguales es, en principio, un caso muy particular y poco frecuente, ya que poblaciones con clusters constituidos por una misma cantidad de elementos es prácticamente imposible. Por ejemplo, no existe una misma cantidad de animales para cada especie en un ecosistema; no se repite con la misma frecuencia cada una de las diferentes palabras que constituyen un texto; no se acuña la misma cantidad de las distintas monedas utilizadas en un país durante un centenario, etc.

La mayoría de los trabajos realizados para poblaciones heterogéneas (es decir, constituidos por clusters no equiprobables) adoptan un enfoque paramétrico. Por ejemplo, Fisher, Corbet y Williams (1943) asumen que para cada cluster, el número de observaciones en la muestra se distribuye según una distribución de Poisson, y el parámetro de dicha distribución se asume que sigue una distribución Gamma. Muchos otros artículos sobre modelos de abundancia de especies en un ecosistema también hacen consideraciones paramétricas. Ver, por ejemplo, McNeil (1973), Engen (1978), Efron y Thisted (1976). Esty (1985) estima el número de clusters en una población heterogénea mediante el concepto de cubrimiento muestral, aunque bajo un modelo paramétrico. Chao y Shen-Ming Lee (1992) propone una técnica de estimación no paramétrica, utilizando también la idea del cubrimiento muestral. Pero hay que subrayar que ninguno de los autores mencionados, como sí hacen algunos autores en el caso equiprobable, estudian cuál es la distribución asintótica del estimador que proponen.

La propuesta de este artículo es justamente el estudio de la distribución asintótica de un estimador para K. Aunque el estimador que aquí se propone es sesgado, lo importante es subrayar la técnica empleada para llegar a dicha distribución, la cual puede ser utilizada próximamente para otros estimadores.

Por tanto, considérese una población cerrada en la cual las observaciones están agrupadas en K clusters. El significado de cerrada hace alusión a que durante el estudio no se producen entradas o salidas de los clusters existentes. A partir de la información obtenida de una muestra aleatoria de tamaño n se propone en el apartado 2 un estimador natural-sesgado para K, $\hat{K}$. El cálculo de su esperanza matemática va a ser importante para cálculos posteriores. Ver el apartado 2.1. Justamente es el apartado 3 el de gran interés. Se estudia la distribución asintótica del estimador propuesto, aplicando un método de Holst (1979). Se prueba que el estimador se distribuye asintóticamente como una normal. En el último apartado se presenta un estudio realizado por simulación para el estimador propuesto. Además se da un ejemplo para un conjunto de datos reales, el cual ha sido aplicado por otros autores. A la vista de los resultados se proponen técnicas de reducción del sesgo del estimador.

2. UN ESTIMADOR NATURAL SESGADO

Asúmase que una muestra aleatoria de tamaño n con reemplazamiento ha sido extraída de la población, la cual está formada por K clusters. La probabilidad de observar el cluster j es $p_j \geqslant 0$, con $j = 1, \ldots, K$ y $\sum_{j=1}^{K} p_j = 1$.

Un estimador natural, sesgado y que bajo-estima K cuando éste es grande con respecto a n es: $\hat{K} = \sum_{j=1}^{K} I_j$, donde

$$I_j = \begin{cases} 1 & \text{si el cluster } j \text{ es observado en la muestra.} \\ 0 & \text{en otro caso.} \end{cases}$$

2.1. Momentos del estimador natural

Son presentados a continuación los operadores esperanza y varianza del estimador $\hat{K}$. El segundo no tiene más interés que saber cuál es la varianza del estimador propuesto. En cambio el primero es de gran importancia por su utilización en el cálculo de la distribución asintótica de $\hat{K}$.

2.1.1. La esperanza de $\hat{K}$

Teorema. *La esperanza del estimador natural $\hat{K}$ viene expresada por*

$$E(\hat{K}) = K - \sum_{j=1}^{K} (1-p_j)^n = K \int_0^{\infty} \left(1 - e^{-x}\right) dF(x),$$

siendo $F(x)$ una función de distribución.

Demostración. Se tiene que:

(1)

$$E(\hat{K}) = E\left(\sum_{j=1}^{K} I_j\right) = \sum_{j=1}^{K} p(I_j = 1) = \sum_{j=1}^{K} [1 - p(I_j = 0)] = K - \sum_{j=1}^{K} (1-p_j)^n.$$

A continuación se demuestra que (1) se puede expresar como:

$$K \int_0^{\infty} \left(1 - e^{-x}\right) dF(x),$$

siendo $F(x)$ una función de distribución, dada en la demostración. ■

El interés que tiene dicha expresión es su utilización en el cálculo de la distribución asintótica.

Considérese la expresión:

$$\frac{\sum_{j=1}^{K} [1-(1-p_j)^n] - \sum_{j=1}^{K} \left[1 - e^{-np_j}\right]}{\sum_{j=1}^{K} \left[1 - e^{-np_j}\right]},$$

donde $0 \leqslant p_j \leqslant 1$ y $\sum_{j=1}^{K} p_j = 1$. Ver el artículo de Harris (1968) donde se utiliza esta expresión, y demostrando que $E(\hat{K}) \cong \sum_{j=1}^{K} \left(1 - e^{-np_j}\right)$. Para ello se aplica el siguiente lema.

Lema. *Si $a_i, b_i > 0, \quad i = 1, 2, \ldots,$ y $\frac{1}{b} = \sup_i \frac{a_i}{b_i}$, entonces $\frac{a}{b} \geqslant \frac{\sum_i a_i}{\sum_i b_i}$. Entonces se*

tiene que:

$$\frac{\sum_{j=1}^{K}\left[1-(1-p_j)^n\right]-\sum_{j=1}^{K}\left[1-e^{-np_j}\right]}{\sum_{j=1}^{K}\left[1-e^{-np_j}\right]} \leqslant \sup_{p_j}\frac{e^{-np_j}-(1-p_j)^n}{1-e^{-np_j}}=\frac{e^{-np}-(1-p)^n}{1-e^{-np}},$$

donde p es justamente una de las p_j's donde se alcanza dicho supremo.

Como $(1-p)^n = e^{n\log(1-p)} = e^{-np-\frac{np^2}{2}}$, entonces

$$\frac{e^{-np}-(1-p)^n}{1-e^{-np}} \leqslant \frac{e^{-np}-e^{-np\frac{np^2}{2}}}{1-e^{-np}} \leqslant \frac{e^{-np}\left(1-e^{-\frac{np^2}{2}}\right)}{1-e^{-np}}.$$

Considérese dos casos posibles para p: cuando $p < 1/\sqrt{n}$ y cuando $p \geqslant 1/\sqrt{n}$ ($n \longrightarrow \infty$ en ambos casos).

Si $p \geqslant \dfrac{1}{\sqrt{n}}$, entonces

$$\frac{e^{-np}-(1-p)^n}{1-e^{np}} \leqslant \frac{e^{-np}}{1-e^{-np}} \leqslant \frac{e^{-\sqrt{n}}}{1-e^{-\sqrt{n}}},$$

que tiende a cero cuando $n \longrightarrow \infty$. Por consiguiente, y teniendo en cuenta la expresión de partida, se tiene que:

$$\sum_{j=1}^{K}\left[1-(1-p_j)^n\right] \cong \sum_{j=1}^{K}\left[1-e^{-np_j}\right]. \tag{2}$$

Si $p < \dfrac{1}{\sqrt{n}}$, considérese

$$(1-p)^n = e^{n\log(1-p)} = e^{-np-\frac{np^2}{2}} \cong e^{-np-\frac{np^2}{2(1-x)^2}} \qquad (0 \leqslant x \leqslant p).$$

Entonces:

$$\frac{e^{-np}-(1-p)^n}{1-e^{-np}} \leqslant \frac{e^{-np}-e^{-np-\frac{np^2}{2(1-(1/\sqrt{n}))^2}}}{1-e^{-np}} = \frac{e^{-np}\left(1-e^{-np-\frac{-n^2p^2}{2(\sqrt{n}-1)^2}}\right)}{1-e^{-np}}$$

Llamando $h_n(p)$ a esta última expresión, calculando la función derivada e igualando a cero, se obtiene:

$$h_n'(p) = \left(\frac{-ne^{-np}\left(1 - e^{\frac{-n^2p^2}{2(\sqrt{n}-1)^2}}\right)}{(1-e^{-np})^2} + \frac{e^{-np}\left(\frac{n^2 2p}{2(\sqrt{n}-1)^2} e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right)}{(1-e^{-np})^2} \right)(1-e^{-np}) -$$

$$- \frac{\left(e^{-np}\left(1 - e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right) n(e^{-np})\right)}{(1-e^{-np})^2} = \frac{-ne^{-np}\left(1 - e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right)(1-e^{-np})}{(1-e^{-np})^2} +$$

$$+ \frac{e^{-np}(1-e^{-np})\frac{n^2 2p}{2(\sqrt{n}-1)^2} e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}}{(1-e^{-np})^2} - \frac{e^{-np}\left(1 - e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right) ne^{-np}}{(1-e^{-np})^2} =$$

$$= \frac{-ne^{-np}\left(1 - e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\right) + e^{-np}(1-e^{-np})\frac{n^2 2p}{2(\sqrt{n}-1)^2} e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}}{(1-e^{-np})^2} = 0.$$

Dividiendo por ne^{-np} y sacando factor común a $e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}$,

$$-1 + e^{\frac{n^2p^2}{2(\sqrt{n}-1)^2}}\left(1 + p\frac{n}{(\sqrt{n}-1)^2}(1-e^{-np})\right) = 0.$$

Tomando logaritmos,

$$\frac{-n^2p^2}{2(\sqrt{n}-1)^2} + \log\left[1 + p\frac{n}{(\sqrt{n}-1)^2}(1-e^{-np})\right] = 0.$$

Como el segundo sumando se puede aproximar por

$$p\frac{n}{(\sqrt{n}-1)^2}(1-e^{-np}),$$

haciendo $\log(1+p) = p - (p^2/2) + (p^3/3) - \cdots \cong p + 0(p)$, entonces

$$\frac{-n^2p^2}{2(\sqrt{n}-1)^2} + p\frac{n}{(\sqrt{n}-1)^2}(1-e^{-np}) \cong 0.$$

Así, $p\frac{n}{(\sqrt{n}-1)^2}\left(1-e^{-np}\right) \cong \frac{n^2p^2}{2(\sqrt{n}-1)^2}$, que es equivalente a decir

$$1-e^{-np} \cong \frac{np}{2}.$$

Al resolver dicha ecuación computacionalmente se obtiene que el máximo de $h_n(p)$ es $p \cong (1,\ 6/n)$. Por consiguiente,

$$h_n\left(\frac{1,6}{n}\right) \cong \frac{e^{-1,6}\left(1-e^{\frac{-2,56}{2(\sqrt{n}-1)^2}}\right)}{1-e^{-1,6}} \longrightarrow 0, \quad \text{cuando} \quad n \longrightarrow \infty.$$

De esta manera se llega a la misma conclusión que (2), es decir,

$$\sum_{j=1}^{K}\left(1-(1-p_j)^n\right) \cong \sum_{j=1}^{K}\left(1-e^{-np_j}\right).$$

Con esto,

$$E(\hat{K} \cong \sum_{j=1}^{K}\left(1-e^{-np_j}\right).$$

Supóngase ahora que cuando K y n tiende a infinito, con las p_j's distintas, la distribución empírica de $np_1, np_2, \ldots, np_k$, definida como

$$F_n(x) = \frac{1}{K}\sum_{j=1}^{K} I(np_j \leqslant x)$$

converge en probabilidad a $F(x)$ sobre $(0, \infty)$. $I(A)$ es la conocida función indicadora. Entonces, se tiene que

$$E(\hat{K}) \cong \sum_{j=1}^{K}\int_0^{\infty}\left(1-e^{-x}\right)dF(x) = K\int_0^{\infty}\left(1-e^{-x}\right)dF(x).$$

Se define X_j como una variable aleatoria que indica el número de veces que se ha observado el cluster j en la muestra, con $j = 1, \ldots, K$; y considérese el siguiente lema de Holst (1979).

Lema. *$P(x_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_k = x_k) = p(Y_1 = x_1, Y_2 = x_2, \ldots, Y_k = x_k \Big/ \sum_{j=1}^{K} Y_j = n\Big)$, donde $\{Y_n\}$ son variables aleatorias independientes de Poisson con media np_j.*

Entonces:

$$E(\hat{K}) = E\left(\sum_{j=1}^{K} I(X_j > 0)\right) = E\left(\sum_{j=1}^{K} I(Y_j > 0)\right) = \sum_{j=1}^{K} Prob(Y_j > 0) =$$

$$= \sum_{j=1}^{K} [1 - ProbI(Y_j = 0)] =$$

$$= \sum_{j=1}^{K} \left[1 - e^{-np_j}\right] \cong \sum_{j=1}^{K} \int_0^{\infty} \left[1 - e^{-np_j}\right] dF(x) =$$

$$= K \int_0^{\infty} \left[1 - e^{-np_j}\right] dF(x),$$

justamente lo que se quería demostrar.

2.1.2. La varianza de $\hat{K}$

Se tiene que:

$$\text{var}(\hat{K}) = E(\hat{K}^2) - E^2(\hat{K}).$$

La esperanza de $\hat{K}$ ha sido calculada anteriormente. Ahora queda por determinar quién es $E(\hat{K}^2)$.

$$E(\hat{K}^2) = \sum_{j=1}^{K}\sum_{l=1}^{K} p(I_j = I_l = 1) = \sum_{j=1}^{K}\sum_{l=1}^{K} \left\{p(I_j = 1) + p(I_l = 1) - p(I_j \text{ o } I_l = 1)\right\}.$$

Como $p(I_j = 1) = 1 - p(I_j = 0) = 1 - (1 - p_j)^n$, y la probabilidad de elegir el cluster j o el cluster l es $p_j + p_l$ $(j \neq l)$, entonces

$$p((I_j = 1) \quad \text{o} \quad (I_l = 1)) = 1 - (1 - (p_j + p_l)^n), \quad j \neq 1.$$

Pero si $l = j$, entonces, $(I_1 = 1) \cup (I_l = 1) = (I_l = 1)$, y

$$p((I_l = 1) \quad \text{o} \quad (I_l = 1)) = 1 - (1 - p_j)^n.$$

Por consiguiente:

$$E(\hat{K}^2) = 2K\sum_{l=1}^{K} p(I_l = 1) - \sum_{\substack{j=1 \\ j \neq l}}^{K}\sum_{l=1}^{K} \left\{p(I_j = 1) + p(I_l = 1)\right\} - \sum_{l=1}^{K} p(I_l = 1) =$$

$$= (2K-1)\left(K-\sum_{l=1}^{K}(1-p_l)^n\right)-K(K-1)+\sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n =$$

$$= K^2(2K-1)\sum_{l=1}^{K}(1-p_l)^n+\sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n.$$

Por tanto,

$$\begin{aligned}\operatorname{var}(\hat{K}) &= K^2-(2K-1)\sum_{l=1}^{K}(1-p_l)^n+\\ &+\sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n-\left[K-\sum_{j=1}^{K}(1-p_j)^n\right]^2=\\ &= K^2-(2K-1)\sum_{l=1}^{K}(1-p_l)^n+\sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n-\\ &-\left[K^2+\left(\sum_{j=1}^{K}(1-p_j)^n\right)^2-2K\sum_{j=1}^{K}(1-p_j)^n\right]=\\ &= \sum_{l=1}^{K}(1-p_l)^n+\sum_{\substack{j=1\\ j\neq l}}^{K}\sum_{l=1}^{K}(1-p_j-p_l)^n-\left\{\sum_{j=1}^{K}(1-p_j)^n\right\}^2.\end{aligned}$$

Hay que notar que dicha expresión coincide por la dada por McNeil (1973).

3. DISTRIBUCIÓN ASINTÓTICA

A continuación se prueba la normalidad asintótica de $\hat{K}$ mediante el método de Holst (1979) (ver también Esty (1985)).

Teorema. *La distribución asintótica de la expresión*

$$K^{-1/2}\left(\hat{K}-E\left(\hat{K}\right)\right)$$

converge a una distribución normal de media cero y varianza σ_1^2, *la cual está dada en la demostración.*

Demostración. Nótese que $\hat{K} = K - N_0$, donde N_0 es una variable aleatoria que indica el número de clusters no observados en la muestra que se define como $N_0 = \sum_{j=1}^{K} I(X_j = 0)$ y $E(N_0)$, utilizando la variable aleatoria indicatriz

$$(3) \qquad Z_j^{(i,n)} = \begin{cases} 1 & \text{si el cluster } j \text{ ocurre } i \text{ veces en la muestra.} \\ 0 & \text{en otro caso.} \end{cases}$$

es igual a $\sum_{j=1}^{K} (1-p_j)^n$. Entonces:

$$\hat{K} - E(\hat{K}) = \hat{K} - \left[K - \sum_{j=1}^{K} (1-p_j)^n\right] = -N_0 + \sum_{j=1}^{K} (1-p_j).$$

Sea:

$$f(X_j) = [I(X_j = 0) - (1-p_j)^n].$$

Se define

$$Z_M = \sum_{j=1}^{K} f(x_j), \quad M < K.$$

Obsérvese que si M es todo K,

$$Z = Z_M = \sum_{j=1}^{K} f(X_j) = \sum_{j=1}^{K} [I(X_j) - (1-p_j)^n] = N_0 - \sum_{j=1}^{K} (1-p_j)^n.$$

Ahora, el problema consiste en encontrar la distribución asintótica de $Z = \sum_{j=1}^{K} f(X_j)$. Para ello se va a seguir el método de Holst (1979), demostrando que la función característica de

$$K^{-1/2}\left(N_0 - \sum_{j=1}^{K} (1-p_j)^n\right)$$

converge a una distribución normal de media cero y varianza σ_1^2, dada en la demostración. Para ello se prueba primero a continuación cuál es la distribución asintótica de $K^{-1/2} Z_M$.

Considérese de nuevo el lema enunciado en el apartado anterior:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = P\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k \Big/ \sum_{j=1}^{K} Y_j = n\right),$$

donde $\{Y_j\}$ son variables aleatorias independientes de Poisson con media np_j. Entonces:

$$E\left\{e^{isK^{-1/2}\sum_{j=1}^{M} f(X_j)}\right\} = E\left\{e^{isK^{-1/2}\sum_{j=1}^{M} f(Y_j)} \Big/ \sum_{j=1}^{K} Y_j = n\right\} \quad (M < K).$$

■

Considérese ahora el siguiente lema de Holst (1979).

Lema. *Si Si* (U, V) *es un vector bidimensional con* U *entero, entonces*

$$E\left(e^{is}/U = n\right) = \frac{1}{2\pi P(U=n)} \int_{-\pi}^{+\pi} E\left(e^{iu(U-n)+isV}\right) du.$$

Entonces,

$$E\left(e^{isK^{-1/2}\sum_{j=1}^{M} f(X_j)} \Big/ \sum_{j=1}^{K} Y_j = n\right) =$$

$$\frac{1}{2\pi P\left(\sum_{j=1}^{K} Y_j = n\right)} \int_{-\pi}^{+\pi} E\left(e^{\sum_{j=1}^{K} (Y_j - np_j) + isK^{-1/2}\sum_{j=1}^{K} f(Y_j)}\right) du.$$

Ahora bien, como $E\left(\sum_{j=1}^{K} Y_j\right) = \sum_{j=1}^{K} np_j = n$ y $n! = e^{-n}\sqrt{2\pi n}n^n$, entonces

$$P\left(\sum_{j=0}^{n} Y_j = n\right) = e^{-n}\frac{n^n}{e^{-n}\sqrt{2\pi n}n^n} = \frac{1}{\sqrt{2\pi n}}.$$

Haciendo el cambio de variable $t = u\sqrt{n}$,

$$E\left(e^{isK^{-1/2}\sum_{j=1}^{M} f(X_j)} \Big/ \sum_{j=1}^{K} Y_j n\right) =$$

$$= \frac{1}{2\pi\frac{1}{\sqrt{2\pi n}}}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}} E\left(e^{itn^{-1/2}\sum_{j=1}^{K}(Y_j-np_j)+isK^{-1/2}\sum_{j=1}^{M}f(Y_j)}\right) n^{-1/2}\,dt =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}} E\left(e^{itn^{-1/2}\sum_{j=1}^{K}(Y_j-np_j)+isK^{-1/2}\sum_{j=1}^{M}f(Y_j)}\right) dt.$$

Sea

$$H_n(s) = \frac{1}{\sqrt{2\pi}}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}} h_{1n}(s,\,t)\,h_{2n}(t)\,dt,$$

donde

$$h_{1n}(s,\,t) = \prod_{j=1}^{M} E\left(e^{itn^{-1/2}(Y_j-np_j)+isK^{-1/2}f(Y_j)}\right)$$

y

$$h_{2n}(s,\,t) = \prod_{j=M+1}^{K} E\left(e^{it(Y_j-np_j)n^{-1/2}}\right).$$

Ahora,

$$h_{2n}(t) = \prod_{j=M+1}^{K}\sum_{m=0}^{\infty} e^{itn^{-1/2}(m-np_j)}\,e^{-np_j}\frac{(np_j)^m}{m!} =$$

$$= \prod_{j=M+1}^{K} e^{-itn^{-1/2}np_j}\sum_{m=0}^{\infty} e^{itn^{-1/2}m}\,e^{-np_j}\frac{(np_j)^m}{m!} =$$

$$= \prod_{j=M+1}^{K} e^{-itn^{-1/2}p_j}\,e^{-np_j}\sum_{m=0}^{\infty}\frac{\left(e^{itn^{-1/2}}np_j\right)^m}{m!} =$$

$$= \prod_{j=M+1}^{K} e^{-itn^{-1/2}p_j}\,e^{-np_j}\,e^{np_j e^{itn^{-1/2}}} =$$

$$= \prod_{j=M+1}^{K} e^{-itn^{-1/2}p_j}\,e^{-np_j\left(e^{itn^{-1/2}}-1\right)}$$

$$= e^{-itn^{-1/2}\sum_{j=M+1}^{K}p_j}\,e^{n\sum_{j=M+1}^{K}p_j\left(e^{itn^{-1/2}}-1\right)}$$

Como $e^{itn^{-1/2}} = 1 + (it/\sqrt{n}) - (t^2/2n) + 0(n)$, entonces:

$$h_{2n}(t) = e^{-itn^{-1/2}\sum_{j=M+1}^{K} p_j} e^{n\sum_{j=M+1}^{K} p_j\left(1+(it/\sqrt{n})-(t^2/2n)-1\right)} =$$

$$= e^{-itn^{-1/2}\sum_{j=M+1}^{K} p_j} e^{n\sum_{j=M+1}^{K} p_j itn^{-1/2} - n\sum_{j=M+1}^{K} p_j (t^2/2n)} =$$

$$= e^{-\sum_{j=M+1}^{K} p_j (t^2/2n)}$$

Considerando $h_{1n}(s, t)$, se tiene:

$$h_{1n}(s, t) = \prod_{j=1}^{M} E\left(e^{itn^{-1/2}(Y_j - np_j) + isK^{-1/2} f(Y_j)}\right) = \prod_{j=1}^{M} g_j(s, t),$$

donde

$$g_j(s, t) = E\left(e^{itn^{-1/2}(Y_j - np_j) + isK^{-1/2} f(Y_j)}\right) =$$

$$= E\left(e^{itn^{-1/2}(Y_j - np_j) + isK^{-1/2} I(Y_j = 0)}\right)\left(e^{-isK^{-1/2}(1-p_j)^n}\right).$$

Ahora bien, el primer factor es igual a:

$$E\left(e^{itn^{-1/2}(Y_j - np_j) + isK^{-1/2} I(Y_j = 0)}\right) =$$

$$= e^{-itn^{-1/2} p_j + isK^{-1/2}} e^{-np_j} +$$

$$+ e^{-itn^{-1/2} p_j} e^{-np_j}\left\{\sum_{R=0}^{\infty} \frac{(np_j)^R \left(e^{itn^{-1/2}}\right)^R}{R!} - 1\right\} =$$

$$= e^{-np_j} e^{-itn^{-1/2} p_j}\left(e^{isK^{-1/2}} - 1\right) + e^{-itn^{-1/2} p_j} e^{-np_j} e^{np_j e^{itn^{-1/2}}}$$

El primer sumando se puede poner de la forma:

$$e^{-np_j}\left(1 - itn^{-1/2} p_j - \frac{nt^2 p_j^2}{2}\right)\left(isK^{-1/2} - \frac{s^2}{2K}\right) =$$

$$= e^{-np_j}\left\{isK^{-1/2} - \frac{s^2}{2K} + \frac{ts}{n^{1/2} K^{1/2}}(np_j)\right\} + 0\left(K^{-1}\right).$$

Y el segundo sumando se puede escribir recordando el desarrollo de $h_{2n}(t)$, como $e^{(-p_j/2)t^2}$.

Por consiguiente,

$$E\left(e^{itn^{-1/2}(Y_j-np_j)+isK^{-1/2}I(Y_j=0)}\right) =$$

$$= e^{-np_j}\left\{isK^{-1/2} - \frac{s^2}{2K} + \frac{ts}{n^{1/2}K^{1/2}}(np_j)\right\} + e^{(-p_j/2)t^2} + 0\left(K^{-1}\right).$$

El segundo factor, teniendo en cuenta que

$$(1-p_j)^n = e^{\log(1-p_j)^n} \cong e^{-np_j},$$

$$e^{-isK^{-1/2}(1-p_j)^n} = 1 - isK^{-1/2}e^{-np_j} - \frac{s^2}{2K}e^{-2np_j} + 0\left(K^{-1}\right).$$

De esta forma,

$$\begin{aligned} g_j(s,t) &= e^{-(p_j/2)t^2}\times \\ &\times \left\{1 + e^{(p_j/2)t^2}e^{-np_j}\left(isK^{-1/2} - \frac{s^2}{2K} + \frac{ts}{n^{1/2}K^{1/2}}(np_j)\right\}\times\right. \\ &\times \left\{1 - isK^{-1/2}e^{-np_j} - \frac{s^2}{2K}e^{-2np_j} + 0\left(K^{-1}\right)\right\} \end{aligned}$$

Haciendo $e^{(p_j/2)t^2} = 1 + 0(p_j t^2)$,

$$\begin{aligned} g_j(s,t) &= e^{-(p_j/2)t^2}\left\{1 - isK^{-1/2}e^{-np_j} - \frac{s^2}{2K}e^{-2np_j} + e^{-np_j}iKsK^{-1/2} - \right. \\ &- \left.\frac{s^2}{2K}e^{-np_j} + \frac{ts}{n^{1/2}K^{1/2}}e^{-np_j}(np_j) + e^{-2np_j}s^2K^{-1}\right\} = \\ &= e^{-(p_j/2)t^2}\left\{1 + \frac{s^2}{2K}e^{-2np_j} - \frac{s^2}{2K}e^{-np_j} + \frac{ts(np_j)}{n^{1/2}K^{1/2}}e^{-np_j}\right\} \end{aligned}$$

Por tanto:

$$\begin{aligned} \prod_{j=1}^{M} g_j(s,t) &= e^{-t^2/2\sum_{j=1}^{M} p_j}\times \\ &\times \prod_{j=1}^{M}\left\{1 - \frac{s^2}{2K}e^{-np_j} + \frac{s^2}{2K}e^{-2np_j} + \frac{ts(np_j)}{n^{1/2}K^{1/2}}e^{-np_j}\right\} \cong e^{-t^2/2\sum_{j=1}^{M} p_j}\times \end{aligned}$$

$$\times \quad e^{-\sum_{j=1}^{M}\left[(s^2/2K)\,e^{-np_j}-(s^2/2K)\,e^{-2np_j}\right]+\sum_{j=1}^{M}\left(tsnp_j/n^{1/2}K^{1/2}\right)e^{-np_j}}$$

Entonces:

$$H_n(s) = \frac{1}{\sqrt{2\pi}}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}} e^{-t^2/2}\left(\sum_{j=1}^{M}p_j+\sum_{j=M+1}^{K}p_j\right) e^{\sum_{j}^{M}\left(tsnp_j/n^{1/2}K^{1/2}\right)e^{-np_j}}\,dt\times$$

$$\times \quad e^{-s^2/2\left\{(1/K)\sum_{j=1}^{M}e^{-np_j}-(1/K)\sum_{j=1}^{M}e^{-2np_j}\right\}} =$$

$$= \int_{-\pi n^{1/2}}^{+\pi n^{1/2}} \frac{1}{\sqrt{2\pi}} e^{-1/2\left\{t-\frac{\sum_{j}^{M}np_je^{-np_j}}{n^{1/2}K^{1/2}}s\right\}^2} dt\times$$

$$\times \quad e^{-(s^2/2)\left\{\sum_{j=1}^{M}(np_j)\,e^{-np_j}\right\}^2/nK} \quad e^{(-s^2/2)\sum_{j=1}^{M}\left\{(1/K)e^{-np_j}-(1/K)e^{-2np_j}\right\}}$$

Si el número de clusters en la población y el tamaño de la muestra son sumamente grandes, tomando el límite de $H_n(s)$ cuando n y K tienden a infinito, se tiene:

$$\lim_{n,K\longrightarrow\infty} H_n(s) = \lim_{n,K\longrightarrow\infty}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}} \frac{1}{\sqrt{2\pi}} e^{-1/2\left\{t-\frac{\sum_{j}^{M}np_je^{-np_j}}{n^{1/2}K^{1/2}}s\right\}^2} dt\times$$

$$\times \lim_{n,K\longrightarrow\infty}\left\{ e^{-(s^2/2)\left\{\sum_{j=1}^{M}(np_j)\,e^{-np_j}\right\}^2/nK}\times\right.$$

$$\left.\times \quad e^{-(s^2/2)\sum_{j=1}^{M}\left\{(1/K)e^{-np_j}-(1/K)e^{-2np_j}\right\}}\right\}$$

Aplicando el teorema de convergencia dominada:

$$\lim_{n,K\longrightarrow\infty}\int_{-\pi n^{1/2}}^{+\pi n^{1/2}}\frac{1}{\sqrt{2\pi}}e^{-1/2\left\{t-\frac{\sum_j^M np_je^{-np_j}}{n^{1/2}K^{1/2}}s\right\}^2}dt=$$

$$=\int_{-\pi n^{1/2}}^{+\pi n^{1/2}}\lim_{n,K\longrightarrow\infty}\left\{\frac{1}{\sqrt{2\pi}}e^{-1/2\left\{t-\frac{\sum_j^M np_je^{-np_j}}{n^{1/2}K^{1/2}}s\right\}^2}\right\}dt=$$

$$=\int_{-\pi n^{1/2}}^{+\pi n^{1/2}}\frac{1}{\sqrt{2\pi}}e^{-1/2t^2}dt=1.$$

Por consiguiente:

$$\lim_{n,K\longrightarrow\infty}H_n(s)=$$

$$=\lim_{n,K\longrightarrow\infty}e^{(-s^2/2)\left\{\sum_{j=1}^M(1/K)e^{-np_j}\sum_{j=1}^M(1/K)e^{-2np_j}-\sum_{j=1}^M(1/nK)\left((np_j)e^{-np_j}\right)^2\right\}}=$$

$$=e^{-(s^2/2)}\lim_{n,K\longrightarrow\infty}\left\{\sum_{j=1}^M(1/K)e^{-np_j}\sum_{j=1}^M(1/K)e^{-2np_j}-\sum_{j=1}^M(1/nK)\left((np_j)e^{-np_j}\right)^2\right\}$$

Así, $K^{-1/2}Z_M$ se distribuye asintóticamente como una $N(0,\sigma_M^2)$, con

$$\sigma_M^2=\lim_{n,K\longrightarrow\infty}\left\{\frac{1}{K}\sum_{j=1}^M e^{-np_j}-\frac{1}{K}\sum_{j=1}^M e^{-2np_j}-\frac{1}{nK}\left\{\sum_{j=1}^M np_je^{-np_j}\right\}^2\right\}.$$

Ahora bien, $Z=Z_M+Z_{MC}$, siendo $Z_{MC}=\sum_{j=M+1}^K f(X_j)$. Se puede probar exactamente igual que $K^{-1/2}Z_{MC}$ se distribuye asintóticamente según una $N(0,\sigma_{MC}^2)$, donde

$$\sigma_{MC}^2=\lim_{n,K\longrightarrow\infty}\left\{\frac{1}{K}\sum_{j=M+1}^K e^{-np_j}-\frac{1}{K_j}\sum_{j=M+1}^K e^{-2np_j}-\frac{1}{nK}\left\{\sum_{j=M+1}^K np_je^{-np_j}\right\}^2\right\}.$$

Así,

$$K^{-1/2}\left(N_0 - \sum_{j=1}^{K}(1-p_j)^n\right) \longrightarrow N(0, \sigma_1^2),$$

donde $\sigma_1^2 = \sigma_M^2 + \sigma_{MC}^2$, tal que

$$\sigma_1^2 = \lim_{n,K \to \infty} \left\{ \frac{1}{K}\sum_{j=1}^{K} e^{-np_j} - \frac{1}{K}\sum_{j=1}^{K} e^{-2np_j} - \frac{1}{nK}\left\{\sum_{j=1}^{K} np_j e^{-np_j}\right\}^2 \right\}$$

y, por consiguiente,

$$\left(\sum_{j=1}^{K}(1-p_j)^n - N_0\right) \longrightarrow N(0, \sigma_1^2(\hat{K})),$$

donde $\sigma_1^2(\hat{K}) = K\sigma_1^2$.

Supóngase ahora, como se hizo en el apartado anterior, que la distribución empírica de $np_1, np_2, \ldots, np_k$, definida como

$$F_n(x) = \frac{1}{K}\sum_{j=1}^{K} I(np_j \leqslant x)$$

converge débilmente a $F(x)$ sobre $(0, \infty)$. Entonces,

$$\begin{aligned}
\sigma_1^2 &= \frac{1}{K}\sum_{j=1}^{K}\int_0^\infty e^{-x}\,dF(x) - \frac{1}{K}\sum_{j=1}^{K}\int_0^\infty e^{-2x}\,dF(x) - \\
&- \left(K\int_0^\infty (Kx)\,dF(x)\right)^{-1}\left(\sum_{j=1}^{K}\int_0^\infty (xe^{-x})\,dF(x)\right)^2 = \\
&= \int_0^\infty \left(e^{-x}(1-e^{-x})\right)dF(x) - \left(K\int_0^\infty (Kx)\,dF(x)\right)^{-1}\left(\sum_{j=1}^{K}\int_0^\infty (xe^{-x})\,dF(x)\right)^2,
\end{aligned}$$

ya que:

$$n = E\left(\sum_{j=1}^{K} X_j\right) = E\left(\sum_{j=1}^{K} Y_j\right) = \sum_{j=1}^{K} np_j \cong \sum_{j=1}^{K}\int_0^\infty x dF(x) = K\int_0^\infty x dF(x).$$

Una cota superior aparente para la varianza asintótica de $\hat{K}$ es:

$$\begin{aligned}
\sigma_1^2(\hat{K}) &= K\left(\int_0^\infty (1-e^{-x})\,dF(x) - \left(K\int_0^\infty (Kx)\,dF(x)\right)^{-1}\left(K\int_0^\infty (xe^{-x})\,dF(x)\right)^2\right) = \\
&= K\int_0^\infty (1-e^{-x})\,dF(x) - \left(K\int_0^\infty (Kx)\,dF(x)\right)^{-1} K\left(K\int_0^\infty (xe^{-x})\,dF(x)\right)^2.
\end{aligned}$$

Sea la variable N_i que se define como el número de clusters que se observan exactamente i veces en la muestra. A partir de la variable aleatoria indicatriz definida en (3),

$$E(N_i) = \sum_{j=1}^{K} \binom{n}{i} p_j^i (1-p_j)^{n-i} \cong \sum_{j=1}^{K} e^{-np_j} \frac{(np_j)^i}{i!} \cong \sum_{j=1}^{K} \int_0^{\infty} \left(e^{-x} \frac{x}{i!}\right) dF(x),$$

tal que

$$i!E(N_i) \cong K \int_0^{\infty} \left(e^{-x} \frac{x}{i!}\right) dF(x),$$

y como $E(\hat{K}) \cong K \int_0^{\infty} (1-e^{-x})\, dF(x)$, se obtiene que,

$$\sigma_1^2(\hat{K}) = E(\hat{K}) - \frac{E^2(N_i)}{E\left(\sum_{j=1}^{K} X_j\right)}.$$

Por consiguiente, un estimador de $\sigma_1^2(\hat{K})$ es

$$\hat{\sigma}_1^2(\hat{K}) = \hat{k} - (n_1^2/n),$$

al reemplazar las esperanzas por los valores observados.

4. RESULTADOS NUMÉRICOS

Antes de presentar los resultados numéricos es necesario indicar que, aunque el objetivo principal de este artículo es la propuesta de presentar una técnica de obtención de una distribución asintótica para un estimador del número de clusters en una población, es necesario y conveniente resaltar cómo corregir $\hat{K}$. Existen técnicas jackknife y bootstrap que corrigen y ajustan el estimador según el sesgo cometido. En Prieto (1998, a y b) se presentan justamente estas técnicas como reducción y corrección del sesgo. También han sido utilizadas por Burnham y Overton (1979) para estimar el número de individuos en una población, o por Heltshe y Forrester (1983) para estimar el número de especies en un ecosistema.

Ejemplo 1

Este ejemplo es propuesto por Fisher, Corbet y Williams (1943). 1421 especies fueron cogidas en una trampa - muestra en la localidad de Rothamsted y clasificadas

por especies. Los datos se resumen en la siguiente tabla.

i	1	2	3	4	5	6	7	≥ 8
n_i	35	11	15	14	10	11	5	139

Fisher, Corbet y Williams estimaron el número de especies en 261.9. $\hat{K} = 240$, y la varianza es

$$\hat{\sigma}_1^2(\hat{K}) = \hat{k} - (n_1^2/n) = 240 - \frac{35^2}{1421} \cong 240.$$

Obsérvese que n debe tender a infinito cuando K es sumamente grande. A continuación se presenta un ejemplo por simulación para ver la eficiencia de $\hat{K}$ según las probabilidades de observacion de cada cluster.

Ejemplo 2

Entonces, para comprobar la eficacia del estimador propuesto $\hat{K}$ se ha evaluado mediante métodos computacionales por simulación. La evaluación de $\hat{K}$ ha sido llevada a cabo simulando una muestra aleatoria o bien de tamaño 50 o bien de tamaño 100 de una población de 200 clusters.

Las probabilidades de observar los diferentes clusters han sido consideradas pertenecientes al intervalo [0.0020; 0.01]. Se han considerado 6 casos posibles. En el primer caso se ha considerado las probabilidades iguales. En el segundo los primeros 100 clusters tienen probabilidades 0.004 de ser observados y los 100 siguientes 0.006. Los siguientes casos se van considerando poblaciones más heterogéneas. Cada caso se ha simulado 50 veces y se han tomado el promedio de los resultados.

Los resultados obtenidos indican que:

- $\hat{K}$ siempre bajo estima el valor de K, siendo muy sesgado. Técnicas para corregir el estimador han sido ya mencionadas.
- Para cualquier población, el sesgo cometido por $\hat{K}$ cuando $n = 50$ es siempre mayor que cuando $n = 100$.
- El sesgo de cada estimador aumenta a medida que la población es más heterogénea.
- La varianza de $\hat{K}$ cuando $n = 50$ es más pequeña que cuando $n = 100$. A medida que la población es más heterogénea, la varianza es ligeramente mayor.

Tabla 1

Casos	n	p_j	$\hat{K}$	$\hat{\sigma}_1^2$	$E(\hat{K}-K)$	E.C.M.	$V(\hat{K})$
1	50	$p_j = 0.005$ $j = 1-200$	119	48.91	-81	6593.14	32.14
	100		132	53.41	-68	4662.09	38.09
2	50	$p_j = 0.004$ $j = 1-100$ $p_j = 0.006$ $j = 101-200$	103	44.12	-97	9447.01	38.01
	100		118	48.43	-82	6768.80	44.80
3	50	$p_j = 0.0035$ $j = 1-90$ $p_j = 0.0045$ $j = 91-180$ $p_j = 0.014$ $j = 181-200$	94	39.10	-106	11274.10	38.14
	100		107	43.41	-93	8694.04	45.04
4	50	$p_j = 0.01$ $j = 1-10$ $p_j = 0.004$ $j = 11-100$	82	41.27	-118	13967.20	43.24
	100	$p_j = 0.003$ $j = 101-190$ $p_j = 0.023$ $j = 191-200$	97	46.92	-103	10657.10	48.12

Continúa

Tabla 1 (cont.)

Casos	n	p_j	$\hat{K}$	$\hat{\sigma}_1^2$	$E(\hat{K}-K)$	E.C.M.	$V(\hat{K})$
5	50	$p_j = 0.0035$	67	44.12	−133	17735.20	46.28
		$j = 1-50$					
		$p_j = 0.006$					
		$j = 151-100$					
		$p_j = 0.002$					
		$j = 101-125$					
	100	$p_j = 0.009$	85	48.43	−115	13274.20	49.21
		$j = 126-150$					
		$p_j = 0.005$					
		$j = 151-200$					
6	50	$p_j = 0.006$	56	48.91	−144	20785.20	49.24
		$j = 1-25$					
		$p_j = 0.0025$					
		$j = 26-50$					
		$p_j = 0.009$					
		$j = 51-75$					
	100	$p_j = 0.008$	69	53.41	−131	17216.30	55.39
		$j = 76-100$					
		$p_j = 0.001$					
		$j = 101-125$					
		$p_j = 0.002$					
		$j = 126-150$					
		$p_j = 0.005$					
		$J = 151-175$					
		$p_j = 0.004$					
		$j = 176-200$					

AGRADECIMIENTOS

Quiero agradecer a Dr. Anne Chao (Institute of Statistics in National Tsing Hua University, Hsin-Chu, Taiwan) su importante colaboración en este artículo.

REFERENCIAS

[1] **Burnham, K.P.** y **Overton, W.S.** (1978). «Estimation of the size of a closed population when capture probabilities vary among animals». *Biometrika*, **63, 3**, 625-633.

[2] **Chao, A.** y **Shen-Ming Lee.** (1992). «Estimating the number of classes via sample coverage». *Journal of the American Statistical Association*, **87, N-417**, 211-217.

[3] **Darroch, J.N.** (1958). «The multiple recapture census I: Estimation of a closed population». *Biometrika*, **45**, 343-359.

[4] **Darroch, J.N.** y **Ratcliff, D.** (1980). «A note on capture-recapture estimation». *Biometrika*, **40**, 343-359.

[5] **Efron, B.** y **Thisted, R.** (1976). «Estimating the number of unseen species: How many words did Shakespeare know?». *Biometrika*, **63**, 435-447.

[6] **Engen, S.** (1978). *Stochastic Abundance Models*. London: Chapman-Hall.

[7] **Esty, W.W.** (1983). «A normal limit law for a nonparametric estimator of the coverage of a random sample». *The Annals of Statistic*, **11**, 905-912.

[8] **Esty, W.W.** (1985). «Estimation of the number of classes in a population and the coverage of a sample». *Mathematical Scientist*, **10**, 41-50.

[9] **Fisher, R.A., Corbet, A.S.** y **Williams, C.B.** (1943). «The relation between the number of species and the number of individuals in a random sample of a animal population». *Journal of Animal Ecology*, **12**, 42-58.

[10] **Harris, B.** (1968). «Statistical inference in the classical occupancy problem unbiased estimation of the number of classes». *Journal of the American Statistical Association*, **63**, 837-847.

[11] **Heltshe, J.F.** y **Forrester, N.E.** (1983). «Estimating species richess using the jackknife procedure». *Biometrics*, **39**, 1-11.

[12] **Holst, L.** (1979). «A unified approach to limit theorems for urn models». *Journal of Applied Probability*, **16**, 154-162.

[13] **Holst, L.** (1981). «Some asymptotic results for incomplete multinomial or poisson samples». *Scandinavian Journal of Statistic*, **8**, 243-246.

[14] **Johnson, N.L.** y **Kotz, S.** (1977). *Urn models and their applications: an approach to modern discrete probability theory.* New York: Yohn Wiley.

[15] **Lewontin, R.C.** y **Prout, T.** (1956). «Estimation of the number of different classes in a population». *Biometrics*, **12**, 211-223.

[16] **Marchand, J.P.** y **Schroeck, P.E.** (1982). «On the estimation of the number of equally likely classes in a population». *Communications in a Statistic, Part A-Theory and Methods*, **11**, 1139-1146.

[17] **McNeil, D.R.** (1973). «Estimating an author's vocabulary». *Journal of the American Statistical Association*, **68, N-341**, 92-96.

[18] **Prieto M., J.J.** (1998, a). «¿Cuántos clusters hay en una población?». *Qüestiió*, **22, 1.**

[19] **Prieto M.. J.J.** (1998, b). «Estimación del número de clusters en una población aplicando el jackknife generalizado». *Qüestiió*, **2.**

[20] **Robbins, H.** (1968). «Estimating the total probability of the unobserved outcomes of an experiment». *Annals of mathematical Statistics*, **39**, 256-257.

ENGLISH SUMMARY

A LITTLE LAW FOR A NATURAL ESTIMATOR OF NUMBER OF CLUSTERS IN A POPULATION

J.J. PRIETO MARTÍNEZ
Universidad Carlos III de Madrid*

A natural estimator, $\hat{K}$, is propousted to estimate the number of clusters, K, that there are in a heterogeneous population. A limit law normal is rigorously proved for $\hat{K}$. The proof utilizes a method of Holst (1979). The performance of the estimator is investigated by means of Monte Carlo experiments and it is applied to one real data examples.

Keywords: Clusters, heterogeneous population, limit law normal.

AMS Classification: 1162G05

*Universidad Carlos III de Madrid. Departamento de Estadística y Econometría. c/Madrid, 126. 28903 Getafe (Madrid).

-Received October 1997.
-Accepted April 1998.

Assume that a random sample is drawn from a population with unknown number K o clusters. Denote p_j the probability that any observation belong to the jth cluster, $j = 1, \ldots, K$, $\sum_{j=1}^{K} p_j = 1$.

A natural estimator, with bias, and its value belonging to $(0, K)$ is

$$\hat{K} = \sum_{j=1}^{K} I_j,$$

where

$$I_j = \begin{cases} 1 & \text{if the cluster } j \text{ is observed in the sample.} \\ 0 & \text{otherwise.} \end{cases}$$

The expectation of $\hat{K}$ is

$$E(\hat{K}) = K - \sum_{j=1}^{K} (1 - p_j)^n = K \int_0^{\infty} \left(1 - e^{-x}\right) dF(x),$$

where $F(x)$ is a distribution function. This result is applied in the proof to obtain the asymptotic distribution. A lemma of Holst (1979) is very important to proof it. The variance of $\hat{K}$ is:

$$\operatorname{var}(\hat{K}) = \sum_{j=1}^{K} (1 - p_j) + \sum_{\substack{j=1 l=1 \\ j \neq l}}^{K} \sum_{}^{K} (1 - p_j - p_l)^n - \left(\sum_{j=1}^{K} (1 - p_j)^n \right)^2,$$

which is similar obtained by McNeil (1973).

The principal goal is the limiting normality of the estimator biased $\hat{K}$, which is derived using the method of Holst (1979). The result is important because this method can be used to obtain the asymptotic distribution of a estimator.

If assume that $n \longrightarrow \infty$, then

$$K^{-1/2} \left(\hat{K} - E(\hat{K})\right) \longrightarrow N(0, \sigma_1^2),$$

where σ_1^2 is given in the proof.

The performance of the proposed estimator is investigated by means of Monte Carlo simulations. Some alternatives are showed to correcting and adjusting $\hat{K}$ for its estimated bias.

QÜESTIIÓ, vol. 22, 3, p. 443-468, 1998

ALGUNAS SOLUCIONES APROXIMADAS PARA DISEÑOS SPLIT-PLOT CON MATRICES DE COVARIANZA ARBITRARIAS

G. VALLEJO SECO
J.R. ESCUDERO GARCÍA
Universidad de Oviedo*

El presente trabajo revisa con cierto detalle diversos tipos de análisis para diseños split-plot que carecen del mismo número de unidades experimentales dentro de cada grupo y en los que se incumple con el supuesto de esfericidad multimuestral. Específicamente, adaptando el enfoque multivariado de aproximar los grados de libertad desarrollado por Johansen (1980) y el procedimiento de la aproximación general mejorada corregida basado en Huynh (1980) se muestra cómo obtener análisis robustos y poderosos a la hora de probar los efectos principales y la interacción, así como hipótesis de comparaciones múltiples relacionadas con estos efectos, tanto si se cuenta con una simple variable dependiente asociada con cada una de las medidas repetidas como si se cuenta con más de una.

Some approximate solutions for split-plot designs with arbitrary covariance matrices.

Palabras clave: Diseños multivariados de medidas repetidas, esfericidad multimuestral, medias ponderadas y noponderadas, procedimientos robustos y poderosos.

Clasificación AMS: 62K10, 62J1

*Departamento de Psicología. Universidad de Oviedo. Plaza Feijoo, s/n. 33003 Oviedo. e-mail:gvallejo@sci.cpd.uniovi.es.

-Recibido en diciembre de 1996.
-Aceptado en mayo de 1998.

1. INTRODUCCIÓN

En las Ciencias del Comportamiento y campos metodológicos afines destaca el impulso que a lo largo de los últimos años ha recibido el empleo de los diseños de medidas repetidas y, sobremanera, los diseños split-plot o diseños que implican anidar las unidades experimentales respecto a los niveles de una o más variables y cruzarlos respecto a los niveles de otra u otras. Bajo el cumplimiento de los supuestos de normalidad conjunta multivariada, homogeneidad de las matrices de dispersión e igualdad de las varianzas correspondientes a las diferencias entre las diferentes medidas repetidas, dichos diseños han sido analizados tradicionalmente por medio del modelo mixto univariado de Scheffé (1956). A su vez, cuando se incumple alguno de los dos últimos supuestos, e inclusive ambos a la vez, lo usual es hacer uso, tanto del enfoque univariado con los grados de libertad corregidos como del enfoque multivariado; sobre todo, teniendo en cuenta que si el diseño está equilibrado, ambos enfoques son relativamente robustos a la violación de los supuestos referidos (Huynh, 1978; Keselman y Keselman, 1990; Keselman, Lix y Keselman, 1993; Rogan, Keselman y Mendoza, 1978). En estos casos, la elección entre el enfoque univariado con los grados de libertad ajustados o el enfoque multivariado descansa, amén de otras cuestiones, en consideraciones de potencia (Davidson, 1972). Por ejemplo, asumiendo que disponemos de moderados tamaños de muestra, ligeras desviaciones en el patrón de esfericidad conllevan una mayor potencia del enfoque univariado que del multivariado correspondiente; por el contrario, la situación se va invirtiendo paulatinamente a medida que nos desviamos del patrón de esfericidad requerido.

Sin embargo, como ha sido puesto de relieve en diversos trabajos de simulación, ambos enfoques analíticos dejan de ser robustos cuando el diseño carece del equilibrio adecuado y las matrices de varianzas y covarianzas poblacionales no son combinables por ser heterogéneas. En concreto, Keselman (1993) encuentra que el procedimiento de ajustar los grados de libertad por el valor de la desviación del patrón de esfericidad produce un sesgo considerable a la hora de verificar las hipótesis correspondientes a los efectos principales de la parte intra del diseño. Por su parte, Timm (1975) nos dice que otro tanto ocurre con la significación de dichos efectos cuando se contrastan mediante el enfoque multivariado. Finalmente, tampoco conviene pasar por alto que diversos autores (Belli, 1988; Huynh y Feldt, 1976; Keselman *et al.*, 1993) han descubierto que las tasas de error de Tipo I concernientes a la interacción también se encuentran seriamente distorsionadas bajo ambos enfoques. Así pues, cuando el diseño de medidas parcialmente repetidas no esté adecuadamente balanceado y el supuesto de homogeneidad de las matrices de dispersión se incumpla, conviene ser muy cuidadosos tanto con las inferencias realizadas en relación con la igualdad de las respuestas dadas a lo largo del tiempo, como con las inferencias referidas a la hipótesis que especifica que las diferencias entre los grupos no dependen de las condiciones de observación consideradas; ya que como se ha resaltado, éstas descansan, en buena

medida, en el cumplimiento del susodicho supuesto. Como destaca Wilcox (1987), las pruebas paramétricas convencionales basadas en las distribuciones t o F pueden resultar seriamente afectadas cuando la homogeneidad de las varianzas no se satisface, sobre todo, cuando el número de unidades experimentales difiere de un grupo a otro. En estos casos, las pruebas estadísticas clásicas se pueden convertir, bien en excesivamente conservadoras, o bien en excesivamente liberales, con tasas de error de Tipo I por debajo del 0.01 ó por encima del 0.70 para un nivel de significación nominal del 0.05 (Keselman y Keselman, 1988; Milligan, Wong y Thompson, 1987; Wilcox, Charlin y Thompson, 1986). Para abordar el problema especificado se han intentado varios remedios, no obstante, la solución más habitual consiste en hacer uso del arreglo Behrens-Fisher y en aproximar los grados de libertad mediante el procedimiento de Welch (1938, 1947, 1951). Dicha solución fue aplicada inicialmente por Welch a los diseños de dos o más grupos al azar, y muchos años después, Algina y Olejnik (1994) la han extendido a los diseños factoriales, mientras que Keselman, Carriere y Lix (1993, 1995) lo han implementado en el ámbito de los diseños de medidas repetidas y en el ámbito de los diseños factoriales no ortogonales; todo ello al amparo de la versión generalizada que del procedimiento de Welch y de la posterior mejora de James (1951,1954) realizó Johansen en 1980. En el contexto de los diseños de medidas repetidas otro de los remedios disponibles, y a juicio de Algina (1994) y Algina y Oshima (1994, 1995) muy recomendable, es el enfoque de la aproximación general de Huynh (1978), así como subsiguientes modificaciones efectuadas en el mismo.

2. EL ENFOQUE DE WELCH-JAMES

Por lo que se refiere a la primera solución o enfoque de la aproximación multivariada de los grados de libertad desarrollado por Johansen (1980), en adelante enfoque de Welch-James, resaltar que los descubrimientos empíricos encontrados a lo largo de los últimos años en el contexto de los diseños univariados ponen de relieve la robustez de este enfoque a la hora de probar si existen diferencias entre las medias de dos o más poblaciones cuando las varianzas son heterogéneas y el número de unidades experimentales difiere de un grupo a otro. Más aún, la potencia de este enfoque resulta comparable a la potencia encontrada tras aplicar las pruebas clásicas de t o F cuando el supuesto de homogeneidad de las varianzas es satisfecho (Roth, 1983; Brown y Forsythe, 1974; Hsiung y Olejnik, 1994b, Wilcox *et al.*, 1986) y, por supuesto, mucho más poderoso que éstas cuando, tanto el tamaño de los grupos como el tamaño de las varianzas difiere sustancialmente.

No obstante, conviene matizar que el enfoque Welch-James no mantiene inalterable su robustez a lo largo de todas las situaciones. En concreto, cuando existe una correlación negativa entre el tamaño de los grupos y el tamaño de las varianzas (la varianza más

grande acontece en el grupo de menor tamaño, o viceversa) y los datos presentan un sesgo apreciable, el enfoque bajo consideración no controla la tasa de error de Tipo I al nivel nominal a. Cuando se registra más de una variable dependiente o a las unidades experimentales se las mide en más de una ocasión, la tasa de error de Tipo I también puede llegar a ser considerable si la razón entre el tamaño de los grupos y el número de variables dependientes o bien el número de medidas repetidas es pequeña, las matrices de varianzas y covarianzas son heterogéneas y los datos no se acomodan a una distribución normal multivariada. En las demás situaciones el enfoque de Welch-James proporciona un adecuado grado de control de la tasa de error de Tipo I; posteriormente efectuaremos algún comentario adicional sobre esta cuestión. Con todo, no conviene dejar sin aclarar que el resto de las técnicas, tanto si son de naturaleza paramétrica como de naturaleza no paramétrica, participan de las mismas debilidades y ofrecen un grado de generalización mucho más restringido que el enfoque bajo consideración.

Seguidamente, se expone la modelización del enfoque Welch-James en el ámbito de los diseños split-plot. Para ello, inicialmente, nos ceñiremos al caso en que las $q(k, \ldots, q)$ respuestas recogidas a partir de las $n(i, \ldots, n)$ unidades muestrales independientes estén agrupadas de acuerdo con los $p(j, \ldots, p)$ niveles de una variable de clasificación. Para una situación como la descrita, el modelo lineal general con N unidades experimentales puede escribirse como sigue:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}$$

donde $\boldsymbol{Y}$ es una matriz de respuestas de orden $N \times q$, $\boldsymbol{X}$ es la matriz de diseño de rango pleno de orden $N \times p$, $\boldsymbol{B}$ es una matriz de parámetros no aleatorios de orden $p \times q$ y $\boldsymbol{E}$ es una matriz de errores aleatorios de orden $N \times q$. Si denotamos por $\varepsilon_i'(\varepsilon_{ij1}, \ldots, \varepsilon_{ijk})$ el vector de errores aleatorios correspondiente a la unidad ith, se asume que cada subvector de errores es $N_q(\mathbf{0}, \boldsymbol{\Sigma}_j)$. El hecho de que la forma de $\boldsymbol{\Sigma}_j$ dependa de j indica que todos los vectores de errores aleatorios no tienen la misma matriz de varianzas y covarianzas, $\boldsymbol{\Sigma}$, lo que implica que las matrices no son combinables.

En términos sustantivos las hipótesis de interés del diseño split-plot o diseño de medidas parcialmente repetidas son las siguientes:

1. ¿Existe interacción entre las variables entre e intra del diseño?
2. ¿Difieren entre sí los diferentes grupos de tratamiento?
3. ¿Tienen todas las respuestas el mismo efecto?

Para contrastar dichas hipótesis haremos dos cosas: Por un lado, utilizaremos la prueba estadística de Welch-James, la cual, de acuerdo con Johansen (1980), puede expresarse como sigue:

$$\boldsymbol{T}_{W-J} = (\boldsymbol{R}\bar{\boldsymbol{y}})' (\boldsymbol{RPR}')^{-1} (\boldsymbol{R}\bar{\boldsymbol{y}})$$

donde $\bar{y}$ es un vector de orden $pq \times 1$ obtenido tras concatenar verticalmente las medias de $\bar{y}_j$, $\boldsymbol{R}$ ($\boldsymbol{R} = \boldsymbol{C}' \otimes \boldsymbol{A}'$) es una matriz de contrastes cuyo orden depende de la hipótesis que estemos probando y $\boldsymbol{P}$ es una matriz diagonal de bloques de orden $pq \times pq$ con el jth bloque igual a $\hat{\Sigma}_j/n_j$ $(\hat{\Sigma}_1/n_1, \ldots, \hat{\Sigma}_p/n_p)$. Según Johansen (1980, p.86), el estadístico T_{W-J} dividido por una constante, c, se distribuye aproximadamente como el estadístico F con grados de libertad ν_1 (rango de la matriz $\boldsymbol{R}$) y $\nu_2 = \nu_1 (\nu_1 + 2)/3A$. Las constantes c y A valen, respectivamente

$$c = \nu_1 + 2A - \frac{(6A)}{(\nu_1 + 2)}$$

y

$$A = \frac{1}{2} \sum_{j=1}^{p} \left\{ tr \left[\boldsymbol{P}\boldsymbol{R}' (\boldsymbol{R}\boldsymbol{P}\boldsymbol{R}')^{-1} \boldsymbol{R}\boldsymbol{Q}_j \right]^2 + \left[tr (\boldsymbol{P}\boldsymbol{R}' (\boldsymbol{R}\boldsymbol{P}\boldsymbol{R}')^{-1} \boldsymbol{R}\boldsymbol{Q}_j \right]^2 \right\} \Big/ (n_j - 1)$$

donde tr denota el operador traza y $\boldsymbol{Q}_j$ es una matriz diagonal de bloques de orden $pq \times pq$ con el jth bloque igual a una matriz de identidad de orden $q \times q$ y el resto ceros.

Por otro lado, expresaremos todas las hipótesis expuestas anteriormente mediante una adecuada elección de la matriz de contrastes $\boldsymbol{R}$ y también en términos de los parámetros de la matriz $\boldsymbol{B}$ que sigue:

$$\boldsymbol{B} = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1q} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2q} \\ \mu_{p1} & \mu_{p2} & \cdots & \mu_{pq} \end{pmatrix}$$

Concretando aún más, la hipótesis nula que afirma que las diferencias entre los niveles de la variable de tratamiento no dependen de los niveles de la variable intra considerados, viene dada por

$$H_{0_1}: \quad \begin{pmatrix} \mu_{11} - \mu_{12} \\ \cdots \\ \mu_{1,q-1} - \mu_{1q} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{p1} - \mu_{p2} \\ \cdots \\ \mu_{p,q1} - \mu_{pq} \end{pmatrix}$$

o, bien

$$H_{0_1}: \quad \boldsymbol{R}\boldsymbol{\mu} = \boldsymbol{0}$$

donde $\boldsymbol{R} = \boldsymbol{C}' \otimes \boldsymbol{A}'$ es una matriz de orden $(p-1)(q-1) \times pq$, $\boldsymbol{C}'$ es una matriz de coeficientes de orden $(p-1) \times p$ que determina los elementos de μ a incluir en la hipótesis nula, $\boldsymbol{A}$ es una matriz de orden $q \times (q-1)$ propia de las situaciones multivariadas que permite generar hipótesis entre los diferentes parámetros de respuesta, μ

es un vector de parámetros de orden $pq \times 1$ y $\mathbf{0}$ es un vector nulo de cuyo orden es $pq \times 1$. Las matrices $\boldsymbol{C}'$ y $\boldsymbol{A}$ adoptan la forma que sigue:

$$\boldsymbol{C}' = \begin{pmatrix} 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & -1 \end{pmatrix}; \qquad \boldsymbol{A} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 \end{pmatrix}$$

La H_{0_1} se rechaza al nivel α si

$$T_{w-J}/c > F_{\alpha};\nu_1, \nu_2 \qquad (\nu_1 = p-1 \times q-1)$$

donde $T_{w-J}/c \sim F(\nu_1, \nu_2)$.

De resultar la interacción significativa la hipótesis nula de ausencia de diferencias entre los grupos viene dada por

$$H_{0_2}: \quad \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1q} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2q} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{p1} \\ \mu_{p2} \\ \vdots \\ \mu_{pq} \end{pmatrix}$$

o, simplemente

$$H_{0_2}: \quad \boldsymbol{R\mu} = \mathbf{0}$$

donde $\boldsymbol{R} = \boldsymbol{C}' \otimes \boldsymbol{A}'$ es una matriz de orden $(p-1)q \times pq$, $\boldsymbol{C}'$ tiene la misma forma que en el caso anterior y $\boldsymbol{A}$ es una matriz de identidad de orden $q \times q$.

En ausencia de interacción la hipótesis nula de igualdad de los grupos se reduce a:

$$H_{0_2}^*: \quad \sum_{k=1}^{q} \mu_{1k}/q = \sum_{k=1}^{q} \mu_{2k}/q = \cdots = \sum_{k=1}^{q} \mu_{pq}/q$$

o, bien

$$H_{0_2}^*: \quad \boldsymbol{R\mu} = \mathbf{0}$$

donde $\boldsymbol{R} = \boldsymbol{C}' \otimes \boldsymbol{a}'$ es una matriz de orden $(p-1) \times pq$, $\boldsymbol{C}'$ mantiene la forma aludida y $\boldsymbol{a}$ es un vector de unos de orden $q \times 1$. En ambos casos la hipótesis nula referida a los niveles de la variable *entre* se rechaza al nivel α si

$$T_{w-J}/c > F_{\alpha}; \quad \nu_1, \nu_2$$

Finalmente, la hipótesis nula de igualdad de las respuestas u ocasiones de observación también puede abordarse teniendo en cuenta la significación o no de la interacción. En el primer caso, dicha hipótesis puede expresarse como:

$$H_{0_3}: \quad \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{pmatrix} = \begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{pmatrix} = \cdots \begin{pmatrix} \mu_{1q} \\ \mu_{2q} \\ \vdots \\ \mu_{pq} \end{pmatrix}$$

o, simplemente

$$H_0: \quad \boldsymbol{R}\boldsymbol{\mu} = \mathbf{0}$$

donde $\boldsymbol{R} = \boldsymbol{C}' \otimes \boldsymbol{A}'$ es una matriz de orden $p(q-1) \times pq$, $\boldsymbol{C}'$ es una matriz de identidad de orden $p \times p$ y $\boldsymbol{A}$ es una matriz de contrastes de orden $q \times (q-1)$.

En el segundo caso, esto es, de no existir interacción la hipótesis de igualdad de las respuestas viene dada por

$$H^*_{0_3}: \quad \sum_{j=1}^{p} \mu_{j1}/p = \sum_{j=1}^{p} \mu_{j2}/p = \cdots = \sum_{j=1}^{p} \mu_{jq}/p$$

o, bien

$$H^*_{0_3}: \quad \boldsymbol{R}\boldsymbol{\mu} = \mathbf{0}$$

donde $\boldsymbol{R} = \boldsymbol{C}' \otimes \boldsymbol{A}'$ es una matriz de orden $(q-1) \times pq$, $\boldsymbol{c}'$ es un vector de unos de orden $1 \times p$ y $\boldsymbol{A}$ es una matriz de contrastes de orden $q \times (q-1)$. Bajo cualquiera de las dos situaciones especificadas, la hipótesis nula referida a la igualdad de los niveles de la variable intra se rechaza al nivel α si

$$T_{w-J}/c > F_\alpha; \quad \nu_1, \nu_2$$

donde $T_{w-J}/c \sim F(\nu_1, \nu_2)$.

Antes de concluir con este apartado conviene advertir que el enfoque descrito sólo permite probar la hipótesis nula de igualdad de las respuestas concerniente a medias no ponderadas. El investigador interesado en probar la hipótesis nula referida a medias ponderadas puede utilizar la solución multivariada que del problema Behrens-Fisher nos ofrecen Nel y van der Merwe (1986).

3. EXTENSIÓN DEL ENFOQUE WELCH-JAMES A SITUACIONES MULTIVARIADAS

Con las modificaciones oportunas de las matrices $\boldsymbol{C}'$ y $\boldsymbol{A}$ puede manejarse cualquier tipo de diseño que implique medidas repetidas y que utilice una sola variable depen-

diente. Sin embargo, a nadie se le escapa que existen bastantes situaciones que si bien se acomodan a un diseño de medidas repetidas, éstas son esencialmente de naturaleza multivariada.

Considérese el diseño multivariado de medidas parcialmente repetidas esquematizado en la figura 1 tomada de Vallejo y Menéndez (1997), en el cual hay un factor de agrupamiento p ($j = 1, \ldots, p$), un factor de observaciones repetidas q ($k = 1, \ldots, q$) y las respuestas dadas por el ith sujeto a lo largo de q medidas repetidas bajo cada una de las r variables de medida. De este modo, bajo esta disposición las q primeras columnas corresponden a la primera variable dependiente, las q segundas a la segunda variable dependiente y las q–ésimas la rth variable dependiente.

Trat.	Suj.	VD_1	VD_2	VD_r
A_j	S_i	$\mathbf{y}'_{ij1} = \left[y_{ij1}^{(1)}\, y_{ij1}^{(2)} \cdots y_{ij1}^{(q)}\right]$	$\mathbf{y}'_{ij2} = \left[y_{ij2}^{(1)}\, y_{ij2}^{(2)} \cdots y_{ij2}^{(q)}\right]$	$\cdots\ \mathbf{y}'_{ijr} = \left[y_{ijr}^{(1)}\, y_{ijr}^{(2)} \cdots y_{ijr}^{(q)}\right]$

Figura 1. Disposición de los datos del diseño multivariado de medidas parcialmente repetidas.

Para un diseño como el esquematizado el modelo lineal multivariado con N unidades experimentales puede ser escrito como sigue:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}$$

donde $\boldsymbol{Y} = N \times qr$ es la matriz de respuestas, $\boldsymbol{B} = p \times qr$ es la matriz de parámetros, $\boldsymbol{X} = N \times p$ es la matriz de diseño de rango pleno y $\boldsymbol{E} = N \times qr$ es la matriz de errores aleatorios. Si denotamos por $\varepsilon'_i = 1 \times qr$ el vector de errores aleatorios correspondiente al sujeto ith, es asumido que:

$$\boldsymbol{\varepsilon}'_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_j)$$

donde la matriz de covarianzas $\Sigma_j = qr \times qr$ es una matriz definida positiva. El hecho de que la forma Σ dependa de la población de la que se hayan extraído las n_j unidades, supone que no todos los vectores de errores aleatorios ε tienen la misma matriz Σ y, por ende, son matrices heteroscedásticas.

Bajo esta nueva situación, la construcción de la matriz $\boldsymbol{R}$ para probar las hipótesis del modelo es algo más complicada; no obstante, tal cometido se puede manejar relativamente bien si se prosigue con una notación similar a la utilizada en el caso univariado. Para ilustrar lo dicho se comienza probando la hipótesis nula que especifica que las diferencias entre los grupos no dependen de las condiciones de observación consideradas en el conjunto de las r variables dependientes examinadas simultáneamente,

esto es

$$H_{0_1}: \quad \begin{pmatrix} \mu_{11}^{(1)} - \mu_{12}^{(1)} \\ \vdots \\ \mu_{1q-1}^{(2)} - \mu_{1q}^{(2)} \\ \vdots \\ \mu_{1q-1}^{(r)} - \mu_{1q}^{(r)} \end{pmatrix} = \begin{pmatrix} \mu_{21}^{(1)} - \mu_{22}^{(1)} \\ \vdots \\ \mu_{2q-1}^{(2)} - \mu_{2q}^{(2)} \\ \vdots \\ \mu_{2q-1}^{(r)} - \mu_{2q}^{(r)} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{p1}^{(1)} - \mu_{p2}^{(1)} \\ \vdots \\ \mu_{pq-1}^{(2)} - \mu_{pq}^{(2)} \\ \vdots \\ \mu_{pq-1}^{(r)} - \mu_{pq}^{(r)} \end{pmatrix}$$

o, bien

$$H_{0_1}: \quad \boldsymbol{R}\boldsymbol{\mu} = \mathbf{0}$$

donde $\boldsymbol{R} = \boldsymbol{C}' \otimes (\boldsymbol{A}_q \otimes \boldsymbol{I}_r)'$ es una matriz de orden $(p-1)(q-1)r \times pqr$, $\boldsymbol{C}'$ es una matriz de coeficientes de orden $(p-1) \times p$ que determina los elementos de μ a incluir en la hipótesis nula, $\boldsymbol{A}$ es una matriz de coeficientes de orden $q \times (q-1)$ que nos permite generar hipótesis entre los diferentes niveles de la variable intra, $\boldsymbol{I}_r$ es una matriz de identidad de orden $r \times r$ y μ es un vector de parámetros de orden $pqr \times 1$. La H_{0_1} se rechaza al nivel α si

$$T_{w-J}/c > F_\alpha; \quad \nu_1, \nu_2 \qquad [\nu_1 = (p-1)(q-1)r]$$

donde $T_{w-J}/c \sim F(\nu_1, \nu_2)$. El cálculo de los grados de libertad correspondientes al error se efectúa igual que en el caso univariado descrito con anterioridad.

La ausencia de diferencias entre los grupos puede ser abordada desde un doble punto de vista en función de que la interacción sea o no sea significación. En el primer caso,

$$H_{0_2}: \quad \begin{pmatrix} \mu_{11}^{(1)} \\ \mu_{12}^{(1)} \\ \vdots \\ \mu_{1q}^{(1)} \\ \vdots \\ \mu_{1q}^{(r)} \end{pmatrix} = \begin{pmatrix} \mu_{21}^{(1)} \\ \mu_{22}^{(1)} \\ \vdots \\ \mu_{2q}^{(1)} \\ \vdots \\ \mu_{2q}^{(r)} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{p1}^{(1)} \\ \mu_{p2}^{(1)} \\ \vdots \\ \mu_{pq}^{(1)} \\ \vdots \\ \mu_{pq}^{(r)} \end{pmatrix}$$

o, alternativamente

$$H_{0_2}: \quad \boldsymbol{R}\boldsymbol{\mu} = \mathbf{0}$$

donde $\boldsymbol{R} = \boldsymbol{C}' \otimes (\boldsymbol{A}_q \otimes \boldsymbol{I}_r)$ es una matriz de orden $(p-1)qr \times pqr$, $\boldsymbol{C}'$ tiene la misma forma que en el caso anterior, $\boldsymbol{A}_q$ e $\boldsymbol{I}_r$ son matrices de identidad de orden $q \times q$ y $r \times r$, respectivamente.

A su vez, en el segundo caso, esto es, de no existir interacción

$$H_{0_2}^*: \quad \begin{pmatrix} \sum_{k=1}^{q} \mu_{1k}^{(1)}/q \\ \sum_{k=1}^{q} \mu_{1k}^{(2)}/q \\ \vdots \\ \sum_{k=1}^{q} \mu_{1k}^{(r)}/q \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{q} \mu_{2k}^{(1)}/q \\ \sum_{k=1}^{q} \mu_{2k}^{(2)}/q \\ \vdots \\ \sum_{k=1}^{q} \mu_{2k}^{(r)}/q \end{pmatrix} = \cdots = \begin{pmatrix} \sum_{k=1}^{q} \mu_{pk}^{(1)}/q \\ \sum_{k=1}^{q} \mu_{pk}^{(2)}/q \\ \vdots \\ \sum_{k=1}^{q} \mu_{pk}^{(r)}/q \end{pmatrix}$$

o, alternativamente

$$H_{0_2}^*: \quad \boldsymbol{R\mu} = \mathbf{0}$$

donde $\boldsymbol{R} = \boldsymbol{C}' \otimes (\boldsymbol{I}_r \otimes \boldsymbol{a}')$ es una matriz de orden $(p-1)r \times pqr$, $\boldsymbol{C}'$ e $\boldsymbol{I}_r$ adoptan la misma forma que en el caso de arriba, pero $\boldsymbol{a}$ es un vector de unos de orden $q \times 1$. En ambos casos la H_0 se rechaza al nivel α si

$$T_{w-J}/c > F_\alpha; \quad \nu_1, \nu_2$$

donde $T_{w-J}/c \sim F(\nu_1, \nu_2)$.

Por último, presentamos la prueba de la hipótesis nula multivariada de igualdad de las ocasiones de observación o intervalos temporales. De nuevo dicha hipótesis puede abordarse desde una doble perspectiva en función de que la interacción resulte o no significativa. En el primer caso,

$$H_{0_3}: \quad \begin{pmatrix} \mu_{11}^{(1)} \\ \mu_{21}^{(1)} \\ \vdots \\ \mu_{p1}^{(1)} \\ \vdots \\ \mu_{21}^{(r)} \\ \vdots \\ \mu_{p1}^{(r)} \end{pmatrix} = \begin{pmatrix} \mu_{12}^{(1)} \\ \mu_{22}^{(1)} \\ \vdots \\ \mu_{p2}^{(1)} \\ \vdots \\ \mu_{22}^{(r)} \\ \vdots \\ \mu_{p2}^{(r)} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{1q}^{(1)} \\ \mu_{2q}^{(1)} \\ \vdots \\ \mu_{pq}^{(1)} \\ \vdots \\ \mu_{2q}^{(r)} \\ \vdots \\ \mu_{pq}^{(r)} \end{pmatrix}$$

o, simplemente

$$H_{0_3}: \quad \boldsymbol{R\mu} = \mathbf{0}$$

donde $\boldsymbol{R} = \boldsymbol{C}' \otimes (\boldsymbol{I}_r \otimes \boldsymbol{A}')$ es una matriz de orden $p(q-1)r \times pqr$, $\boldsymbol{C}'$ es una matriz de identidad de orden $p \times p$, $\boldsymbol{A}$ es una matriz de contrastes de orden $q \times (q-1)$ e $\boldsymbol{I}_r$ es una matriz de identidad de orden $r \times r$.

En el segundo caso, esto es, si no existe interacción entre los niveles de la variable *entre* y las condiciones de observación en el conjunto de las variables dependientes consideradas simultáneamente, tenemos

$$H_{0_3}^{*}: \quad \begin{pmatrix} \sum_{j=1}^{p} \mu_{j1}^{(1)}/p \\ \sum_{j=1}^{p} \mu_{j1}^{(2)}/p \\ \vdots \\ \sum_{j=1}^{p} \mu_{j1}^{(r)}/p \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{p} \mu_{j2}^{(1)}/p \\ \sum_{j=1}^{p} \mu_{j2}^{(2)}/p \\ \vdots \\ \sum_{j=1}^{p} \mu_{j2}^{(r)}/p \end{pmatrix} = \cdots = \begin{pmatrix} \sum_{j=1}^{p} \mu_{jq}^{(1)}/p \\ \sum_{j=1}^{p} \mu_{jq}^{(2)}/p \\ \vdots \\ \sum_{j=1}^{p} \mu_{jq}^{(r)}/p \end{pmatrix}$$

o, simplemente

$$H_{0_3}^{*}: \quad \boldsymbol{R\mu} = \mathbf{0}$$

donde $\boldsymbol{R} = \boldsymbol{c}' \otimes (\boldsymbol{A}' \otimes \boldsymbol{I}_r)$ es una matriz de orden $(q-1)r \times pqr$, $\boldsymbol{c}'$ es un vector de unos de dimensión $1 \times p$, $\boldsymbol{A}_q$ es una matriz de contrastes de orden $q \times (q-1)$ e $\boldsymbol{I}_r$ es una matriz de identidad de dimensión $r \times r$. En ambos casos la hipótesis nula se rechaza al nivel α si

$$T_{w-J}/c > F_{\alpha}; \quad \nu_1, \nu_2$$

donde $T_{w-J}/c \sim F(\nu_1, \nu_2)$.

4. EL ENFOQUE DE LA APROXIMACIÓN GENERAL

Como ha sido anticipado, el procedimiento multivariado de ajustar los grados de libertad que acabamos de describir también adolece de limitaciones, y de hecho debiera ser utilizado con ciertas precauciones cuando el tamaño de muestra es reducido, sobremanera, en lo referido a la significación de la interacción. En estas circunstancias los trabajos de simulación llevados a cabo por Algina (1994), Algima y Oshima (1994, 1995) y Coombs y Algina (1996), han puesto de relieve que tanto el procedimiento de la aproximación general mejorada (IGA), como el de la corrección de la aproximación general mejorada (CIGA), proporcionan un mejor control de la tasa de error Tipo I para los efectos principales que el procedimiento Welch-James. Los procedimientos *IGA* y *CIGA* representan tan sólo variaciones del enfoque de la aproximación general (*GA*) desarrollado por Huynh (1978) para contrastar medias ponderadas en los diseños de medidas parcialmente repetidas. Por lo que se refiere a la interacción, los resultados empíricos de Algina y Oshima (1994) y de Keselman, Carriere y Lix (1993) conducen a conclusiones similares para ambos enfoques. Keselman *et al.* (1993) encuentran que el procedimiento Welch-James tiende a ser liberal cuando el tamaño de muestra

es reducido. En concreto, y a tenor de los hallazgos de estos autores, para que la tasa de error Tipo I de los efectos principales no exceda de 0.075 cuando el valor nominal vale 0.05, se requiere que la razón entre n_j y $q-1$ sea de 2 cuando $q=4$ y de 1.7 cuando $q=8$; para la interacción la razón requerida es de 3 cuando $q=4$ y de 3.9 cuando $q=8$. Dichas razones están referidas a datos normales, para datos muestreados desde distribuciones sesgadas las razones deben ser ligeramente superiores. A su vez, Algina y Oshima (1994) encuentran que cuando la razón entre n_j y $q-1$ es pequeña y la matriz de dispersión se aproxima al patrón de esfericidad, el enfoque *GA* tiende a ser conservador; por el contrario, dicho enfoque se vuelve ligeramente liberal a medida que la matriz de dispersión se aleja del patrón de esfericidad. Los efectos referidos se acentúan cuando se trata comprobar la significación de la interacción y conforme los datos se apartan de la normalidad.

El enfoque de la aproximación general desarrollado por Huynh (1978), permite extender el trabajo iniciado por Box en 1954 a todas aquellas situaciones en las cuales, además de incumplirse el supuesto de esfericidad, también se incumple el supuesto de homogeneidad de las matrices de varianzas y covarianzas. De acuerdo con el enfoque *AG* propuesto por Huynh, la hipótesis nula referida a la ausencia de diferencias entre las medias ponderadas de la variable intra se rechaza si $F > \hat{b}F_{\alpha}; \hat{h}', \hat{h}$. Donde $\hat{b}$, $\hat{h}'$ y $\hat{h}$ son constantes generalmente desconocidas, pero que pueden estimarse a partir de la matriz de varianzas-covarianzas muestral, $\hat{\boldsymbol{\Sigma}}$, como sigue:

$$\hat{b} = \frac{N - p\,\mathrm{tr}\left(\boldsymbol{C}^{*'}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^{*}\right)}{\sum_{j=1}^{p}(n_j-1)\,\mathrm{tr}\left(\boldsymbol{C}^{*'}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^{*}\right)}$$

$$\hat{h}' = \frac{\left(\mathrm{tr}\left(\boldsymbol{C}^{*'}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^{*}\right)\right)^2}{\mathrm{tr}\left(\boldsymbol{C}^{*'}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^{*}\right)^2}$$

y

$$\hat{h} = \frac{\left[\sum_{j=1}^{p}(n_j-1)\,\mathrm{tr}\left(\boldsymbol{C}^{*'}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{C}^{*}\right)\right]^2}{\sum_{j=1}^{p}(n_j-1)\,\mathrm{tr}\left(\boldsymbol{C}^{*'}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{C}^{*}\right)^2}$$

donde $\boldsymbol{C}^{*'}$ es una matriz de contrastes ortonormalizados de orden $(q-1)\times q$ y $\hat{\boldsymbol{\Sigma}} = N^{-1}\sum_{j=1}^{p} n_j\hat{\boldsymbol{\Sigma}}_j$.

A su vez, la nulidad de la otra hipótesis afectada por la ausencia de esfericidad multimuestral, esto es, la referida a la ausencia de interacción entre las variables

del diseño de medidas parcialmente repetidas, se rechaza de acuerdo con el enfoque *GA* si $F > \hat{c}F_{\alpha}$; $\hat{h}''$, $\hat{h}$. En este caso, la obtención de los estadísticos $\hat{c}$ y $\hat{h}''$ se complica ligeramente; no obstante, también pueden obtenerse a partir de las matrices de varianzas-covarianzas muestrales como sigue:

$$\hat{c} = \frac{N-p}{p-1} \frac{\operatorname{tr} \boldsymbol{G}\hat{\boldsymbol{S}}}{\sum_{j=1}^{p} (n_j - 1) \operatorname{tr} \left(\boldsymbol{C}^{*\prime} \hat{\boldsymbol{\Sigma}}_j \boldsymbol{C}^{*} \right)}$$

y

$$\hat{h}'' = \frac{(\operatorname{tr} \boldsymbol{G}\hat{\boldsymbol{S}})^2}{\operatorname{tr} (\boldsymbol{G}\hat{\boldsymbol{S}})^2}$$

donde es una matriz diagonal de bloques de orden $pq \times pq$, estando los jth's bloques de la diagonal principal conformada por $\hat{\boldsymbol{\Sigma}}_1/n_1, \hat{\boldsymbol{\Sigma}}_2/n_2, \ldots, \hat{\boldsymbol{\Sigma}}_p/n_p$ y los de las diagonales secundarias por ceros, y $\boldsymbol{G}$ una matriz de bloques de orden $pq \times pq$ con los bloques de la diagonal principales tomando el valor de $n_j(1 - n_j/N)(\boldsymbol{I} - \mathbf{1}\mathbf{1}'/q)$ y los de las diagonales secundarias $-N^{-1}(\boldsymbol{I} - \mathbf{1}\mathbf{1}' \div q) n_j n_j'$.

Para aquellas situaciones en las que el supuesto de esfericidad multimuestral esté próximo a ser satisfecho, Huynh (1978) ha señalado que los estadísticos $\hat{h}$ y $\hat{h}'$ tienden a subestimar el valor de sus parámetros correspondientes. En los casos en los que esto suceda, Huynh recomienda efectuar una mejora en el enfoque *GA* y utilizar el enfoque *IGA* mediante la sustitución de $\hat{h}$ por $\tilde{h} = \tilde{\eta}/\tilde{\delta}$, donde

$$\tilde{\eta} = \sum_{j=1}^{p} \frac{(n_j - 1)^3}{(n_j + 1)(n_j - 2)} \left\{ n_j \left(\operatorname{tr} \boldsymbol{C}^{*\prime} \hat{\boldsymbol{\Sigma}}_j \boldsymbol{C}^{*} \right)^2 - 2 \operatorname{tr} \left(\boldsymbol{C}^{*\prime} \hat{\boldsymbol{\Sigma}}_j \boldsymbol{C}^{*} \right)^2 + \right.$$

$$\left. + \sum_{j} \sum_{j'} \left[(n_j - 1)(n_j' - 1) \left(\boldsymbol{C}^{*\prime} \hat{\boldsymbol{\Sigma}}_j \boldsymbol{C}^{*} \right) \left(\boldsymbol{C}^{*\prime} \hat{\boldsymbol{\Sigma}}_j \boldsymbol{C}^{*} \right) \right] \right\}$$

y

$$\tilde{\delta} = \sum_{j=1}^{h} \frac{(n_j - 1)^2}{(n_j + 1)(n_j - 2)} \left[(n_j - 1) \operatorname{tr} \left(\boldsymbol{C}^{*\prime} \hat{\boldsymbol{\Sigma}}_j \boldsymbol{C}^{*} \right)^2 - \left(\operatorname{tr} \boldsymbol{C}^{*\prime} \hat{\boldsymbol{\Sigma}}_j \boldsymbol{C}^{*} \right)^2 \right]$$

A su vez, en base a la enmienda que Huynh y Feldt (1976) realizaron del estimador $\hat{\varepsilon}$ de Greenhouse y Geisser (1959), Huynh (1978), sugiere reemplazar $\hat{h}'$ por

$$\tilde{h}' = \frac{N\hat{h}' - 2}{N - p - \hat{h}'}$$

y $\hat{h}''$ por

$$\tilde{h}'' = \frac{(p-1)\left[N\hat{h}'' - 2(p-1)\right]}{(N-p)(p-1) - \hat{h}''}$$

En conexión con el procedimiento *IGA*, Algina y Oshima (1994) sugieren efectuar una mejora del mismo en base a la corrección que Lecoutre (1991) realizó del $\tilde{\varepsilon}$ de Huynh y Feldt (1976). De acuerdo con dicha modificación, Algina y Oshima aconsejan reemplazar $\tilde{h}'$ por

$$\tilde{h}^{*'} = \frac{(N-p+1)\hat{h}'-2}{N-p-\hat{h}'}$$

y $\tilde{h}''$ por

$$\tilde{h}^{*''} = \frac{(p-1)\left[(N-p+1)\hat{h}''-2(p-1)\right]}{(N-p)(p-1)-\hat{h}''}$$

Resaltar, por último, que el enfoque de la aproximación general de Huynh (1978), así como las subsiguientes modificaciones que del mismo hemos expuesto más arriba, sirve para contrastar hipótesis nulas referidas a medias ponderadas; sin embargo, al contar con distintas unidades experimentales dentro de cada grupo parece razonable que el investigador contemple este hecho y, por ende, le interese probar hipótesis referidas a medias no ponderadas. De resultar cierto lo dicho, en esta nueva situación, además de tener que computarse los estadísticos F correspondientes a la igualdad de las respuestas y a la ausencia de la interacción entre las variables efectuando un *AVAR* no ponderado, los argumentos de los valores críticos bF_α; $\hat{h}'$, $\hat{h}$ y cF_α; $\hat{h}''$, $\hat{h}$ deben definirse en términos de la matriz de dispersión $\boldsymbol{\Sigma}^*$ $\left(\boldsymbol{\Sigma}^* = \sum_{j=1}^{p} n_j^{-1}\boldsymbol{\Sigma}_j \Big/ \sum_{j=1}^{p} n_j^{-1}\right)$, en vez de la matriz $\boldsymbol{\Sigma}$ definida anteriormente.

5. DETERMINACIÓN DE LAS DIFERENCIAS ENTRE LOS EFECTOS DEL DISEÑO EN AUSENCIA DE ESFERICIDAD MULTIMUESTRAL

Una vez que se han detectado diferencias en los efectos del diseño es necesario determinar qué comparaciones o contrastes son los responsables de las susodichas diferencias. Para alcanzar la meta reseñada, en lo que resta de este trabajo vamos a hacer dos cosas: por un lado, ofrecer alguna prueba estadística que utilice un término de error específico para cada uno de los contrastes y, por otro, describir algún procedimiento de los muchos que existen actualmente para efectuar comparaciones, que al tiempo que mantiene controlada la probabilidad de cometer un error de Tipo I al nivel de significación elegido para la familia de contrastes, no esté exento de potencia de prueba.

Como ha sido reiterado, en aquellas investigaciones que resulte significativa la interacción lo sensato sería que el investigador concentrase todos sus esfuerzos en el análisis de ésta. Para tal cometido, los investigadores más propensos a las cuestiones de carácter aplicado recomiendan centrarse en el análisis de las diferencias entre los

niveles de una variable en función de cada uno de los niveles de la otra; es decir, en el análisis de los efectos principales simples; por el contrario, los investigadores más atraídos por las cuestiones de carácter teórico recomiendan analizar exclusivamente los contrastes de la interacción. A nuestro modo de ver, si obviamos lo fácil que suele resultar la interpretación de los efectos principales simples, desde un punto de vista estrictamente teórico, el segundo enfoque resulta más apropiado. La razón es sencilla, ya que para un contraste particular la hipótesis que está siendo considerada es consistente con la prueba global del efecto de la interacción.

Actualmente, existen diversos tipos de contrastes para analizar la interacción (Boik, 1993; Timm, 1994; Gabriel, Putter y Wax, 1973), no obstante, los contrastes producto o interacción contraste-contraste resultan los más fáciles de construir y también de interpretar. Por ejemplo, en un modelo con dos factores, la interacción contraste-contraste se obtiene llevando a cabo el producto Kronecker entre dos vectores, conformando cada uno de éstos un contraste entre los niveles de ese factor principal. De este modo, en un diseño de medidas parcialmente repetidas 2×3 los contrastes producto serían tres; mientras que si el diseño hubiese sido un 3×3 los contrastes producto serían nueve. En ausencia de esfericidad multimuestral, comprobar la significación de cada uno de los contrastes producto requiere hacer uso de un estadístico que nos permita obtener el error estándar de los datos implicados en el contraste de interés. Una forma de alcanzar dicho objetivo es mediante el estadístico t' que sigue:

$$t'_{\Psi} = \frac{\left(\overline{Y}_{.jk} - \overline{Y}_{.jk'}\right) - \left(\overline{Y}_{.j'k} - \overline{Y}_{.j'k'}\right)}{\sqrt{\dfrac{\boldsymbol{c}'\hat{\boldsymbol{\Sigma}}_j\boldsymbol{c}}{n_j} + \dfrac{\boldsymbol{c}'\hat{\boldsymbol{\Sigma}}'_j\boldsymbol{c}}{n_{j'}}}} \sim t\frac{\alpha}{2}, \nu'$$

donde $\hat{\boldsymbol{\Sigma}}_j$ es la matriz de varianzas-covarianzas correspondiente al jth nivel de la variable entre y $\boldsymbol{c}$ es un vector de coeficientes de orden $q \times 1$ que recoge los niveles kth y k'th de la variable intra sujetos. Si bien el estadístico anterior no sigue exactamente una distribución t de Student, puede ser aproximado a la misma calculando los grados de libertad mediante el procedimiento de Satterthwaite (1946) que sigue:

$$\nu' = \frac{\left[\dfrac{\boldsymbol{c}'\hat{\boldsymbol{\Sigma}}_j\boldsymbol{c}}{n_j} + \dfrac{\boldsymbol{c}'\hat{\boldsymbol{\Sigma}}_{j'}\boldsymbol{c}}{n_{j'}}\right]^2}{\dfrac{\left[\boldsymbol{c}'\hat{\boldsymbol{\Sigma}}_j\boldsymbol{c}\right]^2}{n_j^2(n_j-1)} + \dfrac{\left[\boldsymbol{c}'\hat{\boldsymbol{\Sigma}}_{j'}\boldsymbol{c}\right]^2}{n_{j'}^2(n_j-1)}}$$

Otra vía más sencilla de alcanzar los mismos resultados, es mediante el procedimiento Welch-James desarrollado a lo largo del presente trabajo. En este caso concreto, el enfoque Welch-James no es ni más ni menos que una versión generalizada de la prueba t' expuesta más arriba. Esta nueva situación requiere usar el vector de contrastes $\boldsymbol{r}$ que sigue:

$$\boldsymbol{r} = \boldsymbol{c}'_{jj'} \otimes \boldsymbol{a}'_{kk'}$$

donde $\boldsymbol{c}'_{jj'}$ y $\boldsymbol{a}'_{kk'}$ son vectores de coeficientes que contrastan los niveles de las variables entre e intra, respectivamente.

A partir de aquí, al investigador sólo le resta utilizar alguno de los numerosos procedimientos que existen para efectuar comparaciones de medias, y que le permita mantener la tasa de error de Tipo I controlada al nivel nominal para todas las comparaciones que componen la familia de contrastes de la interacción. Pues bien, sendos trabajos de simulación llevados a cabo por Keselman y Lix (1996) y Lix y Keselman (1996), ponen de relieve que tanto el procedimiento secuencial de arriba-abajo de Holm (1979) y modificado por Shaffer (1986), como el procedimiento de abajo-arriba de Hochberg (1988) controlan adecuadamente la tasa de error de Tipo I para los contrastes de la interacción en diseños con características similares a las aquí aludidas; esto es, diseños de medidas parcialmente repetidas con matrices de varianzas y covarianzas heterogéneas y desigual número de unidades experimentales dentro de cada grupo.

A su vez, para contrastar las hipótesis referidas a las comparaciones múltiples efectuadas entre los niveles de la variable intra sujeto cuando el diseño, además de tener un número arbitrario de sujetos dentro de cada grupo, no satisface el supuesto de esfericidad multimuestral, también conviene utilizar algún estadístico que permita calcular el error estándar a partir de los datos implicados en el contraste de interés. Una forma de alcanzar dicho objetivo consiste en utilizar el estadístico t' convenientemente adaptado a esta situación como sigue:

$$t'_{\Psi} = \frac{\left(\overline{Y}_{..k} - \overline{Y}_{..k'}\right)}{\sqrt{\frac{\boldsymbol{c}'\hat{\boldsymbol{\Sigma}}_j \boldsymbol{c}}{n_j} + \frac{\boldsymbol{c}'\hat{\boldsymbol{\Sigma}}_{j'} \boldsymbol{c}}{n_{j'}}}} \sim t\frac{\alpha}{2}, \nu'$$

donde $\hat{\boldsymbol{\Sigma}}_j$ denota la matriz de varianzas-covarianzas para el jth nivel de la variable entre y $\boldsymbol{c}'$ es un vector de coeficientes de orden $1 \times q$ referido a los niveles kth y k'th de la variable intra. Como en el caso de la interacción, el estadístico t' no se distribuye exactamente como t de Student, pero puede aproximarse a dicha distribución muestral calculando los grados de libertad ν' mediante el procedimiento de Satterthwaite allí descrito.

Alternativamente, una forma más fácil de obtener pruebas robustas y también poderosas de las comparaciones múltiples para los niveles de la variable intra, consiste en utilizar el manido procedimiento de Welch-James; pues como el lector habrá reparado, la única modificación a lo ya dicho afecta tan sólo a la construcción del vector de contrastes $\boldsymbol{r}$. Para este caso concreto, $\boldsymbol{r} = \boldsymbol{c}'_j \otimes \boldsymbol{a}'_{kk'}$, donde $\boldsymbol{c}_j$ es un vector de unos de dimensión $p \times 1$ y $\boldsymbol{a}_{kk'}$ es un vector de contrastes referido a las diferentes comparaciones entre los niveles de la variable intra cuya dimensión es $q \times 1$.

Señalar, por último, que para la hipótesis de igualdad de las medidas repetidas, los resultados empíricos de Keselman y Lix (1996) ponen de relieve que el procedimiento protegido de Shaffer (1986) proporciona un adecuado control de la tasa de error de Tipo I, al tiempo que se manifiesta ligeramente más potente que el resto de los procedimientos testados. Seguidamente, vamos a efectuar una breve descripción de los procedimientos de Holm-Shaffer y Hochberg que acabamos de reseñar a lo largo de este apartado.

5.1. El procedimiento de Holm-Shaffer

Shaffer (1986) ha propuesto una versión más poderosa del procedimiento de rechazo secuencial desarrollado por Holm (1979). Bajo la modificación desarrollada por esta investigadora uno comienza ordenando por rangos los valores p asociados con el estadístico de la prueba, $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(c)}$. A continuación, el valor p más pequeño es comparado con el valor crítico de Dunn, esto es, α/c, donde $c = q(q-1)/2$ y α denota la tasa de error a controlar; si $p_{(1)} \leqslant \alpha/c$ la hipótesis nula $H_{(1)}$ es rechazada, de lo contrario, detenemos el proceso y declaramos nulas todas las hipótesis, $H_{(1)}, H_{(2)}, \ldots, H_{(c)}$. Si $H_{(1)}$ es rechazada, procedemos a comparar el siguiente valor p más largo, $p_{(2)}$, aquí a diferencia del procedimiento de Holm, en vez de dividir α por el número de pares de comparaciones restantes, $c-1$, Shaffer propone dividirlo por el número de pares de comparaciones que podrían ser iguales, condicionados al rechazo de la hipótesis previa, c_2^*; si $p_{(2)} \leqslant \alpha/c_2^*$, la hipótesis $H_{(2)}$ es rechazada, en caso contrario, detenemos el proceso. El procedimiento continúa del modo expuesto rechazando H_k ($k = 1, 2, \ldots, c$) si $p_{(k)} \leqslant \alpha/c_k^*$, dados $k-1$ rechazos previos. Por ejemplo, cuando $k = 4$ la diferencia entre pares de medias más larga es probada frente a un determinado valor crítico (t, F, q, ...) basada en $\alpha/6$, pues $c = 4(4-1)/2 = 6$. En esta primera fase, tanto el procedimiento de Holm como el de Shaffer son plenamente coincidentes; no obstante, en la segunda etapa el procedimiento de Holm probaría el siguiente par de medias más largo basado en $\alpha/5$, mientras que el procedimiento de Shaffer verificaría dicha diferencia usando $\alpha/3$, pues el rechazo de la primera hipótesis (por ejemplo, $\mu_1 \neq \mu_4$) implica que al menos tres de los restantes pares de medias podrían aún seguir siendo iguales (por ejemplo, $\mu_1 = \mu_2$, $\mu_1 = \mu_3$ y $\mu_2 = \mu_3$ ó $\mu_2 = \mu_4$, $\mu_3 = \mu_4$ y $\mu_2 = \mu_3$). Para la tercera y cuarta comparaciones más largas, el procedimiento de Holm utilizaría $\alpha/4$ y $\alpha/3$, respectivamente; sin embargo, el procedimiento de Shaffer emplearía $\alpha/3$ en ambos casos, ya que el rechazo de la segunda hipótesis (por ejemplo, $\mu_1 \neq \mu_3$) implica que un conjunto de tres hipótesis puede seguir siendo verdadero (por ejemplo, $\mu_2 = \mu_4$, $\mu_3 = \mu_4$ y $\mu_2 = \mu_3$), a su vez, el rechazo de la tercera hipótesis (por ejemplo, $\mu_2 \neq \mu_4$) aún permitiría que un conjunto de tres hipótesis podría seguir siendo verdadero (por ejemplo, $\mu_1 = \mu_2$, $\mu_3 = \mu_3$ y $\mu_3 = \mu_4$). Finalmente, para las dos últimas comparaciones, ambos procedimientos usarían de nuevo el mismo α'; esto es, $\alpha/2$ y $\alpha/1$. En la tabla

1 tomada de Shaffer (1986), se muestra la secuencia de divisores (c_k^*) para k = 3, 4, 5, 6 y 7.

Tabla 1. *Número de posibles pares de medias que pueden permanecer iguales entre k medias con r rechazos.*

(r) número de hipótesis rechazadas

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3	1	1																		
4	6	3	3	3	2	1															
5	10	6	6	6	6	4	4	3	2	1											
6	15	10	10	10	10	10	7	7	7	6	4	4	3	2	1						
7	21	15	15	15	15	15	15	11	11	11	11	10	9	7	7	6	5	4	3	2	1

Shaffer ha tabulado el valor de c_k^* para $k = 3, \ldots, 10$.

En el mismo trabajo de 1986, Shaffer propuso otra modificación del procedimiento de Holm. La única diferencia entre el procedimiento que acabamos de describir y la segunda modificación propuesta por Shaffer, radica tan sólo en el empleo de una prueba de igualdad de medias global. Es decir, que con esta versión protegida de Shaffer se procede a efectuar el contraste de medias una vez verificado que la hipótesis nula global es falsa. Tras el rechazo de la hipótesis nula global, lo cual implica que al menos una comparación es distinta de cero, sin que necesariamente tenga que ocurrir la diferencia entre un simple par de medias, se procede como en el caso descrito pero asumiendo el rechazo de la hipótesis más larga; de este modo, para el ejemplo especificado, si la hipótesis nula global es rechazada, la diferencia más larga sería verificada utilizando $\alpha/3$, en vez de $\alpha/6$, para las hipótesis restantes procederíamos de un modo similar a lo dicho para la versión no protegida. Por consiguiente, si la hipótesis nula global es rechazada, entonces la versión protegida de Shaffer ofrece una mayor potencia del test estadístico asociado con el par de medias que difieren más entre sí.

Si bien la autora a la que nos venimos refiriendo, ha tabulado el valor de c_k^* para distintos niveles de la variable ($k = 3, \ldots, 10$), para efectuar el análisis de los contrastes de la interacción recomienda utilizar la rutina que sigue: Si $H_{(1)}$ es rechazada utilizando el criterio de Dunn, α/c, probamos $H_{(2)}$ empleando como divisor c_2^*, donde $c_2^* = c - (p-1)(q-1)$; esto es, α/c_2^*. A su vez, si $H_{(2)}$ es rechazada, se fija $c_k^* = c_2^*$ para todo $2 \leqslant k \leqslant c - c_2^* - 1$ y asignamos un valor de $c - k + 1$ para todo $k > c - c_2^* + 1$. Por ejemplo, si tenemos un diseño factorial $A \times B$ con $p = 2$ y $q = 4$, el número total de contrastes para la interacción, c, es igual a $p^* \times q^* (p^* = p(p-1)/2$ y $q^* = q(q-1)/2)$. Así pues, de acuerdo con la primera versión del procedimiento de Shaffer, para $c = 1, \ldots, 6$, tenemos los siguientes valores de α':

$$\alpha'_1 = \frac{\alpha}{c_1^*} = \frac{\alpha}{6}$$

$$\alpha'_2 = \frac{\alpha}{c_2^*} = \frac{\alpha}{3} \quad \text{donde} \quad c_2^* = 6-(2-1)\times(4-1)$$

$$\alpha'_3 = \frac{\alpha}{c_3^*} = \frac{\alpha}{3} \quad \text{pues} \quad 2 \leqslant 3 \leqslant 6-3+1$$

$$\alpha'_4 = \frac{\alpha}{c_4^*} = \frac{\alpha}{3} \quad \text{pues} \quad 2 \leqslant 4 \leqslant 6-3+1$$

$$\alpha'_5 = \frac{\alpha}{c_5^*} = \frac{\alpha}{2} \quad \text{ya que} \quad 2 \leqslant 5 \leqslant 6-3+1 \quad \text{es falso. Por tanto} \quad c_5^* = 6-5+1$$

$$\alpha'_6 = \frac{\alpha}{c_6^*} = \frac{\alpha}{1} \quad \text{ya que} \quad 2 \leqslant 6 \leqslant 6-3+1 \quad \text{es falso. Por tanto} \quad c_6^* = 6-6+1$$

Para la versión protegida procederíamos del mismo modo, excepto para la primera diferencia, donde utilizaríamos $\alpha/3$, en vez de $\alpha/6$.

5.2. El procedimiento de Hochberg (1988)

El procedimiento paso a paso de Hochberg (1988) es una de las técnicas secuenciales más simples de cuantas existen. Para implementar este método lo primero que hay que hacer es ordenar por rangos los diferentes valores ρ asociados con el estadístico utilizado para probar las hipótesis correspondientes a los diferentes contrastes; esto es, $\rho_{(1)} \leqslant \rho_{2)} \leqslant \cdots \leqslant \rho_{(c)}$. Bajo este enfoque se parte de la hipótesis que tiene la mayor probabilidad de ser retenida a la que tiene la menor, para ello se comienza testando la hipótesis asociada con el valor ρ más largo. Si $\rho_{(c)} \leqslant \alpha$, rechazamos todas las hipótesis y detenemos el proceso; por el contrario, si $\rho_{(c)} > \alpha$, retenemos $H_{(c)}$ y procedemos a probar $H_{(c-1)}$. Si $\rho_{(c-1)} \leqslant \alpha/2$, además de la hipótesis implicada rechazamos todas las demás; en caso contrario, $H_{(c-1)}$ es retenida. A continuación, $\rho_{(c-2)}$ es comparado con $\alpha/3$, y así se continúa el proceso hasta que alguna hipótesis resulte rechazada. De ser retenidas todas las hipótesis el proceso se finaliza comparando $\rho_{(1)}$ con α/c.

6. A MODO DE CONCLUSIÓN

En el presente trabajo se ponen de relieve algunos de los problemas más importantes a los que tienen que enfrentarse los investigadores que usan los diseños de medidas repetidas, en especial, cuando no se satisface el supuesto de esfericidad multimuestral y el número de unidades experimentales varía a lo largo de los diferentes grupos de

los que consta la variable de tratamiento. A su vez, a lo largo del mismo se describen con cierto grado de detalle varias pruebas que se hallan disponibles actualmente, y cuya utilización permitiría a los investigadores solventar muchos de los problemas a los que irremediablemente les conducen los estadísticos utilizados tradicionalmente en estas circunstancias. Específicamente, mostramos cómo aplicar el enfoque Welch-James para probar hipótesis relacionadas con los efectos de los diseños univariados y multivariados de medidas repetidas. Para lograr este cometido se hace un énfasis especial en la posibilidad de que la interacción resulte significativa, aspecto éste que a pesar de resultar de capital importancia, pasa desapercibido para la mayoría de los investigadores; así como, en la forma de llevar a cabo inferencias para determinar los contrastes responsables de las diferencias obtenidas en los efectos del diseño.

Con todo, y a pesar de ofrecer un amplio abanico de posibilidades de aplicación, el enfoque Welch-James no es la panacea. En concreto, para datos distribuidos normalmente su robustez se va deteriorando a medida que la razón entre el tamaño del grupo más pequeño y el número de medidas repetidas menos uno es inferior a 1.7 ó 2, cuando se prueban hipótesis referidas a los efectos principales y a 3 ó 4, para el caso referido a la interacción. Obviamente, cuando los datos se apartan de la normalidad el deterioro de la robustez de este enfoque se deja sentir para razones inferiores a 3 ó 4, en el caso de los efectos principales y para razones inferiores a 5 ó 6, para el caso de la interacción.

Finalmente, concluimos reseñando que cuando un investigador se encuentre inmerso en alguna situación en las que se incumpla el supuesto de esfericidad multimuestral, los datos estén muestreados desde distribuciones probablemente sesgadas y el tamaño de los grupos sea reducido, lo mejor que puede hacer es reducir su nivel de significación, o bien utilizar alguno de los otros enfoques descrito en este artículo. Sobre manera, si se tiene en cuenta que en estas circustancias los estudios de simulación llevados a cabo por Algina y Oshima (1995) ponen de relieve que tanto el procedimiento *IGA*, como *CIGA*, proporcionan un mejor control de la probabilidad de cometer un error de Tipo I para los efectos principales, que el enfoque Welch-James. En lo que a la interacción se refiere, los trabajos de Algina y Oshima (1994) y de Keselman *et al.* (1993), sugieren la necesidad de seguir desarrollando pruebas alternativas. Por ejemplo, una candidata potencial podría ser la versión generalizada de la prueba de Brown y Forsythe (1974).

7. AGRADECIMIENTOS

El autor desea dejar constancia pública de su agradecimiento al evaluador anónimo, pues la nueva versión del trabajo se beneficia, tanto de sus enjundiosos comentarios como de sus acertadas sugerencias.

REFERENCIAS

[1] **Algina, J.** (1994). «Some alternative approximate tests for a split-plot design». *Multivariate Behavioral Research*, **29**, 315-384.

[2] **Algina, J.** y **Olejnik, S.F.** (1984). «Implementing the Welch-James procedure with factorial designs». *Educational and Psychological Measurement*, **44**, 39-46.

[3] **Algina, J.** y **Oshima, T.C.** (1994). «Type I error rates for Huynh's general approximation and improved general approximation tests». *British Journal of Mathematical and Statistical Psychology*, **47**, 151-165.

[4] **Algina, J.** y **Oshima, T.** (1995). «An improved general approximation test for the main effect in a split-plot design». *British Journal of Mathematical and Statistical Psychology*, **48**, 149-160.

[5] **Belli, G.M.** (1988). *Type I error rates of MANOVA of repeated measures under group heterogenity in unbalanced designs.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

[6] **Boik, A.J.** (1993). «The analysis of two-factor interactions in fixed effects linear models». *Journal of Educational Statistics*, **18**, 1-40.

[7] **Box, G.E.P.** (1954). «Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification». *Annals of Mathematical Statistics*, **25**, 290-302.

[8] **Brown, M.B.** y **Forsythe, A.B.** (1974). «The small sample behavior of some statistics which test the equality of several means». *Technometrics*, **16**, 129-132.

[9] **Coombs, W.T.** y **Algina, J.** (1996). «On sample size requeriments for Johansen's test». *Journal of Educational and Behavioral Statistics*, **21**, 169-179.

[10] **Davidson, M.L.** (1972). «Univariate versus multivariate test in repeated measures experiments». *Psychological Bulletin*, **77**, 446-452.

[11] **Gabriel, K.R., Putter, J.** y **Wat, Y.** (1973). «Simultaneous confidence intervals for product-type interaction contrasts». *Journal of the Royal Statistical Society*, **Serie B, 35**, 234-244.

[12] **Greenhouse, S.W.** y **Geisser, S.** (1959). «On methods in the analysis of profile data». *Psychometrika*, **43**, 161-175.

[13] **Hochberg, Y.** (1988). «A sharper Bonferroni procedure for multiple tests of significance». *Biometrika*, **75**, 800-802.

[14] **Holms, S.** (1979). «A simple sequentially rejective multiple test procedure». *Scandinavian Journal of Statistics*, **6**, 65-70.

[15] **Hsiung, T.** y **Olejnik, S.** (1994). *Type I error rates and statistical power for the James second order test and univariate F test in two-way fixed effects ANOVA models under heterocedasticity and/or normality.* Paper presented at the annual meeeting of the American Educational Research Association, New Orleans, LA.

[16] **Huynh, H.** (1978). «Some approximate tests for repeated measurement designs». *Psychometrika*, **43**, 161-175.

[17] **Huynh, H.** y **Feldt, L.S.** (1976). «Conditions under which mean square ratios in repeated measurements design have exact $F-$distributions». *Journal of the American Statistical Association*, **65**, 1582-1585.

[18] **James, G.S.** (1951). «The comparison of several groups of observations when the ratios of the population variances are unknown». *Biometrika*, **38**, 324-329.

[19] **James, G.S.** (1954). «Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown». *Biometrika*, **41**, 19-43.

[20] **Johansen, S.** (1980). «The Welch-James approximation of the distribution of the residual sum of squares in weighted linear regression». *Biometrika*, **67**, 85-92.

[21] **Keselman, H.J.** (1993). «Stepwise multiple comparisons of repeated measures means under violations of multiple sphericity». En F.M. Hope (Ed.), *Multiple Comparisons, Selection and Applications in Biometry*, 167-186. New York: Marcel Dekker.

[22] **Keselman, H.J.** y **Keselman, J.C.** (1988). «Repeated measures multiple comparison procedures: Effects of violating multisample sphericity in unbalanced designs». *Journal of Educational Statistics*, **13**, 215-236.

[23] **Keselman, H.J.** y **Keselman, J.C.** (1990). «Analysing unbalanced repeated measures designs». *British Journal of Mathematical and Statistical Psychology*, **43**, 265-282.

[24] **Keselman, H.J.** y **Lix, L.M.** (1995). «Improved repeated measures: Stepwise multiple comparison procedures». *Journal of Educational and Behavioral Statistics*, **20**, 83-99.

[25] **Keselman, H.J., Carriere, M.C.** y **Lix, L.M.** (1993). «Testing repeated measures hypotheses when covariance matrices are heterogeneous». *Journal of Educational Statistics*, **18**, 305-319.

[26] **Keselman, H.J., Carriere, M.C.** y **Lix, L.M.** (1995). «Robust and powerful nonorthogonal analyses». *Psychometrika*, **60**, 395-418.

[27] **Keselman, J.C., Lix, L.M.** y **Keselman, H.J.** (1993). *The analysis of repeated measurements: A quantitative research synthesis.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

[28] **Lecoutre, B.** (1991). «A correction for the approximate test in repeated measures designs with two or more independent groups». *Journal of Educational Statistics*, **16**, 371-372.

[29] **Lix, L.M.** y **Keselman, H.J.** (1995). «Approximate degrees of freedom tests: A unified perspective on testing for mean equality». *Psychological Bulletin*, **117**, 547-560.

[30] **Lix, L.M.** y **Keselman, H.J.** (1996). «Interaction contrasts in repeated measures designs». *British Journal of Mathematical and Statistical Psychology*, **49**, 147-162.

[31] **Milligan, G.W., Wong, D.W.** y **Thompson, P.A.** (1987). «Robustness properties of nonorthogonal analysis of variance». *Psychological Bulletin*, **101**, 464-470.

[32] **Nel, D.G.** y **van der Merwe, C.A.** (1986). «A solution to the multivariate Behrens-Fisher problem». *Communications in Statistics-Theory and Methods*, **15**, 3719-3735.

[33] **Rogan, J.C., Keselman, H.J.** y **Mendoza, J.L.** (1979). «Analysis of repeated measurements». *British Journal of Mathematical and Statistical Psychology*, **32**, 269-286.

[34] **Roth, A.J.** (1983). «Robust trend tests derived and simulated: Analogs of the Welch and Brown-Forsythe tests». *Journal of the American Statistical Association*, **78**, 972-980.

[35] **Satterthwaite, F.E.** (1946). «An approximate distribution of estimates of variance components». *Biometrics*, **2**, 110-114

[36] **Shaffer, J.P.** (1986). «Modified sequentially rejective multiple test procedures». *Journal of the American Statistical Association*, **81**, 826-831.

[37] **Timm, N.H.** (1975). *Multivariate Analysis with applications in Education and Psychology*. Monterey, CA: Brooks/Cole.

[38] **Timm, N.H.** (1994). *Analysis of interactions: Another look.* Paper presented at the annual meeting of the American Educational Association, New Orleans, LA.

[39] **Vallejo, G.** y **Menéndez, I.A.** (1997). «Una comparación de enfoques alternativos para el análisis de diseños multivariados de medidas repetidas». *Psicothema*, **9(3)**, 657-666.

[40] **Welch, B.L.** (1938). «The significance of the difference between two-mean when the population variances are unequal». *Biometrika*, **29**, 350-362.

[41] **Welch, B.L.** (1947). «The generalization of "Student's" problem when several different population variances are involved». *Biometrika*, **34**, 28-35.

[42] **Welch, B.L.** (1951). «On the comparison of several mean values: An alternative approach». *Biometrika*, **38**, 330-336.

[43] **Wilcox, R.R.** (1987). «New designs in analysis of variance». *Annual Reviev of Psychology*, **38**, 29-60.

[44] **Wilcox, R.R., Charlin, V.L.** y **Thompson, K.L.** (1986). «New Monte Carlo results on the robutness of the ANOVA F, W and F^* Statistics». *Communication in Statistics-Simulation and Computation*, **15**, 933-943.

ENGLISH SUMMARY

SOME APPROXIMATE SOLUTIONS FOR SPLIT-PLOT DESIGNS WITH ARBITRARY COVARIANCE MATRICES

G. VALLEJO SECO
J.R. ESCUDERO GARCÍA
Universidad de Oviedo*

This paper describes in some detail alternative types of analyses for unbalanced split-plot designs when the assumption of multisample sphericity is violated. Specifically, by adapting a multivariate approximate degrees of freedom approach developed by Johansen (1980) and a revised improved general approximation procedure based on Huynh (1978) it is shown how to obtain robust and powerful analysis to test within-subjects main effect and the between x within interaction, as well as multiple comparison hypotheses related to these effects, so much if is counted with a single dependent variable associated with each repeated measurement as if is counted with multiple dependent variables.

Keywords: Multivariate repeated measures designs, multisample sphericity, weighted and unweighted means, robust and powerful procedures.

AMS Classification: 62K10, 62J1

*Departamento de Psicología. Universidad de Oviedo. Plaza Feijoo, s/n. 33003 Oviedo. e-mail:gvallejo@sci.cpd.uniovi.es.

-Received December 1996.
-Accepted May 1998.

The design of repeated measures is frequently used in the investigations carried out in social, behavioral and health sciences. If the following assumptions are met: multivariate normality, homogeneity of dispersion matrices, equality of the variances corresponding to the differences between repeated measures (circularity or sphericity assumption) and independence among the observations, these designs have been traditionally analyzed by means of Scheffé's (1956) univariate mixed model. In this model, subjects (blocks) are random and treatments are fixed.

If the sphericity assumption is not met, which happens in most psychological data, the investigator has two options to analyze. On one hand, you can use an univariate mixed model with the degrees of freedom adjusted by means of some of the numerous existent Box type correctors (Lecoutre, 1991) and, on the other hand, a multivariate model. This last approach has the advantage of allowing that the variance-covariance matrix adopt any structure., Although it requires, in addition the accomplishment of the assumptions of multivariate normality and homogeneity of covariance, that the number of experimental units be bigger, or at least the same as the one of repeated measures. If the exposed conditions are satisfied, potency is used for the election of the univariate approach with the corrected degrees of freedom or the multivariate approach. For example, Davidson (1972) demonstrates that with moderate sample sizes and the covariance matrix lightly deviated of the required sphericity pattern, the univariate mixed model is always more powerful than the corresponding multivariate approach. The situation is gradually inverted as the dispersion matrix deviates of the required sphericity pattern.

When the multisample sphericity assumption is violated and group sizes is unequal, several investigators (Belli, 1988; Huynh and Feldt, 1976, Keselman, Carriere and Lix, 1993) wonder if both procedures are robust, mainly, if the number of experimental units differs from a group to another of the design. To deal with the specified problem, along the present work different procedures are described. Specifically, we show how to apply the Welch-James approach to prove hypothesis related with within-subjects main effect and the between x within interaction, so much for split-plot designs containing a single dependent variable as for split-plot designs containing multiples dependent variables. To achieve this, a special emphasis is made in the two following aspects:

a) In the hypothesis testing depending on the model (additive or interactive). Although in presence of interaction analytic strategy should be different from that when it is not present, most of investigators do not realize this question and in fact almost all of them testing their hypotheses according to the postulates of an additive model.

b) In the form to accomplish inferences to determine the contrasts responsible for the differences obtained between the levels and combinations of levels of the unbalanced partially repeated measures designs involving heterogeneous covariance

matrices. In order to reach this last goal several multiple comparison procedures are presented in this paper.

However, in spite of offering many application possibilities, the approach Welch-James is enough sensitive to the assumption violations when the group sizes is not sufficiently large. In short, Keselman *et al.* (1993) have shown that the critical factor in determining the robustness of procedure for univariate repeated measures designs in which covariance matrices are heterogeneous and group sizes are unequal is the ratio of the smallest group size to the number of repeated measurements minus one. Particularly, for data with normal distribution their robustness deteriorates as the reason between the size of the smallest group and the number of repeated measures minus one is smaller than 1.7 or 2, when you try to prove hypothesis referred to the main effects. In the case referred to interaction, the deterioration occurs for reasons smaller than 3 or 4. Obviously, when the data are obtained from skewed distributions, the deterioration of the robustness of this approach appears for reasons smaller than 3 or 4, in the case of the main effects and it for reasons smaller than 5 or 6, in the case of interaction. Finally, we conclude pointing out that when an investigator is in some situation in which the assumption of multisample sphericity is not met, the data are to sample from non-normal distributions and the size of the groups is reduced, the best thing to do is to reduce its significance level. It can also be used some of the other approaches described in this article. Especially, if one bears in mind that in these circumstances, the simulation studies carried out by Algina and Oshima (1995) show that the Improved General Approximation and Corrected Improved General Approximation procedures provide a better control of the probability of committing Type I errors for the main effects than the Welch-James approach. With respect to interaction, the works of Algina and Oshima (1994) and of Keselman *et al.* (1993) suggest the necessity to continue developing alternative tests: for example, one possibility is to replace Welch-James procedure at the multivariate version of Brown and Forsythe (1974) procedure. Anyhow, further studies are needed.

Investigació Operativa

QÜESTIIÓ, vol. 22, 3, p. 471-489, 1998

TOUR EULERIÀ SENSE GIRS EN U EN UN GRAF ORIENTAT SIMPLE

D. SOLER FERNÁNDEZ
Universitat Politècnica de València*

Sigui $G = (V,A)$ un graf orientat eulerià simple, hi estudiem la recerca d'un tour eulerià sense girs en U, és a dir, sense recórrer consecutivament parells d'arcs $(u,v),(v,u)$, $u,v \in V$. Desconeguda la complexitat d'aquest problema, generalitzem un resultat d'un cas particular resolt en temps polinomial, proporcionant una condició sota la qual es pot construir en temps polinomial un tour eulerià sense girs en U sobre G. Aquesta condició es basa, a més a més, en l'eliminació de vèrtexs candidats a contenir girs en U en un tour eulerià amb mínim nombre d'ells.

Eulerian tour without U-turns in a simple digraph

Paraules clau: Gir en U, graf orientat U-eulerià, H-subgraf.

Classificació AMS: 68R05,68R10

* Universitat Politècnica de València. Departament de Matemàtica Aplicada. Camí de Vera, 14. 46071 València.

- Rebut l'abril de 1997.
- Acceptat el setembre de 1998.

1. INTRODUCCIÓ I CONCEPTES PREVIS

Sigui $G = (V,A)$ un graf orientat, el problema de trobar-hi un tour eulerià (tour que recorre casdascun dels arcs exactament una vegada), és un problema clàssic i senzill de Teoria de Grafs. És suficient veure que G és connex i que el nombre d'arcs que arriben a v ($d^+(v)$) coincideix amb el nombre d'arcs que surten de v ($d^-(v)$) $\forall v \in V$. Es diu aleshores que G és eulerià i existeixen diverses tècniques per trobar-hi un tour eulerià, potser la més coneguda sigui la regla de Fleury. Aquest problema serveix de base per a la resolució d'altres problemes més complexos amb aplicacions importants al camp de la investigació operativa: transport de mercaderies, recollida d'escombraries, llevaneus, repartiment de correu, etc.

L'estudi es complica si imposem la restricció que el tour eulerià no recorri consecutivament parells d'arcs $(u,v),(v,u)$, $u,v \in V$, és a dir, que no realitzi **girs en U**, girs que són els prohibits per antonomàsia en problemes reals de rutes de vehicles. No tot graf orientat eulerià admet un tour eulerià sense girs en U i si l'admet, les tècniques clàssiques de recerca d'un tour eulerià no garanteixen trobar-ne un sense girs en U. A més a més, el nombre de tours eulerians distints en un graf orientat no és polinomial en la grandària del graf com demostra Fleischner (1983).

Per simplificar les referències en aquest problema, direm que un graf orientat és **U-eulerià** si admet un tour eulerià sense girs en U, al qual anomenarem **tour U-eulerià**. Estudiem en aquest article el problema de reconèixer si un graf orientat és U-eulerià, i ens restringim al cas que el graf orientat sigui simple (sense arcs repetits). Aquest problema l'anomenarem Problema del Graf Orientat U-eulerià (PGOU) i abreviarem l'expressió *graf orientat eulerià simple* per g.o.e.s.

Com veurem en la secció següent, el PGOU es pot resoldre sense necessitat d'enumerar els tours eulerians del graf orientat fins trobar-ne un sense girs en U; el problema rau en què aquesta resolució consumeix un temps exponencial en la grandària del graf orientat. La clau es troba, per tant, en saber si el PGOU és un problema polinomial o NP-complet. Aquesta qüestió no té resposta de moment, malgrat els nostres esforços al respecte, encara que donada la complexitat del PGOU, tot sembla indicar que és NP-complet.

L'objecte d'aquest article és, per tant, donar una condició suficient de g.o.e.s. U-eulerià, verificable en temps polinomial i amb les menors restriccions possibles. Per a tal efecte, generalitzarem un cas particular del PGOU resolt en temps polinomial per Thomassen (1990), el qual exposarem també en la secció següent.

Al llarg d'aquest treball farem ús de les notacions i definicions bàsiques següents, donades totes elles sobre un g.o.e.s. $G = (V,A)$:

– Anomenarem **invers** d'un arc, un arc paral·lel i de sentit contrari que connecta el mateix parell de vèrtexs.

– Denotarem per **U(v)** el nombre de parelles d'arcs inversos incidents amb el vèrtex v de G.

– Anomenarem **vèrtex final** un vèrtex v amb $d^+(v) = U(v) = 1$.

– Direm que els vèrtexs $v = u_0, u_1, \ldots, u_{k+1} = w$, $k \geq 0$ formen un **camí maximal doble** en G si acompleixen les tres condicions següents:

1) $(d^+(v) \neq 2$ o $U(v) \neq 2)$ i $(d^+(w) \neq 2$ o $U(w) \neq 2)$.

2) $(u_i, u_{i+1}), (u_{i+1}, u_i) \in A$ $i = 0, 1, \ldots, k$.

3) Si $k \geq 1$, $d^+(u_i) = U(u_i) = 2$ $i = 1, \ldots, k$.

Si $k \geq 1$ als vèrtexs u_i $i = 1, \ldots, k$ els anomenarem **vèrtexs interiors** del camí maximal doble.

– Sigui $v \in V$ amb $d^+(v) \geq 3$ i sigui T un tour eulerià en G. T passa pel vèrtex v almenys tres vegades. Siguin a_1a_2, a_3a_4 i a_5a_6 tres transicions (girs) de T en v, fetes en aquest ordre, amb $a_1, \ldots a_6 \in A$. Anomenarem **3-intercanvi** l'operació de permutar aquests girs en v pels girs a_1a_4, a_3a_6 i a_5a_2 (veure Figura 1).

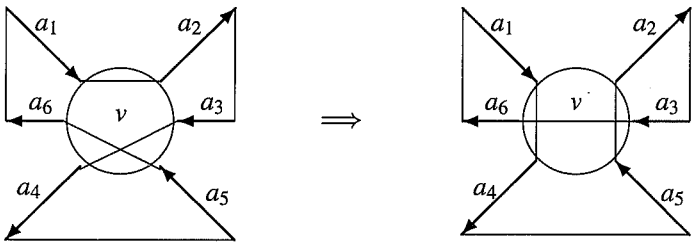


Figura 1. Exemple de 3-intercanvi a un vèrtex v.

Noteu que un 3-intercanvi ens proporciona un nou tour eulerià en G.

2. CAS PARTICULAR I RESOLUCIÓ DEL PGOU EN TEMPS EXPONENCIAL

El PGOU generalitza un altre problema de Teoria de Grafs anomenat Problema del **doble recorregut fort** i que consisteix en: donat un graf no orientat connex, saber si

conté una cadena tancada que recorre cadascuna de les seves arestes exactament dues vegades, una en cada sentit i no de manera consecutiva (sense realitzar girs en U). Aquesta cadena tancada rep el nom de doble recorregut fort.

El problema del doble recorregut fort va ser plantejat per Ore (1951) i el primer resultat teòric sobre aquest problema va ser donat per Troy (1966), qui va demostrar que *donat un graf no orientat G, amb tots els seus vèrtexs amb grau inferior a 4, no conté un doble recorregut fort si el nombre de vèrtexs de G amb grau 3 és múltiple de 4.* Altres autors, com Brenner i Lindon (1984) i Eggleton i Skilton (1984) van tractar el problema del doble recorregut fort, però va ser Thomassen (1990) qui, en les seves paraules, va resoldre el vell problema plantejat per Ore, en demostrar el teorema següent, fent ús d'alguns resultats obtinguts per Xuong (1979):

Teorema 1. *Donat un graf connex no orientat G, els enunciats següents són equivalents:*

1) *G admet un doble recorregut fort.*
2) *G no conté vèrtexs amb grau 1 i conté un arbre generador T, tal que tota component connexa de $G-T$ amb un nombre imparell d'arestes, conté un vèrtex v amb grau major que 3 en G.*

La importància d'aquesta caracterització, de la qual farem ús a la secció cinquena, rau en l'existència d'un algorisme polinomial que troba tal arbre o decideix que no existeix. Thomassen demostra l'existència de l'algorisme anterior basant-se en l'existència d'un algorisme polinomial per resoldre el Problema de la Paritat en matroids gràfiques, algorisme donat per Gabow i Stallmann (1986), encara que aquesta demostració conté algunes falles. Benavent i Soler (1998a) proporcionen una versió corregida d'aquesta demostració i, basant-se en ella, donen un algorisme que en temps polinomial troba el doble recorregut fort o decideix que tal recorregut no existeix.

Per tant, el problema del doble recorregut fort és un cas particular del PGOU —és suficient desdoblar cada aresta del graf no orientat en dos arcs de sentits oposats (veure exemple en Figura 2)— resolt en temps polinomial. Com hem dit a la secció anterior, malgrat els nostres esforços al respecte, no saben si el PGOU és polinomial o NP-complet, però almenys el sabem resoldre en temps exponencial mitjançant la seva transformació al Problema del Circuit Hamiltonià que, recordem, consisteix a saber si un graf orientat $G=(V,A)$ admet un circuit que passi per cadascun dels vèrtexs de V exactament una vegada (circuit hamiltonià). Aquest problema és NP-complet i existeixen diversos mètodes per a la seva resolució. Laporte (1992) recopila els coneguts fins a la data.

El PGOU sobre un graf orientat eulerià no necessàriament simple $G = (V,A)$ es transforma fàcilment al Problema del Circuit Hamiltonià en la forma següent:

Construir un graf orientat $G' = (V',A')$ on V' es correspon biunívocament amb A ($v_a \in V'$ si i sols si $a \in A$) i l'arc (v_a, v_b) pertany a A' si, i sols si, $a = (i,j)$, $b = (j,k)$ i $k \neq i$, és a dir, cada arc de G' es correspon amb un gir no en U en el graf orientat G. És fàcil veure que G admet un tour U-eulerià si, i sols si, G' admet un circuit hamiltonià, essent evident la transformació del circuit hamiltonià en G' al tour U-eulerià en G.

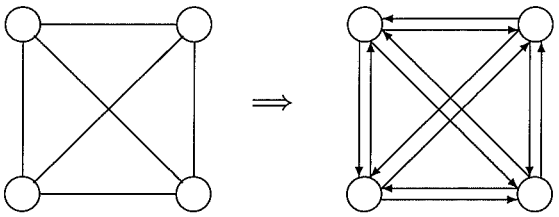


Figura 2. Exemple de transformació d'un problema del doble recorregut fort en un PGOU.

3. H-SUBGRAFS

Un graf orientat simple U-eulerià ha d'acomplir dues condicions bàsiques: no contenir vèrtexs finals i contenir almenys un vèrtex v amb $d^+(v) \neq 2$ o $U(v) \neq 2$ (no ser un cicle doble). En aquesta secció veiem una altra condició necessària que ha d'acomplir tal graf orientat. És per això que donem la definició següent:

Definició 2. *Sigui $G = (V,A)$ un g.o.e.s. i siguin $u,v \in V$ amb $d^+(u) = d^+(v) = 2$ i $U(u)=U(v)=1$ tals que existeix en G un camí maximal doble $u = u_0, u_1, \ldots, u_{k+1} = v$ amb $k \geq 0$. Anomenaren* **H-subgraf** *de G un subgraf de G format per tots els vèrtexs d'aquest camí maximal doble, junt amb tots els arcs incidents i vèrtexs adjacents als vèrtexs esmentats. A u i v els anomenarem* **vèrtexs intermedis** *del H-subgraf.*

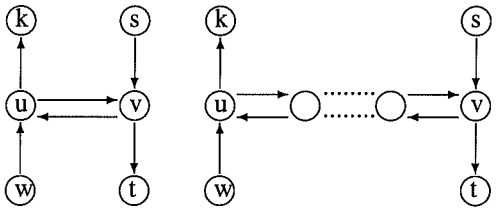


Figura 3. Formes de H-subgraf.

La Figura 3 ens mostra els dos tipus de H-subgrafs que poden aparèixer en un g.o.e.s., segons que $k = 0$ o $k > 0$. Noteu que en aquests H-subgrafs es podrà donar eventualment un dels dos casos següents: a) $k = s$ *i/o* $w = t$ i b) $k = t$ *i/o* $w = s$.

Un tour eulerià amb mínim nombre de girs en U sobre un g.o.e.s. no pot realitzar dos girs en U en un vèrtex interior d'un camí maximal doble que contingui per un extrem un vèrtex v amb $d^+(v) = 2$ i $U(v) = 1$. En cas contrari, creuant ambdós girs en U, obtindríem dos circuits que en creuar-se en v donarien un tour eulerià amb un gir en U menys. De fet, es podrà veure més endavant amb un raonament similar, que un tour eulerià amb mínim nombre de girs en U sobre un g.o.e.s. no contindrà girs en U als vèrtexs interiors d'un camí maximal doble, independentment de com siguin els seus vèrtexs extrems.

Per tant, cada vegada que aparegui un H-subgraf del segon tipus ($k > 0$), es podrà transformar en un altre del primer tipus ($k = 0$) comprimint el camí maximal doble en dos arcs inversos que uneixin els vèrtexs intermedis del H-subgraf. Si abans de realitzar aquesta compressió, existís al graf orientat original l'arc (u,v) o l'arc (v,u) (però no els dos ja que $U(u) = U(v) = 1$), substituïm l'arc (u,v) ((v,u)) per la cadena (u,u'),(u',v) ((v,v'),(v',u)) on v' (u') serà un nou vèrtex amb grau de sortida i grau d'arribada 1. El motiu d'aquesta substitució és que el nou graf orientat continuï essent simple, sense lloc a confusió a l'hora de realitzar o no girs en U.

A partir d'ara suposarem doncs, i sense pèrdua de generalitat, que tots els H-subgrafs tenen els seus vèrtexs intermedis adjacents per dos arcs inversos.

Definició 3. *Considerem un H-subgraf d'un g.o.e.s. i suposem que u i v són els seus vèrtexs intermedis i (u,v),(v,u), (u,k),(w,u),(s,v) i (v,t) els arcs del H-subgraf. Anomenarem* **trencament** *del H-subgraf al desdoblament del seus vèrtexs intermedis en u' i u'' i en v' i v'' respectivament, de manera que obtenim dos camins sense vèrtexs interiors en comú (s,v',u'',k) i (w,u',v'',t) (veure Figura 4).*

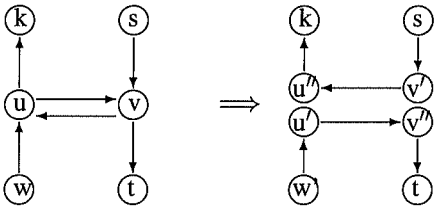


Figura 4. Trencament d'un H-subgraf.

En trencar un H-subgraf, eventualment es podria desconnectar el graf orientat original. Donat un g.o.e.s. G, denotem per H_G el nombre de components fortament connexes

(c.f.c.) que apareixen al graf orientat resultant de trencar tots els seus H-subgrafs. Convenim que $H_G = 1$ si G no té H-subgrafs. H_G està ben definit, ja que no depèn de l'ordre en què es produeix el trencament dels H-subgrafs, en no tenir dos H-subgrafs distints, vèrtexs intermedis comuns.

Teorema 4. *Tot tour eulerià en un g.o.e.s. G té almenys $H_G - 1$ girs en U en vèrtexs intermedis de H-subgrafs.*

Demostració. Suposem que existeix un tour eulerià en G amb k girs en U en vèrtexs intermedis de H-subgrafs, essent $k < H_G - 1$. Si trenquem cadascun dels H-subgrafs on el tour no ha realitzat un gir en U en un dels seus vèrtexs intermedis, no es desconnecta el graf G. Com que cada vegada que es trenca un H-subgraf s'afegeix com a molt una c.f.c. al nombre d'existents, si trenquem tots els H-subgrafs de G tindrem com a màxim $k+1$ c.f.c. amb $k+1 < H_G$, fet que és absurd per definició de H_G.

■

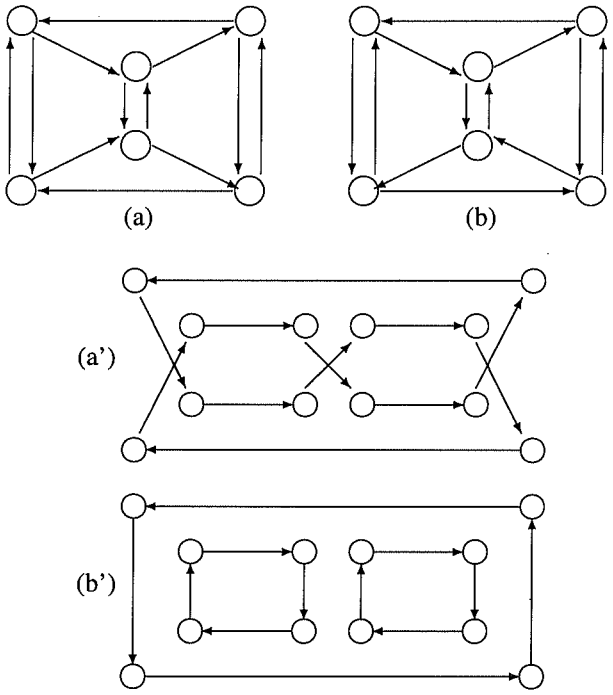


Figura 5. Trencament dels H-subgrafs de dos grafs orientats similars amb resultats diferents.

Per tant, tenim la condició necessària de graf orientat simple U-eulerià següent:

Corol·lari 5. *Si G és un graf orientat simple U-eulerià, aleshores $H_G = 1$.*

En la Figura 5 es constaten clarament els resultats del Teorema 4 per a dos grafs orientats eulerians pràcticament idèntics. Mentre que al graf orientat (a) $H_G = 1$, al graf orientat (b) $H_G = 3$. En trencar els H-subgrafs de (a) obtenim directament un tour U-eulerià. En canvi, és evident que tot tour eulerià en (b) realitzarà almenys dos girs en U per unir les tres c.f.c. resultants de trencar els seus H-subgrafs.

4. VÈRTEXS SENSE GIRS EN U

Suposat que un g.o.e.s. acompleix les tres condicions necessàries per a ésser U-eulerià esmentades a la secció anterior, veurem en aquesta secció en quins tipus de vèrtexs tenim la garantia que un tour eulerià amb mínim nombre de girs en U no hi realitzi girs en U. Per tal motiu necessitarem demostrar alguns resultats.

Teorema 6. *Sigui $G = (V,E)$ un g.o.e.s. i T un tour eulerià en ell. Existeix un tour eulerià T' en G que no realitza girs en U als vèrtexs i amb, bé $d^+(i) > 3$, bé $d^+(i) = 3$ i $U(i) = 1$, i que realitza els mateixos girs que T a la resta de vèrtexs.*

Demostració. Suposem que T realitza el gir en U $(j,i)(i,j)$ i que comença en aquest gir. Sigui $(u,i)(i,v)$ el gir següent per i de T. Considerem els dos casos de l'enunciat:

- Si $d^+(i) > 3$, existeix almenys un altre gir per i de T $(w,i)(i,k)$ amb $k \neq u$, que per l'ordre establert va després del $(u,i)(i,v)$. Si realitzem el 3-intercanvi

$$\{(j,i)(i,j),(u,i)(i,v),(w,i)(i,k)\} \rightarrow \{(j,i)(i,v),(u,i)(i,k),(w,i)(i,j)\}$$

  obtenim un altre tour eulerià sobre G sense almenys el gir en U $(j,i)(i,j)$.

- Si $d^+(i) = 3$ i $U(i) = 1$, sigui $(w,i)(i,k)$ el tercer gir de T per realitzar en i, aquí és evident que $k \neq u$, per tant, si realitzem el mateix 3-intercanvi que al cas anterior, obtenim un altre tour eulerià sobre G sense el gir en U $(j,i)(i,j)$.

Si repetim aquest procediment les vegades que necessitem, obtindrem un tour eulerià T' en G que acompleix les condicions del teorema.

■

Per altra banda, a la Figura 6 donem les diferents formes en què pot aparèixer un vèrtex v amb $d^+(v) = 2$ i $U(v) = 1$ en un g.o.e.s. sense vèrtexs finals:

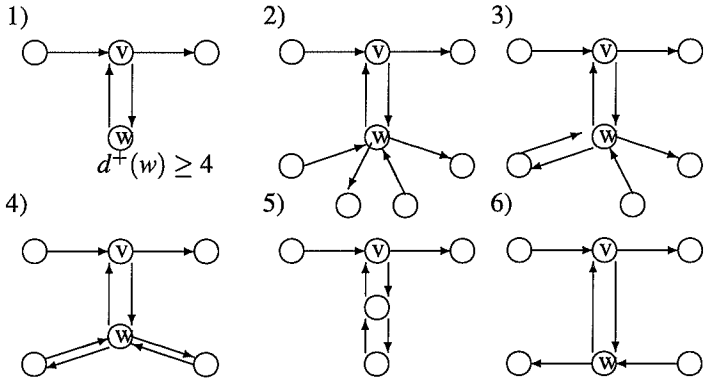


Figura 6. Formes en què pot apareixer un vèrtex v amb $d^+(v) = 2$ i $U(v) = 1$ en un g.o.e.s. sense vèrtexs finals.

En un tour eulerià amb mínim nombre de girs en U en un g.o.e.s., mai apareixerà un gir en U al vèrtex v als casos 1 i 2 de la Figura 6, ja que cas d'aparèixer en v un gir en U, si es creua amb l'altre gir per l'esmentat vèrtex, s'obtenen dos circuits que coincideixen al vèrtex w. A continuació, aquests dos circuits es poden unir en w i pel Teorema 6, un 3-intercanvi adequat faria desaparèixer els girs en U que es formaren en w.

Als casos 3 i 4 de la Figura 6, si apliquem la tècnica esmentada per als casos 1 i 2, si es formara un gir en U en w, un estudi detallat de totes les possibilitats ens diu que no tenim la garantia que un 3-intercanvi el fera desaparèixer.

El cas 5, ja que als vèrtexs interiors d'un camí maximal doble no es realitzen girs en U, es redueix a la resta de casos, depenent de l'altre vèrtex w, extrem del camí maximal doble que comença en v.

El cas 6 ha estat estudiat a la secció anterior.

D'allò que hem dit per als casos 1, 2 i 5 obtenim el resultat següent:

Teorema 7. *Sigui G=(V,A) un g.o.e.s. i sigui* $v = u_0, u_1, \ldots, u_{k+1} = w$ *un camí maximal doble en G, amb* $k \geq 0$, $d^+(v) = 2$, $U(v) = 1$ *i, bé* $d^+(w) \geq 4$, *bé* $d^+(w) = 3$ *i* $U(w) = 1$. *Cap tour eulerià amb mínim nombre de girs en U sobre G, realitza girs en U als vèrtexs d'aquest camí maximal doble.*

Per últim, el resultat següent ens permet eliminar més vèrtexs candidats a contenir girs en U en un tour eulerià sobre un g.o.e.s..

Teorema 8. *Sigui $G=(V,A)$ un g.o.e.s. i sigui $v \in V$ amb $d^+(v)=3$ i $U(v)>1$ tal que existeixen almenys U(v)-1 vèrtexs u connectats a v en G per un camí maximal doble, u acomplint que, bé $d^+(u) \geq 4$, bé $d^+(u)=3$ i $U(u)=1$, bé $d^+(u)=2$ i $U(u)=1$ si —en aquest últim cas— u és un vèrtex de tall. Cap tour eulerià amb mínim nombre de girs en U sobre G, realitza girs en U en v.*

Demostració. Suposem que un tour eulerià amb mínim nombre de girs en U sobre G realitza un gir en U en v. Si un 3-intercanvi no elimina aquest gir en U en v, almenys el transforma (si cal fer el 3-intercanvi) en un altre gir en U $(u,v)(v,u)$ amb u pertanyent a un camí maximal doble d'un dels dos primers casos esmentats a l'enunciat del teorema (el tercer cas no es pot donar ja que el vèrtex extrem del camí maximal doble a què pertany u és de tall). Si creuem aquest gir en U amb un altre gir per v, obtindríem dos circuits que passen per l'altre vèrtex extrem del camí maximal doble a què pertany u (eventualment aquest vèrtex és el propi u). Si creuem ambdós circuits en aquest vèrtex extrem obtindríem un tour eulerià, que si hi tingués girs en U, es podrien desfer mitjançant 3-intercanvis adequats. En qualsevol cas arribem a contradicció per obtenir un tour eulerià amb menys girs en U.

■

Per tant, un tour eulerià amb mínim nombre de girs en U sobre un g.o.e.s. G, suposat G sense vèrtexs finals, amb almenys un vèrtex u amb $d^+(u) \neq 2$ o $U(u) \neq 2$, $H_G = 1$ i trencats tots els seus H-subgrafs, no realitzarà girs en U als tipus de vèrtexs següents, no necessàriament excloents:

(1) vèrtexs v amb $U(v)=0$.

(2) vèrtexs v amb $d^+(v)>3$.

(3) vèrtexs v amb $d^+(v)=3$ i $U(v)=1$.

(4) vèrtexs v interiors d'un camí maximal doble.

(5) vèrtexs v amb $d^+(v)=2$ i $U(v)=1$ si v és vèrtex de tall (és evident).

(6) vèrtexs v amb $d^+(v)=2$ i $U(v)=1$ si v és vèrtex extrem d'un camí maximal doble que té per altre extrem un vèrtex del tipus (2) o (3).

(7) vèrtexs v amb $d^+(v)=3$ i $U(v)>1$ si existeixen almenys $U(v)-1$ vèrtexs connectats a v per un camí maximal doble i pertanyents a alguns dels tipus (2) (3) o (5).

Llavors, els únics vèrtexs on cal estudiar l'aparició de girs en U són aquells v amb $d^+(v) = 3$ i $U(v) > 1$ si no són del tipus (7). Això dóna lloc a la condició suficient de graf orientat simple U-eulerià següent:

Teorema 9. *Sigui G=(V,A) un g.o.e.s. sense vèrtexs finals, amb almenys un vèrtex u amb* $d^+(u) \neq 2$ *o* $U(u) \neq 2$ *i amb* $H_G = 1$. *Si tot vèrtex* $v \in V$ *amb* $d^+(v) = 3$ *i* $U(v) > 1$ *és del tipus (7), G és U-eulerià.*

En cas que un g.o.e.s. contingui vèrtexs v amb $d^+(v) = 3$ i $U(v) > 1$ que no siguin del tipus (7), és quan farem servir el resultat de Thomassen (Teorema 1) en una condició suficient que abraça l'anterior.

5. CONDICIÓ SUFICIENT DE GRAF ORIENTAT SIMPLE U-EULERIÀ

Sigui $G = (V,A)$ un g.o.e.s., li associem dos grafs no orientats G' i G'' definits com segueix:

– $G' = (V,E')$ on l'aresta $(u,v) \in E'$ si i sols si $(u,v)(v,u) \in A$ (cada aresta de G' substitueix un parell d'arcs inversos en G).

– $G'' = (V,E'')$ on l'aresta $(u,v) \in E''$ si i sols si u i v són vèrtexs consecutius en un camí (eventualment cicle) $(u_1,u_2,\ldots,u_k)$ a G tal que $d^+(u_1) = d^+(u_k) = 3$, $U(u_1) > 1$, $U(u_k) > 1$, si $k > 2$ $d^+(u_i) = U(u_i) = 2$ $i = 2,\ldots,k-1$, i tots els arcs d'aquest camí tenen invers (veure exemple en Figura 7).

Sigui $G = (V,A)$ un g.o.e.s., considerem també el conjunt M_G dels vèrtexs $v \in V$ que acompleixen:

1) $d^+(v) = 3$ i $U(v) > 1$.

2) v és vèrtex extrem d'almenys un camí maximal doble amb l'altre vèrtex extrem w acomplint $d^+(w) = 2$, $U(w) = 1$ i w no és vèrtex de tall. Sigui $k(v)$ el nombre d'aquests camins.

3) Existeix un nombre inferior a $U(v) - 1$ de vèrtexs u connectats a v per un camí maximal doble, u acomplint, bé $d^+(u) \geq 4$, bé $d^+(u) = 3$ i $U(u) = 1$, bé $d^+(u) = 2$ i $U(u) = 1$ si -en aquest últim cas- u es vèrtex de tall.

4) Si $k(v) = 1$ i la component connexa a què pertany v en G'' no conté cap altre vèrtex acomplint 1), 2) i 3), aleshores, aquesta component conté un cicle.

M_G és, per tant, un subconjunt del conjunt de vèrtexs v amb $d^+(v) = 3$ i $U(v) > 1$ no del tipus (7), prou restringit per les condicions 2) i 4).

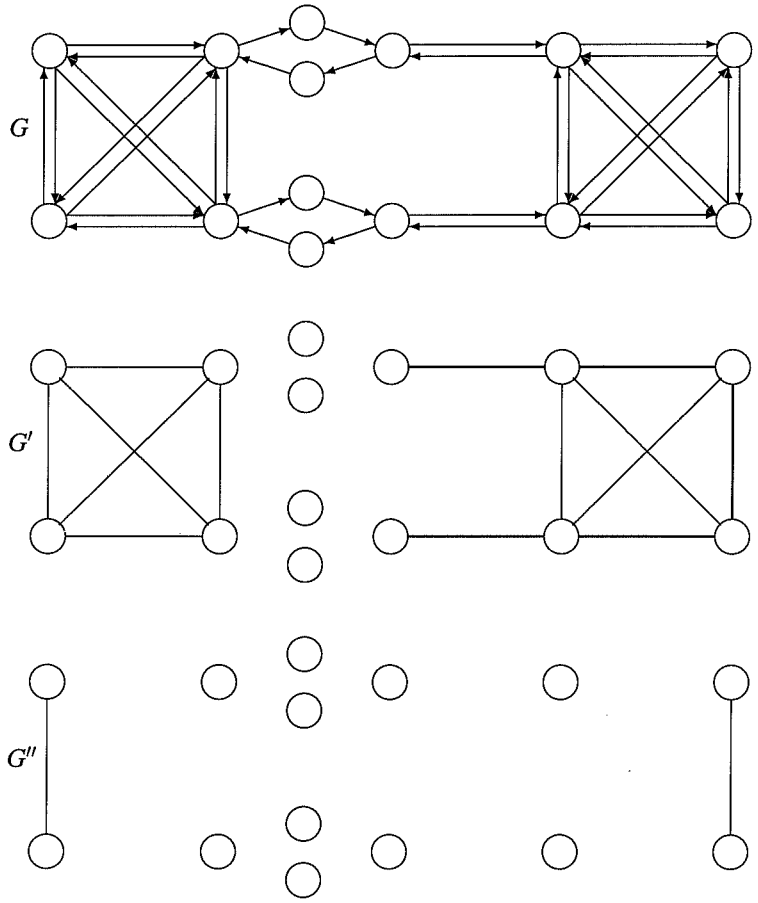


Figura 7

Teorema 10. *Sigui $G = (V,A)$ un g.o.e.s. sense vèrtexs finals, amb $H_G = 1$ i $M_G = \emptyset$. G és U-eulerià si cada component connexa G'_i de G' acompleix una de les dues condicions següents:*

(a) G'_i no és un cicle i admet un arbre generador T'_i tal que tota component connexa de $G'_i - T'_i$ conté un nombre parell d'arestes o un vèrtex de grau major que 3 en G'_i.

(b) Cada component connexa G''_{i_k} de G'' continguda en G'_i, a més de no ser un cicle, admet un arbre generador T''_{i_k} tal que tota component connexa de $G''_{i_k} - T''_{i_k}$ conté un nombre parell d'arestes.

Demostració. Cada component connexa G'_i (G''_{i_k}) de G' (G'') porta associada un subgraf G_i (G_{i_k}) de G generat per les parelles d'arcs inversos a les quals representen les seves arestes. Pel Teorema 1, si G'_i (G''_{i_k}) acompleix (a) ((b)), G_i (G_{i_k}) admet un tour eulerià T_i (T_{i_k}) que només realitza girs en U als seus vèrtexs finals.

Per altra banda, si G'_i no acompleix (a) però acompleix (b), per a cada G''_{i_k} continguda en G'_i, farem el següent: si existeix un camí maximal doble $u = u_0, u_1, \ldots, u_{r+1} = v$ amb $r \geq 0$ en G, acomplint que $u \in G''_{i_k}$ i u_t no pertanyent a cap G''_{i_t} $\forall t > 0$, siguin $w, s \in G''_{i_k}$ els vèrtexs adjacents a u en G''_{i_k} (eventualment $w = s$), ampliem T_{i_k} mitjançant la cadena $w, u, u_1, \ldots, v, u_k, \ldots, u, s$.

Una vegada realitzades totes aquestes ampliacions dels T_{i_k} als subgrafs G_{i_k} que ho precisen, es pot construir un tour eulerià T sobre G que absorbesca als T_j i als T_{i_k} afegint-li la resta d'arcs de G amb el mètode clàssic de recerca d'un circuit d'arcs no usats encara i creuament del mateix amb un altre circuit trobat anteriorment.

Per la forma de construir T, per allò esmentat a la secció anterior sobre els diversos tipus de vèrtexs i per les hipòtesis del teorema, creuaments de girs i 3-intercanvis adequats transformaran T en un tour eulerià que, com a màxim realitzarà girs en U als vèrtexs w que acompleixen $d^+(w) = 2$, $U(w) = 1$, w no és de tall i és vèrtex extrem d'un camí maximal doble amb l'altre vèrtex extrem v acomplint $d^+(v) = 3$, $U(v) > 1$ i v no és del tipus (7). Vegem que els girs en U en aquests vèrtexs també es poden eliminar.

En efecte, si T realitza un gir en U en un vèrtex w acomplint allò dit anteriorment, si creuem aquest gir amb l'altre gir per w, obtenim dos circuits que passen per v. Si creuem aquests dos circuits en v obtenim de nou un tour eulerià. Si aquest realitzés girs en U en v, amb un raonament similar al de la demostració del Teorema 8, amb 3-intercanvis i creuaments de girs adequats, podem, bé fer desaparèixer el gir en U, bé anar movent-lo a vèrtexs contigus. És evident que si no aconseguim que desapareixi, és perquè en anar movent-lo de vèrtex, s'ha format un cicle al subgraf de G'' a què pertany v o ha arribat a un vèrtex de les mateixes característiques que v. Això vol dir que $v \in M_G$, fet que és absurd ($M_G = \emptyset$).

Per tant, T es pot transformar en un tour U-eulerià.

■

De la mateixa demostració del Teorema 10 se segueix que, sota les hipòtesis d'aquest teorema, el tour U-eulerià es pot construir en temps polinomial.

Aquest teorema generalitza el Teorema 9 (les seves hipòtesis impliquen que G'' no conté cap cicle) i el Teorema 1 com a condició suficient (en aquest cas G' seria connex).

Si seguim l'exemple de la Figura 7, G' té dues components connexes. Mentre la component de la dreta acompleix la condició (a) del Teorema 10 (l'arbre es correspon amb les arestes més gruixudes), la de l'esquerra no acompleix (a), però G'' només té una component connexa continguda en ella, que per ser una aresta, és arbre i acompleix la condició (b). Així doncs, G és U-eulerià i un tour U-eulerià sobre ell ve donat en la Figura 8, tot seguint la numeració dels arcs. No entrem als detalls de la seva construcció, la qual fa ús de l'algorisme proposat per Benavent i Soler (1998a) per al cas particular esmentat a la secció 2.

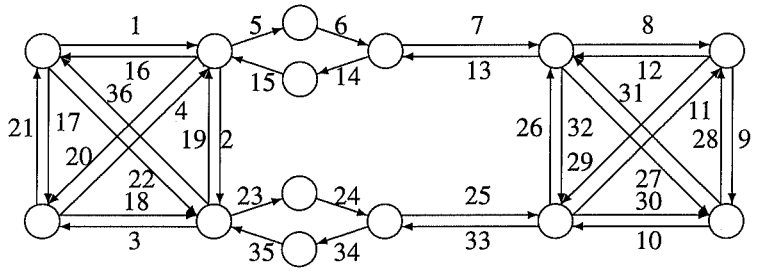


Figura 8. Tour U-eulerià sobre el graf orientat G de la Figura 7.

6. NOTES FINALS

En aquest article hem presentat el problema de trobar, si existeix, un tour eulerià sense girs en U (tour U-eulerià) sobre un graf orientat eulerià simple, el qual generalitza el problema polinomial del doble recorregut fort. Desconeguda la complexitat del nou problema, després d'un estudi detallat del diversos tipus de vèrtexs en un g.o.e.s., presentem al Teorema 10 una condició suficient de g.o.e.s. U-eulerià verificable en temps polinomial i que generalitza la donada per Thomassen (Teorema 1) per a l'existència d'un doble recorregut fort en un graf no orientat. A més a més, també donen una transformació del nou problema al Problema del Circuit Hamiltonià; d'aquesta manera, si no es compleixen les hipòtesis del Teorema 10, encara que en temps exponencial al pitjor del casos, podem saber si un g.o.e.s. és U-eulerià o no amb tècniques conegudes de Teoria de Grafs.

Encara que aquest article és netament teòric, pensem que els resultats i les tècniques aquí donades poden ser molt útils per evitar el major nombre possible de girs en U als vehicles, en problemes reals sobre xarxes de carrers o de carreteres, ja que com hem vist, si s'acompleixen les tres condicions necessàries donades a la secció 3, els girs en U només ocorrerien a determinats vèrtexs v amb $d^+(v) = 3$ i $U(v) > 1$, que si bé poden aparèixer en aquestes xarxes, tret de l'existència de vèrtexs pertanyents a M_G, poc probables en exemples reals, és molt difícil que es violi alguna de les condicions (a) o (b) del Teorema 10, i cas que es violés, l'algorisme per buscar un doble recorregut fort donat per Benavent i Soler (1998a) ens garanteix que (tot seguint el Teorema 10), si G'_i no acompleix (a) i G''_{i_k} no acompleix (b), donat un arbre generador T''_{i_k} de G''_{i_k}, només apareixeria un gir en U per cada component connexa de $G''_{i_k} - T''_{i_k}$ amb un nombre imparell d'arestes. També creiem que aquestes darreres components connexes són poc comuns als grafs provinents de problemes reals.

Pensant també en problemes reals, no volem deixar d'esmentar la possibilitat que el graf orient sigui simple, connex, però no eulerià i plantejar-nos, en aquest cas, la recerca d'un tour que recorri tots els seus arcs sense realitzar girs en U amb un cost extra mínim. Aquest problema és prou diferent al tractat aquí, ja que intervenen dos paràmetres: minimitzar costos al repetir arcs i evitar girs en U. Encara que el seu estudi teòric pugui ser interessant, aquest problema és un cas particular d'un altre amb més aplicacions pràctiques, anomenat Problema del Carter Rural Orientat amb Penalitzacions als Girs i Girs Prohibits i que consisteix a trobar un tour amb mínim cost que travessi determinats arcs evitant els girs prohibits (en particular girs en U). Aquest darrer problema ha estat estudiat per Benavent i Soler (1998b) tant des d'un punt de vista teòric com heurístic.

REFERÈNCIES

[1] **Benavent, E.** and **Soler, D.** (1998a). «Searching for a Strong Double Tracing in a Graph». *Top*, **6, 1**. 123-138.

[2] **Benavent, E.** and **Soler, D.** (1998b). «The Directed Rural Postman Problem with Turn Penalties». Pròxim a aparéixer a *Transportation Science*.

[3] **Brenner, J.L.** and **Lindon, B.C.** (1984). «Doubly Eulerian Trails on Rectangulars Grids». *Journal of Graph Theory*, **8**, 379-385.

[4] **Eggleton, R.B.** and **Skilton, D.K.** (1984). «Double Tracings of Graphs». *Ars Combinatoria*, **17A**, 307-323.

[5] **Fleischner, H.** (1983). «Eulerians Graphs». *Selected Topics in Graph Theory*, **2**, (Academic Press, London-New York) 17-53.

[6] **Gabow, H.N.** and **Stallmann, M.** (1986). «An Augmenting Path Algorithm for Linear Matroid Parity». *Combinatorica*, **6, 2**, 123-150.

[7] **Laporte, G.** (1992). «The Traveling Salesman Problem: An Overview of Exact and Aproximate Algorithms». *European Journal of Operational Research*, **59**, 231-247.

[8] **Ore, O.** (1951). «A Problem Regarding the Tracing of Graphs». *Elem. Math.*, **6, 3**, 49-53.

[9] **Thomassen, C.** (1990). «Bidirectional Retracting-free Double Tracing and Upper Embeddability of Graphs». *Journal of Combinatorial Theory, Series B*, **50, 2**, 198-207.

[10] **Troy, D.J.** (1966). «On Traversing Graphs». *Amer. Math. Monthly*, **73**, 497-499.

[11] **Xuong, N.H.** (1979). «How to Determine the Maximun Genus of a Graph». *Journal of Combinatorial Theory, Series B*, **26**, 217-225.

ENGLISH SUMMARY

EULERIAN TOUR WITHOUT U-TURNS IN A SIMPLE DIGRAPH

D. SOLER FERNÁNDEZ
Universitat Politècnica de Valencia*

Let $G = (V,A)$ be a simple directed Eulerian graph, here we study if G admits an Eulerian tour without U-turns, that is, without traversing consecutively pairs of arcs $(u,v),(v,u)$ $u,v \in V$. This problem can be solved by transforming it into a Hamiltonian circuit problem, but we do not know its complexity despite our efforts, so in this paper we give a condition under which we can find, in polynomial time, an Eulerian tour without U-turns in G. To give this condition, on the one hand, we generalize a result about a particular case solved in polynomial time (the one known as the strong double tracing problem), on the other, we try to identify vertices in which an Eulerian tour with a minimum number of U-turns, will make no U-turns.

Keywords: U-turn, U-Eulerian digraph, H-subgraph.

AMS Classification: 68R05,68R10

* Universitat Politècnica de València. Departament de Matemàtica Aplicada. Camí de Vera, 14. 46071 València.

- Received April 1997.
- Accepted Setember 1998.

Let $G = (V,A)$ be a simple Eulerian digraph, we study if it contains an Eulerian tour without U-turns (without traversing consecutively pairs of arcs $(u,v),(v,u)$, $u,v \in V$). Such a tour is called a U-Eulerian tour and if the answer is positive, we say that G is U-Eulerian.

We do not know the complexity of this problem, but we can solve it in exponential time by transforming it into a Hamiltonian Circuit Problem. The purpose of this paper is then, to give a simple sufficient condition that can be checked in polynomial time, to verify that a simple Eulerian digraph is U-Eulerian.

This condition makes use of a particular case solved in polynomial time, which occurs when the digraph G is the result of splitting each edge of a simple non directed graph G^* into two arcs of opposite direction. In this case, Thomassen (1990) proves that G is U-Eulerian iff G^* has no end-vertices and it contains a spanning tree T^* such that every connected component of $G^* - T^*$ has an even number of edges or contains a vertex v with $d(v) > 3$ in G^*. Checking the existence of tree T^* and constructing (if such be the case) the U-Eulerian tour, takes polynomial time.

Let $v \in V$, we denote by $U(v)$ the number of pairs of arcs $(u,v),(v,u)$, with u adjacent to v. We will suppose without loss of generality that G does not have vertices v with $d^+(v) = U(v) = 2$.

Let $u,v \in V$ with $d^+(v) = d^+(u) = 2$, $U(u) = U(v) = 1$ and $(u,v),(v,u) \in A$, we call H-subgraph, the subgraph of G formed by u,v, its incident arcs and its adjacent vertices. Breaking this H-subgraph means splitting u and v into two respective pairs of vertices $u',u"$ and $v',v"$ such that if $(u,v),(v,u),(u,k),(w,u),(s,v),(v,t) \in A$, we obtain two paths in G without common interior vertices $(s,v',u",k)$ and $(w,u',v",t)$. Let H_G be the number of connected components of G obtained by breaking all its H-subgraphs, if $H_G \neq 1$, then G is not U-Eulerian.

If $H_G = 1$ and G has no final vertices (vertices v with $d^+(v) = U(v) = 1$), then the only vertices that determine if an Eulerian tour in G with a minimum number of U-turns is not U-Eulerian, form a set of vertices v with $d^+(v) = 3$ and $U(v) > 1$ and, besides, there is a subset of such vertices, denoted by M_G which contains "very bad"vertices if U-turns are to be avoided.

Let $G' = (V,E')$ be a non directed graph such that $(u,v) \in E'$ iff $(u,v)(v,u) \in A$ and let $G" = (V,E")$ be another non directed graph such that $(u,v) \in E"$ iff $d^+(u) = d^+(v) = 3$, $U(u) > 1$, $U(v) > 1$ and $(u,v),(v,u) \in A$. The following theorem generalizes the result of Thomassen as a sufficient condition for a U-Eulerian simple digraph and can also be checked in polynomial time:

Theorem. *Let $G = (V,A)$ be a simple Eulerian digraph without final vertices, $H_G = 1$ and $M_G = \emptyset$. G is U-Eulerian if each connected componet G'_i of G' verifies one of the two following conditions:*

(a) G'_i is not a cycle and it contains a spanning tree T'_i such that each connected component of $G'_i - T'_i$ contains an even number of edges or a vertex of degree greater than 3 in G'_i.

(b) Every connected component G''_{i_k} of G'' contained in G'_i is not a cycle and it has a spanning tree T''_{i_k} such that each connected component of $G''_{i_k} - T''_{i_k}$ contains an even number of edges.

Estadística Oficial

QÜESTIIÓ, vol. 22, 3, p. 493-510, 1998

LA VERIFICACIÓ ALEATÒRIA: UNA ESTRATÈGIA PER MILLORAR I AVALUAR LA QUALITAT DE L'ENTRADA DE DADES

J.M. DOMÈNECH MASSONS
J.M. LOSILLA VIDAL
M. POTELL VIDAL
Universitat Autònoma de Barcelona*

S'aborda la problemàtica de la reducció dels errors que es produeixen durant la introducció de les dades i que no es poden controlar mitjançant proteccions automàtiques. Enfront aquest problema, l'estratègia habitual és la «doble entrada» (DE) de les dades, la qual augmenta considerablement el cost de la investigació. Com alternativa a aquesta estratègia es proposa un nou procediment, implementat en el Sistema DAT, que es basa en un procés de «verificació aleatòria» (VA) d'un percentatge del total de dades. A més de reduir el cost, la VA ofereix altres avantatges com el fet de proporcionar una estimació del percentatge d'errors, i oferir un índex d'aptitud i eficàcia dels operadors. A la segona part de l'article es presenten els resultats d'un experiment que recolza la hipòtesi que la VA augmenta l'eficàcia de l'entrada de dades, sense minva de l'eficiència, quan es compara amb situacions en les quals no s'aplica cap control que permeti obtenir un indicador sobre la qualitat de les dades que s'introdueixen.

Random verification: a strategy to improve and assess the quality of data entry.

Paraules clau: Qualitat de les dades; doble entrada; entrada de dades; verificació aleatòria; Sistema DAT.

Clasificación AMS: 62D99

* Dept. de Psicobiologia i de Metodologia de les Ciències de la Salut, Universitat Autònoma de Barcelona.
Investigació realitzada gràcies a l'ajut DGICYT No. PM95-126 del Ministeri d'Educació i Ciència.
Correspondència: J.M. Domènech. Laboratori d'Estadística Aplicada i de Modelització. Universitat Autònoma de Barcelona. Apartat de Correus 40. 08193 Bellaterra. e-mail:josep.m.domenech@uab.es.

- Rebut el març de 1997.
- Acceptat l'octubre de 1998.

INTRODUCCIÓ

En els darrers anys s'ha produït un gran avenç en les aplicacions informàtiques encaminades a facilitar el procés de dades, però es qüestiona si realment aquest tipus d'avenços reverteixen en un increment de la qualitat dels treballs que es publiquen (Cobos, 1995; Freedland i Carney, 1992; Ronel, Varley i Webb, 1993). La falta de coneixement d'alguns procediments que aquestes aplicacions automatitzen, unit a la confiança excessiva en què els resultats són correctes pel sol fet que els presenta un ordinador, poden influir negativament en la qualitat de la informació que es publica.

Les sigles GIGO («Garbage In - Garbage Out») recorden els perills que comporta aquesta situació: si s'entren dades incorrectes en una aplicació informàtica, per exemple per realitzar una anàlisi estadística, l'aplicació no deixarà d'oferir resultats encara que aquests siguin incorrectes. Una expansió més recent de GIGO és «Garbage In, Gospel Out», que és un comentari sarcàstic sobre la tendència a dipositar una confiança excessiva en les dades elaborades per l'ordinador (Dumbill, 1994).

L'«efecte GIGO» es produeix amb més freqüència en investigacions amb grans volums de dades (estudis multicèntrics, longitudinals amb molts seguiments, etc.), degut a la dificultat que representa per l'investigador controlar el correcte funcionament de les mils d'operacions que realitza amb les seves dades a través de diferents aplicacions informàtiques (Barton, Hatcher, Schurig i Marciano, 1991; Moritz *et al.*, 1995).

Ningú discuteix la importància del procés de dades cara a la validesa de les conclusions d'una investigació. No obstant això, hom acostuma a pensar només en l'anàlisi estadística, que és una de les etapes més emblemàtiques però que està molt lluny de ser la única. En aquest treball conceptualitzem el procés de dades com una activitat complexa, les principals etapes d'actuació de la qual són:

1) Disseny de la base de dades (estructura de les taules, proteccions d'entrada, etc.).
2) Entrada de les dades.
3) Preparació de la matriu de dades (generació de variables, selecció de casos, creació de taules rectangulars de dades, etc.).
4) Anàlisis de les dades.

Centrant-nos en les dues primeres etapes del procés de dades, els referents per avaluar la qualitat de les dades són: (*a*) estar exemptes d'errors; (*b*) no contenir valors omesos («missing»), i (*c*) ajustar-se als moments temporals de registre establerts en el disseny (cas d'estudis longitudinals).

Existeixen diferents aspectes que poden afectar la qualitat de les dades i, per tal d'incidir sobre ells, considerem convenient establir un primer criteri de demarcació que vindria donat per la possibilitat d'exercir un control automàtic.

Els *controls automàtics de qualitat* poden expressar la *consistència* de forma *matemàtica* donant lloc a un error, o de forma *probabilística* donant lloc a un avís. Així, un valor fora de rang (p.e. talla d'un adult igual a 7cm.), la inconsistència entre variables que mantenen relacions lògiques (p.e. sexe masculí i nombre d'embarassos diferent de «no aplicable») i el control de valors omesos, donarien lloc a *errors*, mentre que una categoria, una combinació de categories, una relació lògica o un rang de valors poc probables (p.e. adult amb talla superior a 199 cm.) donarien lloc a *avisos*.

Els *errors no controlables de manera automàtica* són aquelles unitats d'informació que compleixen les condicions de rang, lògiques i de valor «missing» anteriors, però que no coincideixen amb el valor original. Aquest tipus d'errors es poden *limitar* incorporant factors que incideixin sobre aspectes atencionals, perceptuals i motivacionals de l'operador encarregat d'introduir les dades (Henning, Sauter, Salvendy i Krieg, 1989; Pocius, 1991; González, 1993; Lalomia, i Sidowski, 1993). Les investigacions realitzades sobre aquests factors s'emmarquen dins l'àmbit d'estudi multidisciplinar de l'«Interacció Home-Ordinador» o HCI («Human Computer Interaction») (Dillon, 1983; Card, Moran i Newell, 1983; Schneiderman, 1992; Johnson, 1992; Carroll, 1993; Wallace i Anderson, 1993).

Una característica dels estudis realitzats en el context de l'HCI és que solen ser específics respecte al tipus d'aplicació informàtica i molt generals pel que fa als objectius que es desitgen aconseguir (p.e. satisfacció, comoditat, eficàcia, eficiència, etc.). En aquest àmbit la major part de treballs s'han centrat en el disseny de la interfície d'usuari, tractant aspectes com els sistemes gràfics orientats a la tasca, la disposició i el contingut dels menús, els tipus de «controls» o mecanismes que l'usuari pot utilitzar per indicar les accions que desitja realitzar, etc. (Wolf, 1992; Croquet, 1994; Howes, 1994; Karwowski, Eberts, Salvendy i Noland, 1994).

Aquest treball se centra en les aplicacions d'entrada de dades i, dins d'aquestes, en la problemàtica de la reducció dels errors d'operador no controlables de forma automàtica, mitjançant un mecanisme anomenat «verificació aleatòria» (VA). La VA forma part d'un conjunt d'aportacions que es deriven del Sistema *DAT* (Domènech i Losilla, 1995). Seguidament es presenten les característiques generals del Sistema, a continuació s'ubica la VA en relació amb altres estratègies que persegueixen objectius similars i, per últim, es descriu un estudi experimental encaminat a avaluar l'eficàcia diferencial de la VA enfront d'altres procediments.

CARACTERÍSTIQUES GENERALS DEL SISTEMA *DAT*

El Sistema *DAT* forma part d'un ampli projecte l'objectiu del qual és desenvolupar un sistema eficient pel procés de dades científiques que permeti el control (des de la seva captura fins a la seva preparació i exportació al sistema d'anàlisi estadística) de tots

els aspectes que afecten a la qualitat de les dades. La Figura 1 ubica, en el context del procés de dades, els tres mòduls que el formen (*EnDat/Lab*, *EnDat* i *ExperDat*).

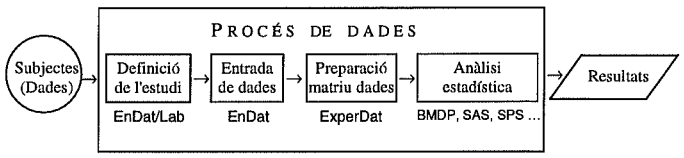


Figura 1. Etapes d'un procés de dades.

S'està preparant la versió comercial del Sistema DAT per a Microsoft Windows (3.1, 3.11, '95 i NT), a partir de les versions beta que des de fa 5 anys s'estan utilitzant de forma intensiva per a la creació, captura i gestió de bases de dades en quatre àmbits: *administració pública* (accidents laborals de Catalunya), *indústria farmacèutica* (assaigs clínics d'un important grup farmacèutic), *laboratori universitari d'estadística aplicada* (estudis per a medicina, psicologia i sociologia) i *investigació epidemiològica* (línia d'investigació sobre epidemiologia psiquiatria infantil, subvencionada amb els ajuts DGICYT PM91-209 i PM95-209).

Les principals característiques del Sistema *DAT* es poden sintetitzar exposant el seu ús en l'esmentada línia d'investigació sobre epidemiologia, que es caracteritza pel registre d'un gran nombre de variables que es corresponen amb els nombrosos ítems de cada instrument d'avaluació clínica. Per tal d'il·lustrar la complexitat i l'enorme volum de dades a manipular, així com la flexibilitat del Sistema DAT per dur a terme la seva gestió, n'hi ha prou amb dir que un d'aquests instruments diagnòstics és l'entrevista clínica estructurada DICA-R, amb més d'un miler de preguntes, que s'administra al cas d'estudi, als seus pares (per recaptar més informació sobre el subjecte) i a cadascun dels seus germans. A més, l'entrevista DICA-R es realitza en dues ocasions (test i retest) i algunes vegades es graven tant les respostes registrades per l'entrevistador com les registrades per un o varis observadors.

El mòdul *EnDat/Lab* permet especificar el disseny relacional de les dades d'aquest estudi, així com incloure totes les proteccions necessàries durant l'entrada de dades, els salts que guien l'entrevistador sobre les preguntes a efectuar, i també definir els camps virtuals amb els diagnòstics DSM-IV, parcials i finals, necessaris per administrar correctament l'entrevista. L'*EnDat/Lab* genera un conjunt d'arxius que contenen aquestes especificacions, i que seran llegits pel mòdul *EnDat*, encarregat de proporcionar la interfície d'usuari per dur a terme l'entrada de dades. Per últim, el mòdul *ExperDat* permet generar noves variables, definir fàcilment *vistes* que reorganitzin l'estructura relacional de la base de dades i preparin matrius rectangulars, les

quals poden implicar fins i tot una nova definició del *cas d'estudi* per adaptar-la al *cas d'anàlisi* d'un determinant objectiu d'investigació. Així, les entrevistes de cada germà, per exemple, a cops es poden considerar rèpliques de l'entrevista del subjecte, mentre que altres vegades constitueixen entrevistes de diferents subjectes (de la Osa, Ezpleleta, Doménech, Navarro i Losilla, 1996).

El Sistema *DAT* és també una plataforma idònia per la recerca en l'àmbit de la HCI, especialment pel que fa a: (*a*) detecció dels aspectes que causen error i/o redueixen la velocitat del treball dels operadors; (*b*) avaluació de noves alternatives que evitin els errors i incrementin l'eficiència, entre les quals es troba la VA, i (*c*) desenvolupament d'instruments indicatius de la destresa dels operadors, que poden facilitar la fase de selecció de personal.

LA VERIFICACIÓ ALEATÒRIA COM ALTERNATIVA O COMPLEMENT A LA DOBLE ENTRADA

Com ja dèiem a la introducció, les implicacions de l'«efecte GIGO» s'aguditzen quan els estudis són d'una certa magnitud. En aquests casos és imprescindible disposar d'un indicador de la qualitat de les dades, el qual és impossible d'obtenir si es realitza únicament una «entrada simple» (SE). En front aquest problema, l'alternativa que es planteja habitualment és la «doble entrada» (DE) de les dades, que implica que s'introdueixi dues vegades cada registre. Aquesta estratègia es concep com l'única que assegura que les dades estiguin lliures d'errors abans de procedir a la seva anàlisi, al mateix temps que permet obtenir un índex de la qualitat del treball dels operadors (Blumenstein, 1993; Gassman, Owen, Kuntz, Martin i Amoroso, 1995; Gibson, Harvey, Everett i Parmar, 1995; Wolf, 1993). S'ha de tenir en compte que si es desitja aprofitar aquest avantatge abans de finalitzar la recollida de dades (cas especialment important en els estudis longitudinals), o si els qüestionaris en paper només poden estar en el lloc on s'introdueixen les dades durant un temps limitat (com succeeix, per exemple, quan existeixen motius de seguretat), la DE no es pot realitzar en diferit, sinó que s'ha d'anar aplicant durant el procés d'entrada de dades (al finalitzar un bloc de registres, al finalitzar el dia, per un operador que fa les funcions de supervisor del què realitza la primera entrada, etc.).

Per tant, el principal desavantatge que comporta l'aplicació de la DE és, sens dubte, el seu elevat cost, tant econòmic com de temps. Per aquest motiu, en la pràctica, només els estudis que compten amb recursos econòmics suficients (p.e. els assaigs clínics) poden aplicar la DE, quedant la majoria d'estudis desproveïts de mecanismes que assegurin la qualitat de les dades, i exposats per tant a l'efecte GIGO.

Enfront d'aquesta situació, Domènech i Losilla (1995) proposen una estratègia de «verificació aleatòria» (VA) de les dades, que consisteix en la doble entrada d'un

percentatge del total de dades, elegides a l'atzar pel sistema informàtic i sol·licitades amb un *retard* suficient per tal d'evitar efectes de record, i en la reproducció dels errors comesos per l'operador durant la primera entrada.

El principal avantatge de la VA és que redueix considerablement el cost de l'entrada de dades mantenint gran part de les característiques positives que s'atribueixen a la DE realitzada en condicions òptimes. És a dir, la VA permet obtenir, en qualsevol moment del procés d'entrada de dades, una estimació del *percentatge d'errors* de les dades i un *índex de l'aptitud i de l'eficàcia dels operadors*.

Un aspecte a destacar aquí és que la VA és compatible amb la DE, perquè es pot utilitzar la VA durant la fase de recollida de dades, i al finalitzar aquesta fase completar la DE introduint els registres restants. Tanmateix, la VA implica necessàriament que sigui el mateix operador qui realitzi la reeintroducció de les dades; no obstant això, té l'avantatge de proporcionar a l'operador un «feed-back» dels errors que es van produint.

També és possible combinar la DE amb la VA en estudis amb estrats de subjectes que per les seves característiques requereixen diferents nivells de control. Per exemple, en l'anomenat registre anual d'accidents laborals, convé aplicar DE o VA del 100% a tots els comunicats d'accidents mortals i greus (aprox. 2.000) i utilitzar un percentatge de VA molt més baix pels accidents lleus (aprox. 160.000), reduint d'aquesta forma el cost global del registre, proporcionant un índex de la seva qualitat i contribuint a aconseguir l'estàndard de qualitat establert.

Aquest «feed-back» de la VA ens permet plantejar la primera hipòtesi del nostre estudi:

Hipòtesi 1: *El «feed-back», inherent a la situació de VA, pot contribuir positivament a incrementar l'eficàcia global del procés d'introducció de dades, és a dir, el nombre d'errors comesos serà menor en les situacions d'entrada de dades en les quals l'operador rep «feed-back».*

Des d'un punt de vista pràctic i atenent només al cost econòmic de l'entrada de dades, també té interès comparar l'*eficiència* dels operadors en situacions en què no rebin «feed-back» sobre els errors que cometen, amb la seva eficiència en situacions en què sí el reben. Però la comparació dels nivells d'eficiència implica contrastar prèviament la següent hipòtesi:

Hipòtesi 2: *El temps que s'inverteix en la introducció de qüestionaris de dades sense errors no és superior al que s'inverteix en la introducció de qüestionaris de dades amb errors, independentment de si la situació comporta o no «feed-back» per a l'operador.*

En efecte, una diferència en el sentit contrari a l'expressat per aquesta hipòtesi, és a dir, major temps requerit per introduir dades sense errors que dades amb errors, implicaria que un augment en l'eficàcia comporta necessàriament una disminució en l'eficiència i, en conseqüència, que no tingui cap valor comparar les eficiències dels operadors en les diferents situacions d'entrada de dades, sinó únicament els seus nivells d'eficàcia. Si, pel contrari, les dades no contradiuen la hipòtesi 2, llavors contrastarem la següent hipòtesi sobre l'eficiència:

Hipòtesi 3: *El «feed-back» associat a la VA no comporta una minora en l'eficiència, és a dir, el nombre de dades per unitat de temps introduïdes sense cometre errors és igual o major en les situacions en les quals l'operador rep «feed-back» que en les que no en rep.*

MÈTODE

Subjectes

Els subjectes són 45 estudiants voluntaris de primer curs de Psicologia de la Universitat Autònoma de Barcelona, amb edats compreses entre els 18 i els 20 anys i un 67% de dones, que s'han dividit a l'atzar en tres grups de mida 15.

Als subjectes se'ls ha ofert, com a recompensa per la seva participació, una llicència d'ús d'un programari informàtic comercial valorat en deu mil pessetes, així com l'accés a l'aula de microinformàtica del nostre laboratori durant el proper curs acadèmic.

S'ha considerat com a condició d'admissió que el subjecte tingui un nivell suficient d'habilitat en l'ús d'un ordinador personal (equivalent al necessari per escriure un document simple amb un programari de tractament de textos).

Material

S'han omplert a mà 100 qüestionaris amb dades fictícies i, per gravar-los, s'ha construït amb el Sistema *DAT* un formulari informatitzat sense proteccions per rangs, condicions lògiques, etc., la qual cosa permet introduir qualsevol valor. Per incrementar la taxa d'errors, s'ha dissenyat la pantalla de manera que la disposició dels camps de captura no coincideixi amb els qüestionaris impresos. L'experiment s'ha realitzant en el nostre laboratori, en grups de 15 subjectes, utilitzant ordinadors PC-486/66, monitor color de 14" i tarja gràfica VGA.

Disseny

S'ha aplicat un disseny de grups a l'atzar amb registre pretest. La variable independent és el tipus de situació d'entrada de dades assignada a cada grup: «verificació aleatòria» (VA), «placebo» (PL) i «control» (CO).

La *condició VA* es caracteritza per una consigna verbal informant que el programari controlarà els errors comesos durant l'entrada de dades i obligarà a repetir la introducció d'alguns formularis. El percentatge de formularis a verificar es fixa en un 15% i el retard en la doble entrada d'aquest percentatge de casos s'estableix en dos formularis.

La condició *PL* es caracteritza només per una consigna verbal informant que durant l'entrada de dades el programari controlarà el nombre d'errors comesos. Aquesta condició és necessària per distingir els efectes específics del «feed-back» que comporta la VA de l'efecte degut a la coneixença que s'estan registrant els errors comesos.

En la *condició CO* els subjectes no reben cap consigna.

Hem definit quatre variables dependents: *mesura de l'eficàcia* (proporció de camps de dades introduïts sense error), *mesura de l'eficiència* (nombre de registres de dades o de qüestionaris per hora de treball introduïts sense error), *temps mitjà per introduir els qüestionaris amb error* (en segons) i *temps mitjà per introduir els qüestionaris sense error* (en segons). El temps de la segona entrada dels qüestionaris revisats mitjançant VA no ha estat considerat perquè aquest temps no existeix en les altres dues condicions.

Procediment

Cada grup de 15 subjectes ha introduït a l'ordinador les dades del conjunt de qüestionaris, en una sessió que ha durat aproximadament dues hores. En la fase de reclutament la tasca a realitzar no es presenta com un experiment sinó com una col·laboració en un treball d'investigació real. Un cop finalitzada la recollida de dades s'ha comunicat als participants els objectius reals de l'estudi (atenent a l'article 34/89 del codi de deontologia del Col·legi de Psicòlegs de Catalunya).

La sessió s'ha estructurat en tres fases: entrenament, línia base i experimental. La *fase d'entrenament* té com a objectius: (1) informar els subjectes de quina serà l'operativa a seguir per entrar els qüestionaris a l'ordinador, (2) anticipar que està previst realitzar parades durant la sessió; (3) incentivar la rapidesa de l'entrada de dades, oferint un premi addicional als subjectes que introdueixin un major nombre de qüestionaris durant la sessió; i (4) que el subjecte practiqui amb el programari informàtic utilitzat

per l'entrada de dades. Les dades obtingudes en aquesta fase no s'han incorporat en les anàlisis dels resultats de l'estudi.

Passats els 15 minuts que ha durat la fase d'entrenament, es realitza una breu parada i, a continuació, s'inicia la *fase de línia base*, que té una durada de 45 minuts i en la qual no es realitza cap mena d'indicació als subjectes.

Finalitzada aquesta fase, s'inicia un nou descans en el qual es donen les consignes verbals corresponents a cadascuna de les tres condicions. La fase experimental té una durada de 60 minuts per als grups CO i PL, i de 75 minuts pel grup VA.

RESULTATS

Per tal de contrastar la primera hipòtesi, relativa a l'efecte del «feed-back» associat a la situació de VA sobre l'eficàcia global de l'operador (EFICACIA), s'ha ajustat un model de regressió considerant com a variable independent la situació d'entrada de dades (GRUP) descomposta en dues variables fictícies codificades respecte a la categoria de referència VA, i incloent com a variable de control el nombre de camps de dades introduïdes sense error durant la línia base (LBEFICA) juntament amb els corresponents termes d'interacció perquè no es pot descartar a priori la seva existència:

$$\begin{aligned} \text{EFICACIA} &= \alpha + \beta_1 \times \text{GRUP}_{\text{CO}} + \beta_2 \times \text{GRUP}_{\text{PL}} + \gamma \times \text{LBEFICA} + \\ &+ \delta_1 \times \text{GRUP}_{\text{CO}} \times \text{LBEFICA} + \delta_2 \times \text{GRUP}_{\text{PL}} \times \text{LBEFICA} + \varepsilon \end{aligned}$$

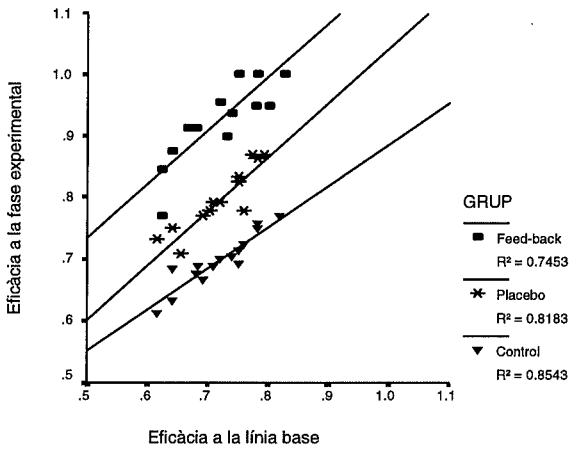


Figura 2. Relació entre l'eficàcia durant l'entrada de dades en la fase de línia base i en la fase experimental.

Els resultats de l'anàlisi (Figura 2) posen de manifest que la interacció entre GRUP i LBEFICA no és estadísticament significativa ($F(2, 39) = 1.03$; $p = 0.37$), motiu pel qual hem estimat un nou model sense interacció. En aquest nou model el terme d'ajust (línia base) produeix canvis molt poc importants en els coeficients b, però s'ha retingut perquè millora la precisió de les seves estimacions (ambdós errors estàndard disminueixen de 0.02 a 0.01).

Els resultats de l'anàlisi indiquen un *increment d'eficàcia* perquè en el grup VA s'han introduït un 27.7% (IC95%: 20.7 a 24.7) més de camps de dades sense error que en el grup CO, i un 13.1% (IC95%: 11.1 a 15.0) més de dades sense error que en el grup PL. Aquests resultats són congruents amb la hipòtesi plantejada, ja que posen de manifest l'efecte positiu sobre l'eficàcia global del procés d'introducció de dades degut al «feed-back» inherent a la situació de VA, distingint aquest efecte de l'atribuïble al fet que els subjectes puguin tenir coneixement (grup PL) que s'estan registrant els seus errors.

Per tal de contrastar la segona hipòtesi, referida a la igualtat entre els temps requerits per a la introducció de les dades sense errors respecte a les dades amb errors, s'han comparat els temps mitjans amb una prova t per a mostres relacionades. Els resultats de l'anàlisi indiquen que si bé el temps mitjà d'introducció de dades amb error en el grup VA ha incrementat significativament en 0.5 seg. ($p = 0.02$), aquest augment manca d'importància pràctica (IC95%: 0.1 a 0.9). Per tant, aquest resultat justifica el plantejament de la tercera hipòtesi.

Per tal de contrastar la tercera hipòtesi sobre l'efecte del «feed-back» inherent a la situació de VA sobre l'eficiència, s'ha procedit a estimar l'anterior model de regressió utilitzant en aquest cas com a variable dependent la mesura de l'eficiència.

Els resultats de l'anàlisi posen de manifest que la interacció entre el grup i la eficiència durant la línia base no és estadísticament significativa ($F(2, 39) = 0.63$; $p = 0.54$). En el model final s'ha inclòs el terme control (línia base) perquè, malgrat no es produeixen canvis apreciables en els coeficients de regressió que mesuren els efectes, millora la seva precisió. Els resultats d'aquesta última anàlisi són congruents amb la hipòtesi de manteniment de l'eficiència perquè el temps mitjà d'introducció de dades sense error, encara que és lleugerament superior en el grup VA respecte al del grup CO, comporta una diferència que no és pràcticament important (d=0.8 seg.; IC95%: (-0.5 a 2.1), ni estadísticament significativa (p=0.22). El mateix passa en el grup PL (d=0.3 seg.; IC95%: (-1.0 a 1.6; p=0.64).

DISCUSSIÓ

Els resultats trobats indiquen que la VA augmenta l'eficàcia del procés d'entrada de dades, sense minvar la seva eficiència, quant es compara amb situacions en les quals

no s'aplica cap control que permeti obtenir un indicador sobre la qualitat de les dades que s'introdueixen. A més, els resultats obtinguts al comparar el grup de VA amb el grup placebo, recolzen la hipòtesis plantejada sobre l'efecte específic del «feed-back» inherent a la situació de VA com a mecanisme explicatiu de la reducció de l'error.

Encara que l'estratègia de VA, com succeeix amb la DE, té un cost de temps més elevat que la SE, les dades aportades en el present estudi suggereixen que aquest increment del cost no és motiu suficient per excloure controls d'aquest tipus durant la introducció de les dades. Segons la nostra opinió, i per evitar en part l'«efecte GIGO», mai s'hauria d'analitzar unes dades de les quals no es disposi, com a mínim, de l'estimació de l'error que contenen, motiu pel qual considerem de gran interès pràctic els resultats trobats sobre l'eficàcia i l'eficiència de l'aplicació de controls mitjançant VA.

Per últim, cal destacar també que el Sistema *DAT* es mostra com una plataforma molt útil per a implantar models de laboratori vàlids, adreçats a avaluar l'impacte d'altres mètodes de reducció dels errors comesos pels operadors durant l'entrada de dades.

REFERÈNCIES

[1] **Barton, C., Hatcher, C., Schurig, K.** i **Marciano, P.** (1991). «Managing data entry of a large-scale interview project with optical scanning hardware and software». 20th Annual Meeting of the Society for Computers in Psychology (1990, New Orleans, Louisiana). *Behavior Research Methods, Instruments, and Computers*, **23 (2)**, 214-218.

[2] **Blumenstein, B.A.** (1993). «Verifying keyed medical research data». *Statistics in Medicine*, **12**, 1535-1542.

[3] **Card, S.K., Moran, T.P.** i **Newell, A.** (1983). *The psychology of human computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

[4] **Carroll, J.M.** (1993). «Creating a design science of human-computer interaction». *Interacting with Computers*, **5(1)**, 3-12.

[5] **Cobos, A.** (1995). «El síndrome "GIGO"». *JANO*, **49(1123)**, 481-482.

[6] **Croquet, M.** (1994). *PC y Robótica. Técnicas de interfaz.* Madrid: RA-MA.

[7] **Col·legi de Psicòlegs de Catalunya.** (1989). *Código deontológico.*

[8] **de la Osa, N., Ezpeleta, L., Doménech, J.M., Navarro, J.B.** i **Losilla, J.M.** (1996). «Fiabilidad entre entrevistadores de la DICA-R». *Psicothema*, **6(2)**, 359-368.

[9] **Dillon, R.F.** (1983). «Human factors in user-computer interaction: An introduction». *Behavior Research Methods, Instruments and Computers*, **15**, 195-199.

[10] **Domènech, J.M.** i **Losilla, J.M.** (1995). *Sistema DAT: gestor de datos científicos. Manual de referencia.* Campus de Bellaterra, Barcelona: Laboratori d'Estadística Aplicada i de Modelització. Universitat Autònoma de Barcelona.

[11] **Dumbill, E.** (1994). *Jargon Lexicon Main Index.* Diccionari en format hipertexte que es pot obtenir a Internet en la següent direcció: http://www-i5.informatik.rwth-aachen.de/mbp/jargon300/main.html.

[12] **Gassman, J.J, Owen, W.W., Kuntz, T.E, Martin, J.P.** i **Amoroso, W.P.** (1995). «Data quality assurance, monitoring, and reporting». *Controlled Clinical Trials*, **16 (2 Suppl.)**, 104S-136S.

[13] **Gibson, D., Harvey, A.J., Everett, V.** i **Parmar, M.K.** (1995). «Is double data entry necessary? The CHART trials. CHART Steering Committee. Continuous, Hyperfractionated, Accelerated Raiotherapy». *Controlled Clinical Trials*, **15 (6)**, 482-488.

[14] **González, H.** (1993). «Psicología de las interfaces: usuario-sistema: teorías y métodos». *Revista Intercontinental de Psicología y Educación*, **6(1-2)**, 35-61.

[15] **Freedland, K.E.** i **Carney, R.M.** (1992). «Data management and accountability in behavioral and biomedical research». *American Psychologist*, **47(5)**, 640-645.

[16] **Henning, R.A., Sauter, S.L., Salvendy, G.** i **Krieg, E.F.Jr.** (1989). «Microbreak length, performance, and stress in a data entry task». *Ergonomics*, **32**, 855-864.

[17] **Howes, A.** (1994). *A model of the adquisition of menu knowledge by examination.* En B. Adelson, S. Dumais y J. Olson (Eds.), *CHI'94*, pp. 445-451. Boston, MA: ACM.

[18] **Johnson, P.** (1992). *Human Computer Interaction: Psychology, Task Analysis and Software Engineering.* Londres: McGraw-Hill.

[19] **Karwowski, W., Eberts, R, Salvendy, G.** i **Noland, S.** (1994). «The effects of computer interface design on human postural dynamics: Special Issue: Festschrift in honour of Professor E. Negel Corlett». *Ergonomics*, **37(4)**, 703-724.

[20] **Lalomia, M.J.** i **Sidowski, J.B.** (1993). «Measurements of computer anxiety: A review». *International Journal of Human Computer Interaction*, **5(3)**, 239-266.

[21] **Moritz, T.E., Ellis, N.K., Villanueva, C.B., Steeger, J.E., Ludwig, S.T.** i **Deegan, N.I.** (1995). «Development of an interactive data base management system for capturing large volumes of data». *Medical Care*, **33** (10 Suppl.).

[22] **Pocius, K.E.** (1991). «Personality factors in human-computer interaction: A review of the literature». *Computers in Human Behavior*, **7(3)**, 103-135.

[23] **Ronel, R.K.; Varley, S.A.** i **Webb, C.F.** (1993). *Clinical data management.* Chichester: John Wiley & Sons.

[24] **Schneiderman, B.** (1992). *Designing the user interface: Strategies for effective human-computer interaction* (2ª ed.). Reading, MA: Addison-Wesley.

[25] **Wallace, M.D.** i **Anderson, T.J.** (1993). «Approaches to interface design». *Interacting with Computers*, **5(3)**, 259-278.

[26] **Wolf, C.G.** (1992). «A comparative study of gestual, keyboard, and mouse interfaces». *Behaviour and Information Technology*, **11(1)**, 13-23.

[27] **Wolf, R.M.** (1993). «Data quality and norms in international studies. Special Issue: Mandatory testing: Issues in policy-driven assessment». *Measurement and Evaluation in Counseling and Development*, **26 (1)**, 35-40.

ENGLISH SUMMARY

RANDOM VERIFICATION: A STRATEGY TO IMPROVE AND ASSESS THE QUALITY OF DATA ENTRY

J.M. DOMÈNECH MASSONS
J.M. LOSILLA VIDAL
M. POTELL VIDAL
Universitat Autònoma de Barcelona*

The paper addresses the problem of the reduction of errors produced during data entry which are not controllable through automatic protections. The habitual strategy in front of this problem is the «double entry» (DE) of data, which considerably increases the research cost. As an alternative to this strategy we propose a new procedure, implemented in DAT System, based on a process of «random verification» (RV) or a global data percentage. Besides of cost reduction, RV offers other advantages, as the fact of providing an error percentage estimation which supports the hypothesis that RV increases the effectiveness of data entry, without decreasing the efficiency, when is compared with situations in which there is no control about quality of entered data.

Keywords: Data quality, double entry, data entry, random verification, DAT System.

AMS Classification: 62D99

*Dept. de Psicobiologia i de Metodologia de las Ciències de la Salut, Universitat Autònoma de Barcelona. Aquesta investigació s'ha realitzat gràcies a l'ajut DGICYT No. PM95-126 del Ministeri d'Educació i Ciència.

-Received March 1997.
-Accepted October 1998.

INTRODUCTION

This work studies data entry applications and, concretely, the problematic of reduction of operators errors that are not automatically controllable, through a mechanism called «random verification» (RV). RV is one of a contribution set derived from DAT System (Domènech & Losilla, 1995), which are presented as a complement to the double data entry (DE) strategy.

Errors not automatically controllable are those units of information that fulfill previous rank, logical and missing value conditions, but that are not coincident with original value. This kind of errors could be limited by incorporating factors that affect on attentional, perceptional and motivational aspects of the operator managing data entry (Henning, Sauter, Salvendy and Krieg, 1989; Pocius, 1991; González, 1993; Lalomia and Sidowski, 1993).

GENERAL CHARACTERISTICS OF DAT SYSTEM

DAT System is an element of a wide project which has as objective the development of an efficient system for scientific data managing, allowing to control (from its capture to its preparation and exportation to the statistical analysis system) all the questions affecting data quality. Figure 1 locates, in data managing process, the 3 modules of DAT System (*EnDat/Lab, EnDat and ExperDat*).

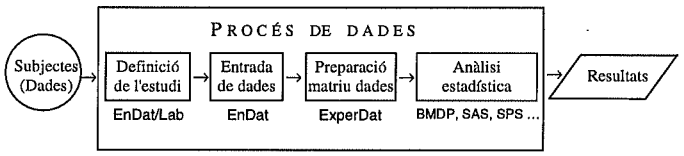


Figure 1

DAT System is also a suitable platform for research in Human Computer Interaction, specially regarding to: (*a*) detection of aspects that cause errors and/or decrease operators' work speed; (*b*) assessment of new alternatives in order to avoid errors and increase efficiency, as RV is, and (*c*)) development of tools indicating operators' kills, which could facilitate workers selection phase.

RANDOM VERIFICATION AS AN ALTERNATIVE OF COMPLEMENT TO DOUBLE ENTRY

The main disadvantage of DE is, without a doubt, its high economic and temporal cost. For this reason, in practice, only studies with huge economic resources (i.e. clinical trials) could apply double entry, being the greater of studies devoid of mechanisms that guarantee data quality.

In front of this situation, Domènech and Losilla (1995) propose a strategy of «random verification» (RV) of data, consisting on double entry of only a percentage of the total data, selected at random by computer system and requested with enough delay as to avoid memory effects, and on the reproduction of the operators' errors committed during first entry.

The main advantage of RV is that considerably decreases data entry cost, preserving a great part of positive characteristics attributed to double entry when is carried out in optimum conditions. That is to say, RV has the advantage of giving to the operator a feedback of errors produced and allows to obtain, in any time of data entry process, an estimation of data errors percentage and an index of operator' aptitude and efficacy.

This feedback of RV allow us to pose the first hypothesis of our study:

Hypothesis 1: *The feedback, inherent to RV situation, could positively contribute to increase the global efficacy of data entry process, that is to say, the number of errors committed will be less in data entry situations in which operator receives feedback.*

From a practical point of view and attending only to the economic cost of data entry, it is also interesting to compare operator' efficiency in situations in which they do not receive feedback about committed errors, with their efficiency in situations in which they do. But the comparison of efficiency levels involves the next hypothesis:

Hypothesis 2: *The time spent on the introduction of data questionnaires without errors will not be superior to the time spent on the introduction of data questionnaires with errors, independently of if the situations implicates or not feedback to the operator.*

Indeed, a difference in the opposed direction to the expressed in the hypothesis, that is to say, a greater time needed to introduce data without errors than data with errors, would implicate that increasing efficacy implies necessarily decreasing efficiency and, consequently, that comparing operators' efficiency in different data entry situations would not have any worth, but only comparing their efficacy levels. Contrarily, if data do not contradict hypothesis 2, we will contrast the next hypothesis about the efficiency:

Hypothesis 3: *The feedback associated to RV does not imply a decrease in efficiency, that is to say, the number of data per time unit introduced without error is equal or greater in situations in which operator receives feedback than in situations in which do not.*

METHOD

Subjects

The subjects are 45 voluntary students of first course of psychology at the Universitat Autònoma de Barcelona, with ages between 18 and 20 years-old and a 67% of woman, which have been divided in three groups of size 15.

Material

100 questionnaires have been hand filled with fictitious data and, to register them, a computerised form without protections has been created with DAT System.

Design

A random groups with pre-test registration design has been applied. The independent variable is the type of data entry situation assigned to each group: «random verification» (RV), «placebo» (PL) and «control» (CO).

RV condition is a verbal instruction informing that software will control the errors committed during data entry, and will oblige to repeat the introduction of some forms. The percentage of forms to verify is set at 15% and the delay in the double entry of these forms is established in two forms.

PL condition is a verbal instruction informing that software will control the number of committed errors. This condition is necessary to distinguish the specific effects of feedback that RV implies from the effect due to the knowledge that committed errors are being registered.

In Co conditions the subjects do not receive any instruction.

We have defined four dependent variables: *efficacy measurement* (proportion of data fields introduced without error), *efficiency measurement* (number of data records or questionnaires per hour of work introduced without error), *average time employed to introduce the questionnaires with error* (in seconds) *and average time employed to introduce the questionnaires without error* (in seconds). The time employed in the

second entry or revised questionnaires in RV has not been considered because this time does not exist in the other two conditions.

Procedure

Each group of 15 subjects has introduced to the computer the data from the whole set of questionnaires, in a structured session with three phases: training, baseline and experimental.

RESULTS

The results of the analysis show an efficacy increment because in RV group a 27.7% (IC95%: 20.7 to 24.7) more of data fields without error than in CO group have been introduced, and a 13.1% (IC95%: 11.1 to 15.0) more of data without error than in PL group.

The results also show that, although the average time of introduction of data with error in RV group has significantly increased in 0.5 seconds ($p = 0.02$), this increment has not realistic importance (IC95%: 0.1 to 0.9).

The results of the last analysis are congruent with the hypothesis of efficiency maintenance because the average time of introduction of data without error, although is slightly greater in RV group than in CO group, implies a difference neither of realistic importance ($d = 0.8$ sec.; IC95%: -0.5 to 2.1) nor statistically significant ($p = 0.22$). The same occurs in PL group ($d = 0.3$ sec.; IC95%: -1.0 to 1.6; $p = 0.64$).

DISCUSSION

The results show that RV increases the efficacy of data entry process, without decreasing its efficiency, when is compared with situations in which there is no control about data quality. Furthermore, the results obtained comparing RV and PL groups support the hypothesis about the specific effect of feedback inherent to RV situation as an explanatory mechanism of error reduction.

Although RV strategy, as happen to DE, has a higher time cost than simple entry, the contributions of this study suggest that this increment is not a sufficient motive to exclude this type of controls during data entry. In our opinion, it should never be analyzed data without having, at least, an estimation of the errors that they contain. For this reason, we consider the results about efficacy and efficiency in the application of controls through RV of great realistic interest.

QÜESTIIÓ, vol. 22, 3, p. 511-526, 1998

A COMPARATIVE STUDY OF MICROAGGREGATION METHODS*

J.M. MATEO-SANZ
J. DOMINGO-FERRER
Universitat Rovira i Virgili**

Microaggregation is a statistical disclosure control technique for microdata. Raw microdata (i. e. individual records) are grouped into small aggregates prior to publication. Each aggregate should contain at least k records to prevent disclosure of individual information. Fixed-size microaggregation consists of taking fixed-size microaggregates (size k). Data-oriented microaggregation (with variable group size) was introduced recently. Regardless of the group size, microaggregates on a multidimensional data set can be formed using univariate techniques on projected data or using multivariate techniques. This paper presents the first method for multivariate fixed-size microaggregation. In addition, a real data set is used to compare the information loss and output data quality of fixed-size vs. data-oriented, and univariate vs. multivariate microaggregation.

Keywords: Statistical disclosure control; Microaggregation; Data-oriented microaggregation; Microdata protection.

AMS Classification: 62H30, 92G30, 68PXX

* This work is partly supported by the Spanish CICYT under grant no. TIC98-0699-C02-02.
** Universitat Rovira i Virgili. Departament d'Enginyeria Informàtica. Autovia de Salou s/n, E-43006 Tarragona. Catalonia, e-mail:{jmateo,jdomingo}@etse.urv.es.
- Received July 1998.
- Accepted November 1998.

1. INTRODUCTION

A *microdata set* is a set of records containing data of individuals being studied, who can be persons, companies, etc. The individual records of a microdata set are stored in a *microdata file*. Each individual j is assigned a *data vector* V_j, also called data record or data set. A data vector is formed by several variables.

Microaggregation is a family of statistical disclosure control techniques for microdata which belong to the data modification category. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if the data vectors correspond to groups of k or more individuals, where no individual dominates (*i. e.* contributes too much to) the group and k is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation.

To obtain microaggregates in a microdata set with n data vectors, these are combined to form g groups of size at least k. For each variable, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting (modified) data vectors can be published.

The partition problem implicit in microaggregation differs from the classical clustering problem whose goal is to split a population into a fixed number of disjoint groups (Hartigan, 1975), regardless of the group size. Partitions resulting from microaggregation cannot consist of groups of size smaller than k; call such partitions k-partitions. To solve the k-partition problem, a measure of similarity between data vectors is needed. Each individual data vector can be viewed as a point and the whole microdata set as a set of multidimensional points. The dimension is the number of variables in data vectors. If data vectors are characterized as points, similarity between them can be measured using a distance.

To be more specific, consider a microdata set with p continuous variables and n data vectors (*i. e.* the result of observing p variables on n individuals). A particular data vector can be viewed as an instance of $X' = (X_1, \cdots, X_p)$ where the X_i are the variables. With these individuals g groups are formed with n_i individuals in the i-th group ($n_i \geq k$ and $n = \sum_{i=1}^{g} n_i$). Denote by x_{ij} the j-th data vector in the i-th group; denote by $\bar{x}_i$ the average data vector over the i-th group, and by $\bar{x}$ the average data vector over the whole set of n individuals.

The within-groups sum of squares SSE is defined as

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i)$$

The between-groups sum of squares SSA is

$$SSA = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})'(\bar{x}_i - \bar{x})$$

The total sum of squares is $SST = SSA + SSE$ or explicitly

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})'(x_{ij} - \bar{x})$$

The optimal k-partition is the one that minimizes SSE (or equivalently, maximizes SSA); sums of squares are usual to measure information loss (Gordon and Henderson, 1977). A measure L of information loss standardized between 0 and 1 can be obtained from

$$L = \frac{SSE}{SST} \tag{1}$$

The rest of this paper is organized as follows. In Section 2, previous work on univariate microaggregation is reviewed (both with fixed-size groups and variable-size groups). Section 3 deals with multivariate microaggregation techniques; a new concept of multivariate fixed-size microaggregation is presented and a method to implement it is specified; also recent work of these authors on multivariate data-oriented microaggregation is briefly recalled. In Section 4, the information loss and output data quality of microaggregation methods are compared based on a real data set; four families of methods are considered: univariate fixed-size, univariate data-oriented, multivariate fixed-size and multivariate data-oriented. Finally, Section 5 is a conclusion and a sketch of future research.

2. UNIVARIATE MICROAGGREGATION

Defays and Nanopoulos (1993) proposed a mathematical algorithm to find an optimal solution for the k-partition problem (minimizing the information loss L). The idea is to choose a suitable set of hyperplanes separating the n data vectors into a number of homogeneous groups. As pointed out by its authors, the proposed algorithm is pretty complicated and difficult to implement in practice. As an alternative, the same paper presents some of the practical alternatives described in Subsection 2.1.

2.1. Univariate fixed-size microaggregation

Practical heuristic microaggregation methods were proposed in Defays and Nanopoulos (1993), in Anwar (1993) and in Defays and Anwar (1995). The partition

mechanism is the same in all such methods: first, data vectors are sorted in ascending or descending order according to some criterion. Then groups of successive k vectors are combined. Inside each group, the effect for each variable is to replace the k values taken by the variable with their average. If the total number of data vectors n is not a multiple of k, the last group will contain more than k data vectors.

Instead of using a multidimensional distance to sort data vectors, all practical methods quoted above perform straightforward one-dimensional sorting (this fact explains why such methods are called univariate). Two main approaches exist: single-axis sorting and individual sorting.

Single-axis sorting methods are good if all variables are highly correlated. If a particular variable is used for sorting, this variable must reflect somehow the size of the data vector. Vectors are sorted in ascending or descending order by the sorting variable, and then groups of k successive vectors are formed. Inside each group and for each variable, values are replaced by the group average. A natural improvement is to sort data vectors by the first principal component of the microdata set rather than by a particular variable. Principal components are transformed variables such that the first principal component is highly correlated with most original variables. An alternative that, like principal components, also takes all variables into account is based on the sum of z-scores: all variables are standardized and, for each data vector, the standardized values of all variables are added. Vectors are subsequently sorted by their sum of z-scores.

If the individual sorting option is chosen, then each variable is considered independently. Data vectors are sorted by the first variable, then groups of k successive values of the first variable are formed and, inside each group, values are replaced by the group average. A similar procedure is repeated for the rest of variables. Individual sorting usually preserves more information than single-axis sorting, but has a higher disclosure risk. Indeed, with individual sorting any intruder knows that the real value of a variable in a data vector in the i-th group is between the average of the $i-1$-th group and the average of the $i+1$-th group; if these two averages are very close to each another, then a very narrow interval for the real value being searched has been determined. Individual sorting also has a conceptual drawback: it does not partition the n data vectors in the microdata set on a data vector basis; instead, microaggregation is done for each variable in turn so that a different partition is obtained for each variable in the microdata set.

2.2. Data-oriented univariate microaggregation

In Domingo and Mateo (1998), microaggregation using variable-sized groups depending on data (data-oriented microaggregation) is introduced in the univariate case.

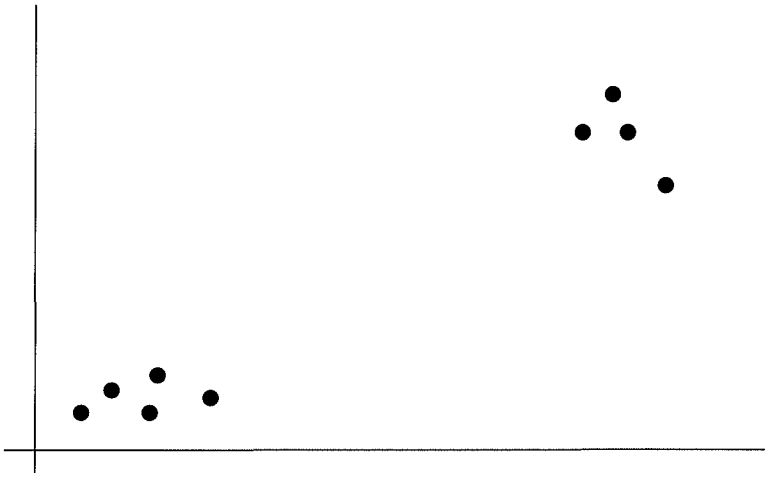


Figure 1. Variable-sized groups versus fixed-sized groups.

The idea is that groups need not consist of exactly k data vectors, but of *at least* k data vectors. Methods yielding variable-sized groups are a bit more complex than fixed-size microaggregation (Defays and Nanopoulos, 1993) but they may take advantage from the distribution of original data to obtain a smaller information loss in comparison with fixed-size microaggregation. Figure 1 is a simple graphical example that illustrates the advantages of variable-sized groups. The figure shows two variables and nine data. If fixed-size microaggregation with $k = 3$ is used, we obtain a partition of the data into three groups, which looks rather unnatural for the data distribution given. On the other hand, if variable-sized groups are allowed then the five data on the left can be kept in a single group and the four data on the right in another group; such a variable-size grouping achieves a smaller information loss.

As discussed above, deterministically finding an optimal solution to the k-partition problem is very difficult. Heuristic methods are the only practical alternative and they should attempt to minimize the information loss L specified by expression (1). Since SST is fixed for a given data set, one should attempt to find a grouping that minimizes SSE.

In Domingo and Mateo (1998) two alternative heuristic approaches to variable-size univariate microaggregation are presented:

- Microaggregation based on genetic algorithms (GA).
- Modified Ward's Algorithm (k-Ward).

Being univariate, both approaches above must be combined with single-axis or individual sorting (described in Section 1) to deal with a multivariate microdata set.

Genetic microaggregation represents k-partitions as binary strings (also called chromosomes) and combines directed and random search to locate global optima.

Hierarchical classification methods can also be used as building blocks for heuristic microaggregation methods yielding variable-sized groups. Ward's method (Ward, 1963) is attractive because it is stepwise optimal: the two groups or data elements joined at each step are chosen so that the increase in the within-groups sum of squares *SSE* caused by their union is minimal. However, Ward's method must be adapted to be useful for microaggregation (k-Ward algorithm). The standard method just builds up a grouping hierarchy, whereas a k-partition of the initial data set is desired. As will be shown in this paper, k-Ward can be naturally turned into a truly multivariate microaggregation method.

k-Ward is a microaggregation method for quantitative data or for qualitative data where a distance has been defined. In what follows, k-Ward will be briefly recalled. The following definitions and results are needed:

Definition 1. *For a given data set, a k-partition P is any partition of the data set such that each group in P consists of at least k elements.*

Definition 2. *For a given data set, k-partition P is said to be finer than k-partition P' if every group in P is contained by a group in P'.*

It is straightforward to check that «finer than» is a partial order relationship on the set of k-partitions of a given data set.

Definition 3. *For a given data set, a k-partition P is said to be minimal with respect to the relationship «finer than» if there is no k-partition $P' \neq P$ such that P' is finer than P.*

Proposition 1. *For a given data set, k-partition P is minimal with respect to the relationship «finer than» if and only if it consists of groups with sizes $\geq k$ and $< 2k$.*

Corollary 1. *An optimal solution to the k-partition problem of a set of data exists that is minimal with respect to the relationship «finer than».*

The proofs of the above results can be found in Domingo and Mateo (1998). Now, Ward's hierarchical classification method can be modified to provide a solution that belongs to the set of candidate optimal solutions characterized by Proposition 1 and Corollary 1. Modified Ward's algorithm (k-Ward) is as follows:

Algorithm 1 ($k-$Ward)

1. *Form a group with the first (smallest) k elements of the data set and another group with the last (largest) k elements of the data set.*
2. *Use Ward's method until all elements in the data set belong to a group containing k or more data elements; in the process of forming groups by Ward's method, never join two groups which have both a size greater than or equal to k.*
3. *For each group in the final partition that contains $2k$ or more data elements, apply this algorithm recursively (the data set to be considered is now restricted to the particular group containing $2k$ or more elements).*

The following property of the above algorithm is proven in Domingo and Mateo (1998); its proof is recalled here because it helps understanding the design of the algorithm.

Property 1 (Convergence). *Algorithm 1 ends after a finite number of recursion steps.*

Demostració. By Step 1 of the above algorithm each new recursion step starts splitting the initial data group into at least two groups; the rule in Step 2 ensures that the group formed by the smallest elements and the group formed by the largest elements are never joined thereafter (because of their sizes). In this way, at the end of a recursion step, the final k-partition consists of at least two groups and is therefore finer than the initial k-partition (consisting of a single group). If there is a group of size $\geq 2k$, then the algorithm is recursively applied to it and strictly smaller groups will be obtained (according to the previous argument). Thus after a finite number of recursion steps a k-partition of the initial data set will be obtained such that the maximal group size is less than $2k$.

■

As explained above, Ward's algorithm is stepwise optimal in what regards information loss. Stepwise optimality does no longer hold for k-Ward, but a good behaviour is expected given that k-Ward is built on top of Ward's method. See Section 4 for computational results.

3. MULTIVARIATE MICROAGGREGATION

We present in this section the multivariate counterparts of the methods described above. We define multivariate microaggregation to be microaggregation on multivariate *unprojected* data. Of course, if multivariate data are one-dimensionally projected using single-axis or individual sorting (see Section 1), the resulting projected data can be microaggregated with the univariate methods described in Section 2. However, single-axis sorting is a rather coarse technique and individual sorting has a higher disclosure risk and does not really perform microaggregation on a data vector basis.

Work presented in Subsection 3.1 on multivariate fixed-size microaggregation is new. In Section 3.2, previous work of these authors on data-oriented multivariate microaggregation is recalled (Mateo and Domingo, 1998; Mateo, 1998).

3.1. Multivariate fixed-size microaggregation: a new proposal

We introduce in this subsection a new family of microaggregation methods. The idea is to form groups of size k *without* projecting multivariate data in one-dimension. Instead, a multivariate distance is used. The basic algorithm can be described as follows:

Algorithm 2 (Multivariate fixed-size microaggregation)

1. *Form one group with the «first» k data vectors and another group with the «last» k data vectors.*
2. *If there are at least $2k$ data vectors which do not belong to the two groups formed in Step 1, go to Step 1 taking as new data set the previous data set minus the groups formed in the previous instance of Step 1.*
3. *If there are between k and $2k-1$ data vectors which do not belong to the two groups formed in Step 1, form a new group with those elements and exit the Algorithm.*
4. *If there are less than k data vectors which do not belong to the groups formed in Step 1, add them to the closest group formed in Step 1.*

The problem remains of how to decide which are the «first» and «last» data vectors in Step 1 of the above algorithm. If single-axis sorting is used to make that decision, Algorithm 2 is equivalent to performing univariate fixed-size microaggregation on projected multivariate data. A truly multivariate criterion is as follows. Define as extreme data vectors the two vectors in the data set which are most distant according to the distance matrix; then, for each of the extreme data vectors,

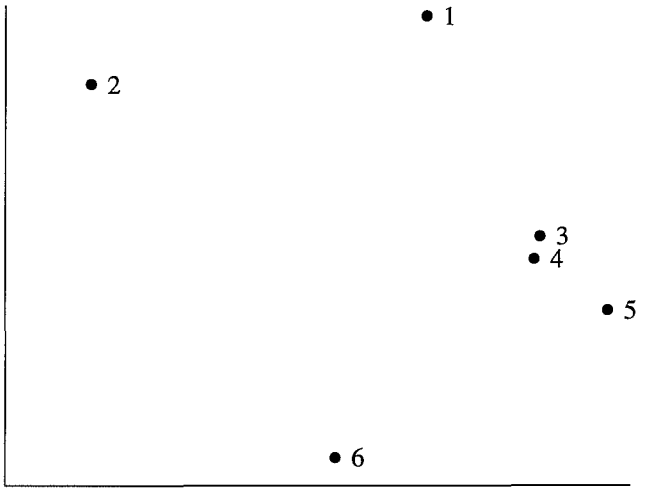


Figure 2. Grouping with the maximum-distance criterion.

take the $k-1$ data vectors closest to it following the distance matrix; in this way, a group with the «first» k data vectors and another group with the «last» data vectors are obtained. This criterion for choosing the «first» and the «last» data vectors will be called maximum-distance (MD) criterion. The grouping resulting from MD may depend on which extreme data vector one starts with, *i. e.* which extreme data vector is taken as the «first» data vector. For example, consider the six two-dimensional data vectors in Figure 2 and take $k = 3$. The two most distant vectors are labeled 2 and 5. Starting from vector 2, the closest vector is vector 1; now the vector closest to the group (1,2) is vector 3. So starting from vector 2, we get the groups (1,2,3) and (4,5,6). But if we choose to start from the other extreme point (vector 5), the closest vector is vector 4; now the vector closest to the group (4,5) is vector 3. So starting from vector 5, we get the groups (3,4,5) and (1,2,6). Anyway, the differences in the information loss resulting from choosing either extreme vector as «first» or «last» are small (see results in Subsection 4.2). Furthermore, the larger the data set, the less likely is the kind of pathological situation depicted in Figure 2.

3.2. Multivariate data-oriented microaggregation

A natural way to obtain multivariate data-oriented methods is to generalize some of the univariate data-oriented methods quoted in Subsection 2.2. Genetic microaggregation methods are not so easy to adapt for dealing with unprojected multivariate data: the main problem comes from the fact that a multidimensional space is only partially

ordered, which makes properly representing multivariate k-partitions as binary strings far from obvious.

Luckily enough, the strong point of k-Ward is that it can be easily adapted into a multivariate k-Ward algorithm to directly work with multidimensional (unprojected) data vectors. The reason is that the underlying Ward's method was actually designed as a multivariate clustering algorithm. Thus to obtain a multivariate version of k-Ward, only Step 1 of Algorithm 1 needs to be adapted. Basically, what is needed is a multivariate sorting criterion specifying what is meant by the «first» k data vectors and the «last» k data vectors.

Unlike for multivariate fixed-size microaggregation, it makes sense to use single-axis sorting to determine which are the first and last k vectors in Step 1 of Algorithm 1. The resulting algorithm is *not* equivalent to data-oriented univariate microaggregation on projected data. The multivariate data vectors can be ranked according to their first principal component, their sum of z-scores or a particular variable. The maximum-distance criterion may also be used as an alternative to avoid one-dimensional projection, and then the grouping depends on which extreme vector is taken as first.

4. COMPUTATIONAL RESULTS

The performance of the multi-dimensional microaggregation methods discussed in this work has been compared using a data set of 834 companies in the Tarragona area for which 13 variables have been collected: fixed assets (V1), current assets (V2), treasury (V3), uncommitted funds (V4), paid-up capital (V5), short-term debt (V6), sales (V7), labour costs (V8), depreciation (V9), operating profit (V10), financial outcome (V11), gross profit (V12) and net profit (V13). The data set corresponds to year 1995.

The methods considered in the comparison include univariate fixed-size microaggregation (denoted by UFS), univariate data-oriented microaggregation using k-Ward (denoted by UDO), multivariate fixed-size microaggregation (MFS), and multivariate data-oriented microaggregation using k-Ward (denoted by MDO). For UFS and UDO (step 1 of k-Ward) several sorting criteria have been considered: a particular variable (PV), sum of z-scores (SZ) and first principal component (FPC). For MFS, only truly multivariate sorting criteria such as maximum-distance (MD) make sense (see Subsection 3.1). For MDO (step 1 of multivariate k-Ward), the PV, SZ, FPC and MD criteria have been considered. In the rest of this section, the sorting criterion appears as a subscript of the microaggregation method. When using the PV criterion, a range of results is obtained depending on which particular variable is used; if a single result is reported in what follows, it corresponds to the *best variable*.

4.1. Computing time

The whole data set of 834 data vectors was microaggregated on a Pentium MMX at 166MHz running under the Linux operating system. The following results for the microaggregation time (excluding sorting time) were obtained:

- With UFS, the computing time is negligible for any $3 \leq k \leq 5$ regardless of the sorting criterion used.
- With UDO, the computing time is about 25 seconds, regardless of the group size and the sorting criterion used.
- With MFS, the computing times is also about 25 seconds.
- With MDO, the computing time is between 35 and 39 seconds.

The above results exhibit no significant differences. It should be noted here microaggregation is usually performed off-line and even a few hundred seconds of computing time would be acceptable. Thus, it can be concluded that the computing time is not an issue for either method, even if MDO turns out to be somewhat slower than the rest of methods. The really interesting comparison is in terms of information loss and data quality.

4.2. Information loss and data quality

To compare the information loss caused by multivariate fixed-size microaggregation and k-Ward, we have considered the loss L (see expression (1)). It must be pointed out here that the value of L depends on the units used for the variables in the microdata set. Such an undesirable property can be neutralized if all variables are standardized prior to microaggregation: if variable V_i takes a value x, then x is replaced by $(x-\bar{v}_i)/s_{v_i}$, where $\bar{v}_i$ and s_{v_i} are, respectively, the average and the standard deviation of the values taken by V_i. Results presented throughout this section have been obtained on standardized variables. Table 1 shows the percentage values of L obtained for UFS, UDO, MFS and MDO.

For the FPC and SZ sorting criteria, two losses are given (unless both are very similar). The first one is obtained when data are sorted in ascending order following the criterion; the second loss is obtained when data are sorted in descending order.

A range of losses is given for the PV criterion. With this criterion, the information loss depends on the particular variable chosen for sorting. Thus the lower limit of the range corresponds to the best variable (leading to a minimal loss) and the upper limit to the worst variable. It can be seen that the range for MDO is narrower than

for UDO. In this sense, multivariate data-oriented methods are more *robust* than their univariate counterparts.

Table 1. *Comparison of percentage information loss*

Method	$100L$ ($k=3$)	$100L$ ($k=4$)	$100L$ ($k=5$)
UFS_{PV}	30.11 - 48.48	34.14 - 57.0	37.59 - 60.83
UFS_{SZ}	28.92	32.15 or 32.08	35.20 or 32.56
UFS_{FPC}	23.87 or 23.89	30.62 or 25.99	33.29 or 30.74
UDO_{PV}	31.17 - 53.93	35.28 - 58.23	39.06 - 61.18
UDO_{SZ}	30.13	32.56	34.65
UDO_{FPC}	25.35	31.71	32.08
MFS_{MD}	15.60	19.27	22.67
MDO_{PV}	16.23 - 21.61	20.46 - 29.45	22.38 - 31.77
MDO_{SZ}	19.16	24.31	27.6
MDO_{FPC}	15.87	21.58	23.69
MDO_{MD}	16.01 - 16.75	21.13 - 21.24	21.83 - 22.77

For the MDO method with MD criterion, a range is also given. The reason is that each time Step 1 of the multivariate k-Ward method is run, one must decide which of two extreme data vectors is the «first» one and which is the «last» one. Thus, if Step 1 is run m times, one could in principle obtain as many as 2^m different losses. However, it can be seen that the MD criterion is pretty robust in that the difference between the worst and best losses is very small.

Table 2 compares average standard deviations of all variables under each method and for three group sizes. Original data have been standardized, so they have standard deviation 1; microaggregated data cannot contain more information, *i. e.* more variability, so standard deviations for variables are less than or equal to 1. The closer the average standard deviation to 1, the better is a method. It can be seen that multivariate methods perform better than univariate methods; a closer look reveals that MFS is slightly better than MDO.

For each method and for each group size k, Table 3 reflects the impact of the various microaggregation methods on the first principal component. The table gives:

ΔR: Percentage change in the average correlation of variables with the first principal component. Using the original data set, the percentage change is 0; therefore, the smaller this figure, the better is a method.

ΔW: Percentage change in the average weight of variables on the first principal component. The smaller this figure, the better is a method.

%FPC: Percentage proportion of variability of the whole microaggregated data set explained by the first principal component. In the original data set, the first component explains 63.4% of the total variability. The more similar the percentage explained to 63.4, the better is a method.

Table 2. *Comparison of average standard deviations*

Method	$k=3$	$k=4$	$k=5$
UFS_{PV}	.83	.80	.78
UFS_{SZ}	.84	.82	.81
UFS_{FPC}	.87	.86	.83
UDO_{PV}	.83	.80	.77
UDO_{SZ}	.83	.81	.80
UDO_{FPC}	.86	.82	.82
MFS_{MD}	.92	.90	.88
MDO_{PV}	.91	.89	.88
MDO_{SZ}	.90	.87	.85
MDO_{FPC}	.92	.88	.87
MDO_{MD}	.92	.89	.88

Table 3. *Impact of microaggregation on the FPC*

Method	$k=3$			$k=4$			$k=5$		
	ΔR	ΔW	%FPC	ΔR	ΔW	%FPC	ΔR	ΔW	%FPC
UFS_{PV}	15.8	11.7	81.4	20.2	15.7	85.8	22.0	17.0	87.8
UFS_{SZ}	18.8	14.5	84.3	21.3	15.0	88.0	22.6	16.3	89.4
UFS_{FPC}	16.3	13.1	81.3	17.1	13.4	82.9	21.7	16.1	88.3
UDO_{PV}	16.6	12.3	82.4	21.3	16.6	86.9	23.7	18.4	89.8
UDO_{SZ}	19.8	14.1	86.0	22.1	15.3	88.9	24.4	17.5	91.3
UDO_{FPC}	17.7	14.1	83.0	23.2	16.6	90.0	23.6	17.1	90.4
MFS_{MD}	7.7	7.1	71.9	9.1	8.1	73.7	9.6	8.0	74.3
MDO_{PV}	8.2	8.0	72.6	10.4	9.5	75.1	10.6	9.0	75.5
MDO_{SZ}	10.1	9.1	74.7	13.7	10.6	79.2	14.7	11.5	80.1
MDO_{FPC}	8.2	8.0	72.6	11.4	9.9	76.7	13.2	10.9	78.6
MDO_{MD}	7.8	7.9	72.1	10.7	9.5	75.8	10.5	9.0	75.3

Table 4. *Impact of microaggregation methods on correlations*

Method	$k=3$		$k=4$		$k=5$	
	$\overline{\Delta r}$	$s_{\Delta r}$	$\overline{\Delta r}$	$s_{\Delta r}$	$\overline{\Delta r}$	$s_{\Delta r}$
UFS_{PV}	.20	.08	.26	.12	.28	.13
UFS_{SZ}	.24	.10	.28	.11	.29	.12
UFS_{FPC}	.20	.09	.22	.10	.28	.11
UDO_{PV}	.21	.09	.27	.13	.30	.15
UDO_{SZ}	.26	.10	.19	.12	.32	.13
UDO_{FPC}	.22	.10	.30	.12	.31	.13
MFS_{MD}	.10	.05	.12	.06	.12	.06
MDO_{PV}	.10	.06	.13	.07	.14	.07
MDO_{SZ}	.13	.06	.18	.07	.19	.08
MDO_{FPC}	.10	.06	.15	.07	.17	.07
MDO_{MD}	.10	.05	.14	.06	.13	.06

Table 4 summarizes the impact of the various microaggregation methods on the correlations between variables. With 13 variables, the number of unordered variable pairs is $13!/(11!2!) = 78$. Let r_{ij} be the linear correlation coefficient between variables i and j for original data; let r_{ij}^m be the correlation coefficient between the same variables once data have been microaggregated using method m. For each method m, Table 4 gives the average $\overline{\Delta r}$ and the standard deviation $s_{\Delta r}$ of the 78 discrepancies $|r_{ij}^m - r_{ij}|$. The smaller the average discrepancy and the smaller the discrepancy variability, the better is a method.

5. CONCLUSION AND FUTURE RESEARCH

The results shown in Tables 1 through 4 for the microdata set tested can be summarized as follows:

- Multivariate methods behave significantly better than univariate methods. The reason is that in univariate microaggregation there are two sources of information loss: one-dimensional data projection and microaggregation. In multivariate microaggregation, the only source of information loss is microaggregation itself.
- Among univariate methods, UFS_{FPC} is slightly better than the rest for this data set.
- Among multivariate methods, the new method MFS_{MD} is better than the rest for this data set.

- The MD sorting criterion seems to be the best one for multivariate microaggregation. It is very robust, gives best results and requires little computation.

An interesting line of future research would be to repeat the comparative study performed in this paper for a very large data set. Some changes in the implementation of MFS and MDO may be necessary to deal with very large multivariate data sets: the distance matrix between data vectors cannot be prestored and thus distances between data vectors must be computed when they are needed. This introduces a considerable overhead and may be a computing time penalty for MFS. The reason is that MFS normally tends to create more groups than MDO, and thus requires more distance computations to complete the microaggregation process.

Other research topics include the development of alternative heuristics for multivariate microaggregation. For example, one could think of adapting genetic algorithms to deal with unprojected multivariate data.

ACKNOWLEDGMENTS

We thank the Registre Mercantil de Tarragona for providing the data set used to carry out the tests described in Section 4.

REFERENCES

[1] **Anwar, N.** (1993). *Micro-Aggregation - The Small Aggregates Method,* internal report, Luxembourg: Eurostat.

[2] **Defays, D.** and **Nanopoulos, P.** (1993). «Panels of enterprises and confidentiality: the small aggregates method», in *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys,* Ottawa: Statistics Canada, 195-204.

[3] **Defays, D.** and **Anwar, N.** (1995). «Micro-aggregation: a generic method», in *Proceedings of the 2nd International Symposium on Statistical Confidentiality,* Luxembourg: Eurostat, 69-78.

[4] **Domingo-Ferrer, J.** and **Mateo-Sanz, J.M.** (1998). *Practical Data-Oriented Microaggregation for Statistical Disclosure Control,* Research Report DEI-RR-98-005, Department of Computer Science, Tarragona: Universitat Rovira i Virgili.

[5] **Gordon, A.D.** and **Henderson, J. T.** (1977). «An algorithm for Euclidean sum of squares classification», *Biometrics,* **33**, 355-362.

[6] **Hartigan, J.A.** (1975). *Clustering Algorithms,* New York: Wiley.

[7] **Mateo-Sanz, J.M.** (1998). *Contributions to Statistical Disclosure Control*, Faculty of Mathematics, Universitat de Barcelona (in Catalan).

[8] **Mateo-Sanz, J.M.** and **Domingo-Ferrer, J.** (1998). «A method for data-oriented multivariate microaggregation», in *Proceedings of Statistical Data Protection'98*, Amsterdam: IOS Press (to appear).

[9] **Ward, J.H.** (1963). «Hierarchical grouping to optimize an objective function», *Journal of the American Statistical Association*, **58**, 236-244.

Biometria

QÜESTIIÓ, vol. 22, 3, p. 529-538, 1998

THE USEFULNESS OF DISCRIMINATION BASED ON DISTANCES ON HUMAN EVOLUTION

C. ARENAS*
D. TURBÓN**
Universitat de Barcelona

The reconstruction of human history from the fossil record often runs up against incomplete or differential preservation of specimens. In anthropological studies a large number of variables are usually taken and missing values can be a problem. Here we analyze three population samples of extinct aborigines from Tierra del Fuego. The first sample, with sex and ethnic group known, is used to compare the step-wise discriminant analysis and the discriminant analysis based on distances. With the second sample a first approach to the assignation of poorly documented specimens in relation to sex or ethnic group is presented here by comparing the results from the two discriminant methods. A third sample of skulls with ethnic group and sex unknown is used to illustrate the advantages of distance-based discriminant analysis to solve the problem of allocating individuals when some values are missing.

Keywords: Distance discrimination; step-wise discrimination; missing values; Tierra del Fuego aborigines.

* Facultat de Biologia. Departament d'Estadística. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona.
** Departament de Biologia Animal (Antropologia). Facultat de Biologia. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona.

- Received February 1997.
- Accepted April 1998.

1. INTRODUCTION

One important aim in anthropology is to reconstruct extinct human groups, and thus to find relationships between them, to analyse differences and similarities with other extinct groups or with modern groups and to establish a possible common origin. For these reasons it is essential that the material studied should be accurately identified. Moreover accurate identification is necessary in order to allocate problematic individuals. However, the anthropological reconstruction of extinct human groups from archaeological sites is usually conditioned by the state of preservation of the remains, usually skull and long bones, and the statistical treatment often has to deal with a variable set of missing values. Another source of difficulty comes from the evolutionary context. For instance, the size of the bones could be problematic when determining sex of remains belonging to neighbouring ethnic groups, as some females from an hypothetically robust group could be erroneously classified as males from another group. This is the case when dealing with skulls of aborigines from Tierra del Fuego (particularly the Ona, pedestrian hunters-gatherers in Isla Grande) which show great osteological robusticity, all of them probably corresponding to a Paleoindian stock (Lahr 1995). This great robusticity prevents the distinction between the Ona female skulls and male skulls of the sea-canoe aborigines Yamana and Alakaluf. All these aborigines were decimated upon contact with Europeans, leading to their virtual extinction between the turn of the nineteenth century and the early twentieth century. In particular, the Ona were moved away from their original land, where they mixed with other ethnic groups.

This study is a tentative classification of a number of Fueguian skulls from different European and American museums and collections (Turbón 1995). Some cases are of uncertain attribution because of their physical displacement, complicated by the difficulty of discriminating the robust Ona females from the sea-canoe males. Sometimes more than one possible identification is given or contemporany anthropologists contradict former identifications.

Our skulls were further classified in three groups depending on previous identification. One with completely identified skulls (ethnic group and sex); another with no sure sex identification, and a third group with poor ethnic and sex identification. The aim of this study is to clarify the identification of these last two groups. A discriminant analysis is proposed, but as usual in the analysis of measurements of human skulls the following difficulties arise. When a broken skull was found, only some measurements could be taken. Thus, the data were incomplete and then several choices are available to compute the missing values. In what follows, some of the more usual solutions found in the literature are commented. One choice is to remove all cases for which the data are incomplete, which often reduces the number of samples dramatically and could exclude particular cases of critical importance for the analysis. A second choice is the replacement of missing values, either by the group mean or by values

obtained through multiple regression, which is not satisfactory in our case since some subgroups contain only a few specimens. Another possibility is the suppression of the variables for which a large number of values are missing, which would involve a significant reduction of information. Furthermore, the longer skulls could be prone to damage or poor preservation (Rao 1989) and a distinction must be made between measurements taken on well preserved skulls and those from damaged skulls. In this study we work under the hypothesis that the distribution of missing values is random. Finally, other difficulties can arise when classical discriminant analysis is applied. For example, if there many variables classical step-wise discriminant analysis is appropriate. However, the step-wise forward method with the F criterion for including a variable requires the data to be normally distributed. Furthermore the variables selected by this method are not always optimal (see Mc Cabe 1975). Moreover it is possible that the new individuals to be classified present missing values in the variables selected (sometimes in all of them), so correct allocation is not possible. In order to avoid some of these problems, this study uses the discriminant analysis based on distances introduced by Cuadras (1989). First a brief description of the method is presented and some interesting properties are discussed. After describing our data, a discriminant analysis and the assignation of some skulls with problems in the sex or ethnic group identification are performed first using the step-wise discriminant analysis, and then using the distance-based method. The results given by these methods are compared and we discuss some of their advantages and disadvantages.

2. MATERIAL AND METHOD

The material studied consists of 162 Fueguian skulls belonging to three ethnic groups: Yamana, Alakaluf and Ona (Table 1). A total of 65 biometrical traits were measured following W.W. Howells' technique (Howells 1973, 1989). These measurements (see Howells 1973 for details) are useful in the identification of the sex and ethnic group. This material is classified in three groups. Sample S_1 contains skulls with sure sex and ethnic group identification; sample S_2 contains skulls with sure ethnic group identification but doubtful sex identification; sample S_3 contains skulls with doubtful sex and ethnic group identification.

The distance-based (DB) discriminant analysis was introduced by Cuadras (1989) and it has recently been explained in detail (Cuadras 1989,1991,1992; Arenas *et al.* 1994). Its goal and rule of classification may be briefly summarised as follows.

Given some groups $\prod_i$ $(i = 1,\ldots,k)$ and a selected distance function $\delta(\cdot,\cdot)$ between individuals, then the rule of classification for a new individual x is,:

«allocate x to $(i = 1,\ldots,k)$ if and only if $f_i(x) = \min\{f_1(x),\ldots,f_k(x)\}$»,

where

$$f_i(x) = \frac{1}{n_i} \sum_{l=1}^{n_i} \delta_{li}^2 - \frac{1}{2n_i^2} \sum_{l,j=1}^{n_i} \delta_{lji}^2$$

and n_i, the sample size of group $\prod_i$; δ_{lji}^2 the square distance between objects l and j of group $\prod_i$ and δ_{li}^2 the square distance between object l of group $\prod_i$ and the new object x.

Cuadras *et al.* (1997a) proved that each $f_i(x)$ can be interpreted as the proximity of x to $\prod_i$. Thus the DB rule assigns an individual to the nearest group (see also Cuadras *et al.* 1997b). It can also be showed that it is equivalent to the linear discriminant rule, the quadratic discriminant rule or the euclidean discriminant rule, if an appropriate distance function is taken. Furthermore, as it is based on a distance, it can be applied to binary, qualitative or mixed variables by using a suitable distance.

It is clear that the results of the distance-based discriminant analysis depend on the distance selected. In this study Gower's distance (Gower, 1971) was chosen. This distance is obtained by assigning a score $0 \leqslant s_{ijk} \leqslant 1$ and a weight w_{ijk} for variable k.

The expression of this distance is given by $d_{ij} = 1 - \dfrac{\sum_k s_{ijk} w_{ijk}}{\sum_k w_{ijk}}$ where for continuous variables $s_{ijk} = \Sigma\left(1 - \left|x_{ik} - x_{jk}\right| / G_k\right)$, G_k is the range of the kth continuous variable.

For qualitative or binary variables s_{ijk} is 1 for matches between states and 0 for mismatches. The weight w_{ijk} is set to 1 when a comparison is considered valid for variable k and to 0 when the value of variable k is unknown for one or both observational units. As proved in Montanari (1994) it is a suitable distance for the treatment of data with missing values because it seems to be the least biased and reproduces the original cluster structure.

Summarising, distance-based discrimination has the following advantages:

- It works with mixed variables.
- It allows to work with a large number of variables.
- It can deal with missing values.
- It allocates new individuals with missing values.
- It does not need calculation of any inverse-matrix, so it is robust to the problem of ill-conditioned covariance matrices.
- It does not need any hypothesis about the distribution (normality) of data.

For all these reasons, a distance-based discriminant analysis might be preferable to classical linear discrimination or step-wise discriminant analysis in some cases. In this study, a comparison between the DB-method and the step-wise forward method using the F criterion is carried out. For calculating the probability of miss-classification the leave-one-out method is used. The analysis with the step-wise method is performed using the BMDP package. The DB method is implemented in the package of multivariate analysis Multicua (Arenas *et al.* 1991, 1993, 1998). A version for a large number of data was written by F. Oliva in SAS/IML.

3. RESULTS

In our data (Table 1), 75% of the skulls measured had a variable number of missing values, affecting 68% of the 65 variables observed. As mentioned above, the skulls were classified in three different subsamples, S_1 = completely known; S_2 = no sure sex identification and S_3 = ethnic group and sex unknown.

Table 1. *Description of the data: number of skulls for males (M) and females (F)*

	S_1		S_2	S_3
	M	F		
Yamana	31	22	11	
Alakaluf	11	10	2	
Ona	19	6	32	
Total	99		45	18

First we consider the data of sample S_1. The results of three DB discriminant analysis on the three groups of S_1 compared with those derived from a classical step-wise discriminant analysis are shown in Table 2. Table 3 shows the sex assignation given by the DB model and the step-wise model for the skulls of group S_2. The results of a new discriminant analysis when both samples S_1 and S_2 (with the final assignation) are put together are presented in Table 4. Finally using all the skulls of group S_1 an analysis is made in order to assign individuals of the S_3 group. The results of the discriminant analysis and the assignations are given in Tables 5 and 6 respectively.

In the first analysis (Table 2) a higher percentage of correct classification is obtained by the step-wise method, confirming the well known efficiency of this procedure. However, when we try to assign individuals of S_2, the advantages of the DB-discriminant analysis are clear (Table 3). With the step-wise discriminant analysis, some skulls cannot be allocated because they have missing values in the variables selected. From these results, it is clear that although the classical step-wise discriminant analysis initially gives better results, the DB-discriminant analysis is more useful for new

assignations. Furthermore, if the variables selected by the step-wise discriminant analysis are removed in order to work only with variables for which the elements of S_2 have no missing values, the assignation for all skulls is then possible although the probability of misclassification becomes greater (0.06 for Yamana; 0.2 for Ona). A new discriminant analysis is performed when samples S_1 and S_2 are put together. As Table 4 shows, with the step-wise method some skulls of known sex and ethnic group (from S_1) are now incorrectly classified. Finally when a discriminant analysis is performed with the skulls of group S_1 (see Table 5) the step-wise method , as before, gives a better classification than the DB rule, but again problems arise when skulls of S_3 are assigned (see Table 6). In this case using the second assignation, it is impossible to assign the skulls of S_3. These skulls present missing values in the variables selected by the step-wise discriminant analysis. If the variables selected by the step-wise are removed then the probability of bad classification (0.42) by the step-wise discriminant analysis becomes greater than the probability of bad classification using the DB-discriminant analysis (0.232).

Table 2. *Results of a DB-discriminant analysis and classical step-wise analysis on S_1*

DB-discriminant analysis						Step-wise discriminant analysis					
Matrix of misclassification						Matrix of misclassification					
		M	F	Prob. misclassif.	number variables			M	F	Prob. misclassif.	variables selected
Yamana	M	27	4	0.132	65	Yamana	M	31	0	0	8
	F	3	19				F	0	22		
Alakaluf	M	10	1	0.095	65	Alakaluf	M	11	0	0	10
	F	1	9				F	0	10		
Ona	M	17	2	0.28	65	Ona	M	18	1	0.04	2
	F	5	1				F	0	6		

Table 3. *Results of the assignation of skulls from group S_2*

Assignation according to the Museum (initial assignation) and biometrical assignation (DB and step-wise). ?= have missing values in the selected variables and no assignation is possible.

	Initial assignation	DB assignation	Step-wise assignation
Yamana	4M	3M 1F	0M 3F 1?
	7F	7F	1M 6F
Alakaluf	1M	1M	1F
	1F	1F	1M
Ona	23M	18M 5F	4M 17F 2?
	9F	3M 6F	5M 4F

Table 4. *Results of a DB-discriminant analysis and classical step-wise analysis on S_1 and S_2*

DB-discriminant analysis						Step-wise discriminant analysis					
Matrix of misclassification						Matrix of misclassification					
		M	F	Prob. misclassif.	number variables			M	F	Prob. misclassif.	variables selected
Yamana	M	30	4[1]	0.125	65	Yamana	M	30	2[4]	0.06	4
	F	4[2]	26				F	2[4]	29		
Alakaluf	M	11	1[1]	0.087	65	Alakaluf	M	10	2[4]	0.174	22
	F	1[1]	10				F	2[4]	9		
Ona	M	33	7[3]	0.210	65	Ona	M	25	3[4]	0.091	6
	F	5[3]	12				F	2[4]	25		

1 the same incorrectly assigned skulls as in the first analysis (Table 2)
2 three of the incorrectly assigned skulls in the first analysis (Table 2) and the skull initially assigned as M and finally assigned as F.
3 skulls of sample S_2.
4 skulls with ethnic group and sex known (from S_1) that now are incorrectly classified.

Table 5. *Results of a DB-discriminant analysis and classical step-wise analysis on S_1 group.*

DB-discriminant analysis Matrix of misclassification							Step-wise discriminant analysis Matrix of misclassification						
	1	2	3	4	5	6		1	2	3	4	5	6
1	21	3	3	0	4	0	1	29	0	0	1	1	0
2	3	16	0	3	0	0	2	0	22	0	0	0	0
3	2	0	7	1	1	0	3	0	0	9	1	1	0
4	0	1	1	8	0	0	4	0	0	1	9	0	0
5	4	0	3	0	10	2	5	1	0	1	0	16	1
6	0	1	1	0	3	1	6	0	0	0	0	0	6

1=Yamana M; 2=Yamana F;3= Alakaluf M; 4=Alakaluf F; 5=Ona M; 6=Ona F.

Prob. misclassification	number variables	Prob. misclassification	variables selected
0.361	65	0.08	6

Table 6. *Results of the assignation of skulls from group S_3*

Assignation according to the Museum (initial assignation) and biometrical assignation (DB and step-wise). ?= have missing values in the selected variables and no assignation is possible.

Initial assignation	DB assignation	Step-wise assignation
18 skulls	10 are reconfirmed 8 change	?

4. CONCLUSIONS

Palaeontological studies based on quantitative variables often deal with heterogeneous samples that do not belong to empirical biological populations and include incomplete data sets. Furthermore, isolated specimens are usually considered in the comparative analyses establishing phylogenetic relationships. The DB discriminant analysis could be a valuable statistical tool as it works with morphological distances and avoids the missing values problem at the same time. This is particularly useful when substitution of missing values is impossible or inadvisable if a step-wise analysis method were initially chosen, which in this case is actually more efficient than the DB method. Whether substitution of missing values, a combination of both techniques, or direct application of the DB rule is the right choice depends on the information sought, since the three choices respectively present as many advantages as disadvantages. However the above results indicate that it seems that in order to make new assignations, it is better to use the DB-discriminant analysis than a classical step-wise discriminant analysis when there are missing values. So it is clear that the DB-discriminant analysis has some advantages when some values are missing. A summary of some advantages and disadvantages of the DB discriminant method with respect to the classical step-wise method is presented bellow. The DB-rule can use qualitative, quantitative, binary or mixed variables without any transformation. The step-wise method uses quantitative variables, and can also use qualitative variables although a codification as binary variables is needed. If the number of variables is large with respect to the number of individuals, the DB-rule can work with all of them. The step-wise method has to select some of them and this selection is not always optimal. The step-wise method usually gives better allocation for predetermined groups than the DB-rule. With the step-wise method if the new individual to allocate has missing values in the selected variables, assignation is not possible. The DB-rule deals with missing values and can allocate individuals with values of this kind.

5. ACKNOWLEDGEMENTS

We thank Prof. Cuadras for his helpful suggestions and R. Rycroft (S.A.L. University of Barcelona) for the English correction. This work has been supported in part by grants from Spanish DGCYT (PB93-0021 and PB96-1004) and the Generalitat de Catalunya (GRC96-UB2506 and SGR97-00183).

6. REFERENCES

[1] **Arenas, C.** and **Bernal, M.** (1994). «Multivariate approach to the classification of *genus Dianthus L. (Caryophyllaceae)*». *Selected topics on stochastic modelling*, R. Gutierrez y M.J. Valderrama Editores, World Scientific, Singapore.

[2] **Arenas,C.; Cuadras, C.M.** and **Fortiana, J.** (1991). *Multicua. Paquete no estandard de análisis multivariante, versión 0.75.* Publicacions del Departament d'Estadística, n° 4, Barcelona, Spain.

[3] **Arenas, C.; Cuadras, C.M.** and **Fortiana, J.** (1993). *Multicua. Paquete no estandard de análisis multivariante, versión 0.77.* Publicacions del Departament d'Estadística, n° 4, Barcelona, Spain.

[4] **Arenas, C.; Cuadras, C.M.** and **Fortiana, J.** (1998). *Multicua. Paquete no estandard de análisis multivariante, versión 0.77 ampliada.* Publicacions del Departament d'Estadística (nueva colección), n° 1, Barcelona, Spain.

[5] **Cuadras, C.M.** (1989). «Distance analysis in discrimination and classification using both continuous and categorical variables». *Statistical Data Analysis and Inference.* (In: Y. Dodge, ed.) Elsevier North Holland, Amsterdam pp. 459–473.

[6] **Cuadras, C.M.** (1991). «A distance approach to Discriminant Analysis and its Properties». *Universitat de Barcelona, mathematics preprint series*, n° 90.

[7] **Cuadras, C.M.** (1992). «Some examples of distance based discrimination». *Biometrical Letters*, **29 (1)**, 3–20.

[8] **Cuadras, C.M.; Fortiana, J.** and **Oliva, F.** (1997a). «The proximity of an individual to a population with applications to discriminant analysis». *Journal of Classification*, **14**, 117–136.

[9] **Cuadras, C.M.; Atkinson, R.A.** and **Fortiana, J.** (1997b). «Probability densities from distances and discriminant analysis». *Statistics & Probabilities Letters*, **33**, 405–411.

[10] **Gower, J.C.** (1971). «A general coefficient of similarity and some of its properties». *Biometrics*, **27**, 857–874.

[11] **Howells, W.W.** (1973). *Craneal Variation in Man. A study by Multivariate Analysis of Patterns of Difference Among Recent Human Populations.* Papers of the Peabody Museum Harvard University, vol. 67, 259 pp.

[12] **Howells, W.W.** (1989). *Skull Shapes and the Map. Craniometric Analysis in the Dispersion of Modern Homo.* Papers of the Peabody Museum Harvard University, vol.79, 189 pp.

[13] **Lahr, M.M.** (1995). «Patterns of modern human diversification. Implications for Amerindian origins». *Yearbook of Physical Anthropology*, **38**, 163–198.

[14] **Mc Cabe, G.P., Jr.** (1975). «Computations for variable selection in discriminant analysis». *Technometrics*, **17**, 103–109.

[15] **Montanari, A.** and **Mignari, S.** (1994). «Notes on the bias of dissimilarity indices for incomplete data sets: the case of archaelogical classifications». *Qüestiió*, **18**, 39–49.

[16] **Rao, C.R.** (1989). *Statistics and Truth.* International Co-oper. Pub. House, Fairland, USA.

[17] **Turbón, D.** (1995). «Antropología de los aborígenes fueguinos». Estévez J. and Vila A. Eds.: *Encuentros en los conchales fueguinos.* C.S.I.C. Madrid, 275–289.

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 EN EL VOLUM 22 N. 2

PROBLEMA N. 70

1) Demostrem que les distribucions marginals són uniformes

$$H(u, 1) = u^{(1-\theta)}(\min\{u, 1\}) = u^{1-\theta}u^{\theta} = u \qquad o < u < 1$$

$F_U(u) = u, \quad 0 < u < 1$, és la funció de distribució uniforme $(0, 1)$.

Anàlogament $F_V(v) = v, \qquad 0 < v < 1.$

2) Si $\theta = 0$ és $H(u, v) = u \cdot v = F_U(u) \cdot F_V(v) \implies U, V$ són estocàsticament independents.

Si $\theta = 1$ és $H(u, v) = \min\{u, v\}$. Sabem que

$$\begin{aligned} H(u, v) &= P(U \leqslant u, V \leqslant v) \leqslant P(U \leqslant u) \\ H(u, v) &= P(U \leqslant u, V \leqslant v) \leqslant P(V \leqslant v) \end{aligned}$$

Però si $H(u, v) = \min\{u, v\}$, posant $u = v$ tenim a més que

$$H(u, u) = P(U \leqslant u) = P(V \leqslant u)$$

$$\implies [U \leqslant u, V \leqslant u] = [U \leqslant u] = [V \leqslant u] \implies U = V$$

Per tant la correlació entre U i V és 1.

3) $$\begin{aligned} \lim_{u=v\to 1} \frac{H(u, v) - u \cdot v}{\min\{u, v\} - u \cdot v} &= \lim_{u\to 1} \frac{(u^2)^{1-\theta} \cdot u^{\theta} - u^2}{u - u^2} \\ &= \lim_{u\to 1} \frac{u^{2-\theta} - u^2}{u - u^2} \\ &= \lim_{u\to 1} \frac{(2-\theta)u^{1-\theta} - 2u}{1 - 2u} = \theta \end{aligned}$$

C.M. Cuadras
Universitat de Barcelona

PROBLEMA N. 71

Sigui $Y = a_1 X_1 + \cdots + a_p X_p$ una combinació lineal de $X = (X_1, \ldots, X_p)$. Sigui $\boldsymbol{a} = (a_1, \ldots, a_p)'$, $\boldsymbol{u}_i = (0, \ldots, 1, \ldots, 0)'$ un vector unitari. Es verifica:

$$\sum_{i=1}^{p} \boldsymbol{u}_i \boldsymbol{u}_i' = \boldsymbol{I}$$

on $\boldsymbol{I}$ és la matriu identitat, i

$$\boldsymbol{u}_i' \boldsymbol{R} \boldsymbol{u}_i = 1 \qquad i = 1, \ldots, p.$$

Podem suposar les variables X_i amb variància 1. La correlació (al quadrat) entre X_i, Y és

$$\rho^2(X_i, Y) = \frac{(\boldsymbol{u}_i' \boldsymbol{R} \boldsymbol{a})^2}{(\boldsymbol{u}_i' \boldsymbol{R} \boldsymbol{u}_i)(\boldsymbol{a}' \boldsymbol{R} \boldsymbol{a})} = \frac{(\boldsymbol{u}_i' \boldsymbol{R} \boldsymbol{a})^2}{(\boldsymbol{a}' \boldsymbol{R} \boldsymbol{a})}$$

Tenim que:

$$\begin{aligned} \sum_{i=1}^{p} (\boldsymbol{u}_i' \boldsymbol{R} \boldsymbol{a})^2 &= \sum_{i=1}^{p} (\boldsymbol{a}' \boldsymbol{R} \boldsymbol{u}_i)(\boldsymbol{u}_i' \boldsymbol{R} \boldsymbol{a}) \\ &= \boldsymbol{a}' \boldsymbol{R} \left(\sum_{i=1}^{p} \boldsymbol{u}_i \boldsymbol{u}_i' \right) \boldsymbol{R} \boldsymbol{a} \\ &= \boldsymbol{a}' \boldsymbol{R} \boldsymbol{I} \boldsymbol{R} \boldsymbol{a} \\ &= \boldsymbol{a}' \boldsymbol{R}^2 \boldsymbol{a} \end{aligned}$$

La suma de correlacions (al quadrat) és

$$\sum_{i=1}^{p} \rho^2(X_i, Y) = \frac{\boldsymbol{a}' \boldsymbol{R}^2 \boldsymbol{a}}{\boldsymbol{a}' \boldsymbol{R} \boldsymbol{a}}$$

El màxim d'aquest quocient és el més gran valor propi λ_1 de $\boldsymbol{R}^2$ respecte de $\boldsymbol{R}$

$$\boldsymbol{R}^2 \boldsymbol{a}_1 = \lambda_1 \boldsymbol{R} \boldsymbol{a}_1 \quad \Longrightarrow \quad \boldsymbol{R} \boldsymbol{a}_1 = \lambda_1 \boldsymbol{a}_1$$

Per tant, $\boldsymbol{a}_1$ és el primer vector propi de $\boldsymbol{R}$, $Y = Y_1$ és la primera component principal i

$$\sum_{i=1}^{p} \rho^2(X_i, Y_1) = \frac{\boldsymbol{a}_1' \boldsymbol{R}^2 \boldsymbol{a}_1}{\boldsymbol{a}_1' \boldsymbol{R} \boldsymbol{a}_1} = \frac{\lambda_1^2}{\lambda_1} = \lambda_1$$

és màxim.

C.M. Cuadras
Universitat de Barcelona

PROBLEMES PROPOSATS

PROBLEMA N. 72

Demostrar que si X, Y son variables aleatorias independientes igualmente distribuidas, entonces la distribución de $X-Y$ no puede ser uniforme en el intervalo $(-1, +1)$.

C.M. Cuadras
Universitat de Barcelona

PROBLEMA N. 73

Sea $H(x, y)$ una función de distribución bivariante continua, con marginales $F(x) = H(x, \infty)$, $G(y) = H(\infty, y)$.

Se cumple la desigualdad

$$H^-(x, y) \leqslant H(x, y) \leqslant H^+(x, y)$$

donde

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

son las llamadas *cotas de Fréchet.* Probar las siguientes relaciones entre H^- y H^+

$$\begin{aligned} H^+(x, y) &= F(x) - H^-\left(x, G^{-1}(1 - G(y))\right) \\ H^+(x, y) &= G(y) - H^-\left(F^{-1}(1 - F(x)), y\right) \end{aligned}$$

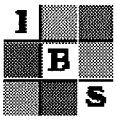
C.M. Cuadras
Universitat de Barcelona

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

Presidente:	Emilio A. Carbonell Guevara
Vicepresidente:	Guadalupe Gómez i Melis
Secretario y Tesorero:	Fernando López Santoveña
Vocal en calidad de Miembro del Consejo de la IBS:	Carles Cuadras Avellana
Vocales:	Rosa Estarelles Rodríguez
	Martin Ríos Alcolea
	Juan Luis Chorro Gascó
	José Luis González Andújar
	Alex Sánchez Pla
	María Jesús Bayarri García
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 el primer semestre de 1998 se ha celebrado un curso sobre «Regresión Logística» y otro sobre «Modelos Mixtos con S-Plus».

Actualmente se está estudiando la programación de dos nuevos cursos: «Modelos de Supervivencia» y «Métodos Gráficos de Análisis de Datos».

VII Conferencia Española de Biometría
Reunión bianual de la Sociedad Española de Biometría
10-12 de marzo, 1999. Palma de Mallorca.
http://iris1.uib.es/people/biome99/biometria99@clust.uib.es

The TES Institute
Training of European Statisticians

TES OR THE ENHANCEMENT OF STATISTICS

The TES Institute is a non-profit association created in November 1996 by the Directors General of the National Statistical Institutes of ten Member States of the European Union and of the four Member States of the European Free Trade Association.

As an international post-graduate vocational training institute for statisticians, the mission of the TES Institute is to create truly European vocational training and staff development opportunities at post-graduate level. The annual training programmes are designed for target groups ranging from young statisticians to executives of National Statistical Institutes.

Over the last five years, the objectives of the TES Institute were *(a)* to provide training courses and seminars of short duration at post-graduate level, thus offering statisticians opportunities to perfect their skills, *(b)* to provide a forum for mutual consultation on vocational training for statisticians, *(c)* to give assistance in the training for new European statistical projects, thus disseminating standards, European methods and classifications, and to help future members to prepare for their integration into the European Statistical System, and *(d)* to promote exchanges of skills and experiences.

TYPES OF TRAINING

The TES Institute offers two types of training: the *Core Programme* and the *Special Courses Programme*. Both of them are covering the following specialisation areas:

(1) *Data Collection and Survey Methodology*, (2) *Economic Statistics*, (3) *Social Statistics* and (4) *Publication, Dissemination and Use of Statistics*, complemented by the *Common Courses* (general «statistical» culture) and the *Support Courses* (statistical analysis, computing techniques and statistical management).

The Core Programme

The annual training programmes are designed for public sector statisticians in the Member States of the EU and EFTA. However, the programmes are also open to the National Statistical Institutes of countries in Central and Eastern Europe, the Mediterranean Basin, some other selected countries and private sector statisticians.

The leading principle for the definition of the annual programmes and the detailed content of each of the courses is that choices are made on the basis of the training needs in the Member States of the EU and the EFTA. The *Core Programme* is subsidised by the Statistical Office of European Commission (Eurostat).

The Special Courses Programme

The *Special Courses Programme* is designed for public sector statisticians from countries outside the EU and the EFTA. At this moment its main clients are statisticians from the Central European countries and the countries of the Mediterranean Basin. The *Special Courses Programme* is not subsidised.

The *Special Courses Programme* consists of courses which are either repeats of existing courses in the *Core Programme* or tailor-made courses at the request of a country or a group of countries. Whenever a course from the *Core Programme* is repeated, the content is reviewed and adapted to the specific needs of the participants.

TRAINING ORGANISATION

Obviously, vocational training programmes of this kind imply the involvement of different groups of people.

- The training staff of the TES courses consists of *an international pool of trainers* (highly qualified practitioners and university professors). This pool is responsible for the design and the execution of the courses and the development of the course material, which are annually reassessed and updated where necessary. Each year the organisation of the programme requires the intervention of about hundred trainers.

- The National Statistical Institutes of the European Union (EU), the European Free Trade Association (EFTA), the Central and Eastern European Countries (ECO[1]) and the Mediterranean Basin countries[2] have nominated *TES Correspondents* for the TES programme. This network of TES Correspondents is responsible for the communication between their Institute and the TES Institute, for the co-ordination of the registration of candidates and for the dissemination of information about TES courses.
- The main task of the staff of the TES Institute is to assist in the design, to organise and to give logistic support to the international vocational training programme and to provide information about existing training possibilities in the various countries in Europe. To carry out these activities, the TES Institute comprises a permanent team of eleven people.

CURRENT ACTIVITIES

Training Core and Special Courses Programmes

Over the last five year, the TES Institute organised some 110 courses in the framework of the *Core Programme* open to EU, EFTA and ECO statisticians. Besides, *Special Courses* have also been organised especially for statisticians from ECO countries.

On the whole, about 600 participants have been trained every year through these two kind of training programme.

Moreover, since 21st May 1997, when the first Eurostat *Task Force on Training* for MEDSTAT countries was organised, the TES Institute officially extended its training activities to the Mediterranean Basin countries.

Ad hoc projects

Recently, a seminar on «Education Statistics» has been organised at the request of Eurostat (Vienna, January 1998).

Moreover, two ad hoc courses (5 days) for the statistical service of Malta: the course «Enterprise Statistics» and the course «National Accounts Statistics in Practice». Both courses were executed end 1997.

---

[1] Albania, Bulgaria, Czech Republic, Estonia, Hungary, FYROM (Former Yugoslav Republic of Macedonia), Latvia, Lithuania, Poland, Romania, Slovakia and Slovenia.

[2] Algeria, Cyprus, Egypt, Israel, Jordan, Lebanon, Malta, Morocco, Palestine, Syria, Tunisia, Turkey.

In a near future, the TES Institute will carry out:

- Two seminars on «Statistics and Policy Making» at the request of the Palestinian Statistical Training Centre (Gaza, Westbank). These seminars will be financed through UN funds and are carried out in co-operation with UNITAR, the UN Institute for Training and Research, Geneva. Again this seminar may be introduced, in a modified form, in the future regular programme.
- The first part of a «Summer School on Social Statistics» in July 1998 ($5\frac{1}{2}$ days) in Siena. The other two parts are planned for summer 1999 and 2000.
- Specific training will be provided to statisticians from the Ukrainian Statistical Office. The training will cover *Business Registers, Enterprise Statistics and National Accounts.*

Consulting

Apart from the well-known training activities, the TES Institute has recently set up a procedure for consulting activities.

In this context, the TES Institute is requested to provide support to the set-up of a training centre for statisticians in a specific country in Ukraine. This support will consist of *(a)* training of future trainers, as above mentioned and *(b)* consulting for curriculum development for their future training programmes.

Publication

Articles The TES Institute is regularly publishing articles on its current activities in periodic Newsletter of several Statistical Institutes and some statistical journals as *Qüestiió* (Quaderns d'Estadística i Investigació Operativa), edited by the Institut d'Estadística de Catalunya.

Manuals

The TES Institute decided to set up a TES Manual series for the manuals to be developed in the specialisation areas covered by the vocational training programme (i.e. *Data Collection and Survey Methodology, Economic Statistics, Social Statistics and Publication, Dissemination and Use of Statistics*).

By the end of this year, the following publications will be available at the TES Institute:

- «*Introduction to Demographic Data Analysis*» by Mr Otto Andersen from Statistics Denmark. This manual will cover subjects like: data in demography, lexis diagram, fertility, mortality, etc. As the manual will be available at the TES Institute.

- «*Index Number Theory and Practice*» by Prof. Marco Martini from the University of Milano.

Please inform the TES Institute if you are interested in the above mentioned publications (articles and/or manuals) and detailed information will be forwarded to you.

Other activities

The TES Institute has recently been associated with the Maastrich School of Management as training and advice provider in the framework of their MBA.

In order to receive an informative leaflet on the MBA, please address your request to the TES Institute (see below for contact person). Further requests will be treated by MSM directly.

GENERAL INFORMATION

For further information about courses, new brochure, publications, specific projects and so on, please contact Ms Valérie Vandewalle:

by phone:	+352/29 85 85 34
fax:	+352/29 85 29/30
e-mail:	vvandewalle@tes-institute.lu
website:	http://www.tes.institute.lu.institute
for mail:	5, rue Guillaume Kroll
	L - 1882 Luxembourg

Activities of the TES Institute from January to June 1999

The new 1998-1999 training programme is planned from January 1998 to June 1999 and comprises over 25 courses organised whole over Europe.

The planning of courses until beginning 99 is the following:

- The first course of next year will deal with the *theory and practice of Regional Accounts*. Mr. Grünewald (Eurostat) will hold the course in English in the premises of the regional institute of Münich from 11 to 13 January 99.
- The TES Institute will host the course of Mr. Lancetti (Eurostat) who will present in English the principles of a *survey for the service sector* from 18 to 20 January. Through this course, participants will be provided with pragmatic ideas and methods

for the implementation of surveys. Particular emphasis will be put on enterprise statistics for the service sector.

- From 25 to 28 January, Mr. Sautory (INSEE) will train on the techniques for using auxiliary information to improve estimation of data acquired by sampling surveys and thus improve estimation techniques. The course will be held in English in the premises of INSEE-Alsace.
- A new course on *Regression modelling for Complex Surveys data* will be organised in Luxembourg from 27 to 29 January. Through this course, Mr. Skinner (University of Southampton) will introduce the principles and practice of regression modelling for survey data obtained using complex sampling scheme. Participants will also be introduced to the implementation of appropiate methods in the STATA software package.

Please note that all the courses offered by the TES Institute are divided between both theoretical and practical sessions. The lectures will provide the theoretical background necessary for further developments and the practical sessions will offer you the opportunity to discuss and exchange experience with colleagues from other countries and other working areas.

WORKSHOP
GIE

Jornades internacionals sobre

Generació d'informació estadística: qualitat i limitacions

Barcelona,
30 de novembre - 1 de desembre de 1998

XARXA TEMÀTICA

ENQUESTES I QUALITAT DE LA INFORMACIÓ ESTADÍSTICA

Col·laboren:

- Institut d'Estadística de Catalunya (IDESCAT)
- CIRIT, Generalitat de Catalunya
- Service pour la Science et la Technologie de l'Ambassade de France

Organitzen:

- Departament d'Econometria, Estadística i Economia Espanyola, Universitat de Barcelona
- Departament d'Estadística i Investigació Operativa, Universitat Politécnica de Catalunya

Generació de la informació estadística: qualitat i limitacions

La preocupació per la qualitat de les dades en les enquestes i en la recollida d'informació estadística és essencial en la producció de resultats fiables. L'estudi dels fenòmens socio-econòmics no pot ignorar la rellevància de la generació de les dades i els límits de la utilització de la informació estadística recollida per sondeig o per qualsevol altra font.

Es pretén donar al Workshop un enfocament pràctic i afavorir la sinergia entre investigadors i professionals del sector públic i de l'empresa.

A qui s'adreça el Workshop?

Aquestes jornades s'adrecen a estadístics i usuaris de l'estadística, investigadors i professionals del món institucional o empresarial, preocupats per la qualitat i fiabilitat de la informació estadística.

Comitè de programa:

Manuela Alcañiz (UB, Barcelona)
Tomàs Aluja (UPC, Barcelona)
Joan M. Batista (URL, Barcelona)
Mónica Bécue (UPC, Barcelona)
Germà Coenders (UG, Girona)
Michael Greenacre (UPF, Barcelona)
Montse Guillén (UB, Barcelona)
Enric Ripoll (IDESCAT, Barcelona)
Frederic Utzet (UAB, Bellaterra)
Sylvie Viguier-Pla (U. Perpignan, Perpignan)

Informació i inscripcions:

GIE Mónica Bécue
Dept. d'Estadística i Investigació Operativa
Facultat de Matemàtiques i Estadística (UPC)
Pau Gargallo, 5; 08028 Barcelona
Fax +34 93 401 51 55; e-mail: monica@eio.upc.es

PROGRAMA

Dilluns, 30 de novembre

09:00 Recepció dels participants

10:15 Presentació de les jornades
Jordi Oliveras, Director de l'Institut d'Estadística de Catalunya (IDESCAT)

Màrius Rubiralta, Vicerector de Recerca de la Universitat de Barcelona (UB)

Antoni Marí, Vicerector de Recerca de la Universitat Politècnica de Catalunya (UPC)

George Carry, Agregat cultural de l'Ambaixada Francesa

10:45 **Pausa-Cafè**

Sessió 1 Disseny i el·laboració de mostres estadístiques

11:00 Utilisation d'information auxiliaire dans les échantillons statistiques.
Yves Tillé, ENSAI-INSÉE-CREST

11:45 Diseño de muestras en encuestas de población y hogares.
Juana Porras, INE

12:30 Diseño de encuestas de opinión.
Valentín Martínez, CIS

Sessió 2 Operacions censals i explotació de registres administratius

15:00 Utilisation des données des recensements et des registres pour les études démographiques.
Laurent Toulemon, INSÉE-INED

15:45 Utilización de la documentación aduanera y la declaración INTRASTAT para la elaboración de la balanza comercial y la estadística de comercio exterior.
Reyes Crespo, D.G. Aduanas

16:30 **Pausa-Cafè**

16:45 Aprofitament estadístic dels registres administratius de naturalesa fiscal a Catalunya.
Àlex Costa, IDESCAT

17:30 Comunicacions lliures

Dimarts, 1 de desembre

Sessió 3 Disseny de qüestionaris i recollida de dades

09:00 Design of questionnaires.
Willem Saris, Universitat de Amsterdam

09:45 Qualité de l'information dans les enquêtes.
Ludovic Lebart, CNRS.ENST

10:30 **Pausa-Cafè**

11:00 Punts forts i dèbils de la investigació per enquesta.
Enric Renau, UPF-Institut DEP

11:45 Comunicacions lliures

Sessió 4 Dades longitudinals i dades de panels

15:00 Exclusion sociale et abstention électorale: utilisation du fichier électorel et de l'échantillon démographique permanent pour le suivi des individus qui s'abstiennent systématiquement dans les votes électoraux.
François Heran, INSÉE-INED

15:45 Les objectifs et problèmes de la collecte de données longitudinales et biographiques.
Laurent Toulemon, INSÉE-INED

16:45 **Pausa-Cafè**

16:30 La inserció professional dels universitaris: estudi longitudinal del seu rendiment en els anys d'estudis universitaris i itineraris professionals al llarg dels cinc primers anys de treball.
Josep Masjoan i ***Jesús Vivas***, Departament de Sociologia, Universitat Autònoma de Barcelona

17:15 Comunicacions lliures

Lloc:

Palau de les Heures
Universitat de Barcelona
Campus de la Vall d'Hebron
Metro: L3 (Estació Montbau)
Autobusos: 27, 60, 73, 76, 85, 173

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 LaTeX 2ε [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 LaTeX 2ε 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

nQuery Advisor® Release 2.0

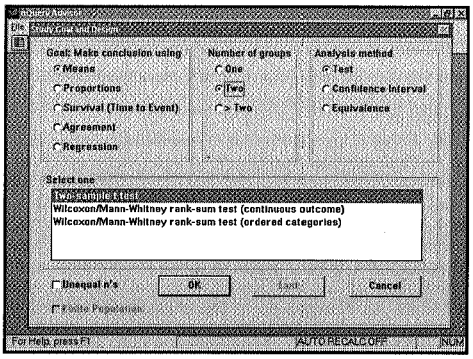
STUDY PLANNING SOFTWARE

Sample size and Power determination

nQuery Advisor is a very useful aid to investigators and statisticians in planning research studies to ensure efficient use of resources. Determination of the correct sample size and power are important aspects of study planning.

nQuery Advisor provides simple, reliable, and efficient methods for determining these values. It provides expert computational assistance and extensive coverage of sample size and power determination problems as well as sample size answers for confidence interval and equivalence analyses. It allows for either equal or unequal group sample sizes for most problems.

nQuery Advisor Release 2.0 features major new enhancements designed to make planning your research studies even easier. New tables have been included with expanded options for planning sample sizes for survival studies, sampling from finite populations, cluster sampling, and for analyses using nonparametric tests. The manual and on-line help have been expanded providing more details on the use of nQuery Advisor for repeated measures and crossover designs, equivalence demonstrations, and survival analysis.



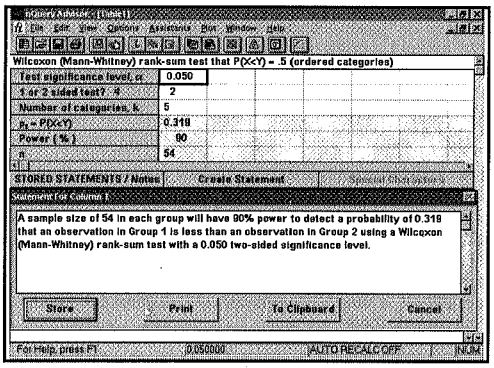
Among the new features introduced in version 2.0 are Non-Parametric tests, and sampling from a finite population.

For researchers planning clinical trials or other studies with survival or time to an event as an endpoint, nQuery Advisor Release 2.0 introduces a new sample size table for Survival Analysis. This table allows user specification of survival curves which are not exponential in shape, or hazard ratios which are not constant throughout, and user specification of complex patterns of accrual or dropout.

Social science researchers will appreciate the new options for planning studies for sampling from a finite population. nQuery Advisor Release 2.0 now provides sample size tables with a finite population correction for a range of study designs with testing and confidence interval goals.

The design of community intervention studies often requires randomization of clusters of subjects. nQuery Advisor Release 2.0 assists in estimation of the cluster standard deviation to use in determining the number of clusters which need to be sampled.

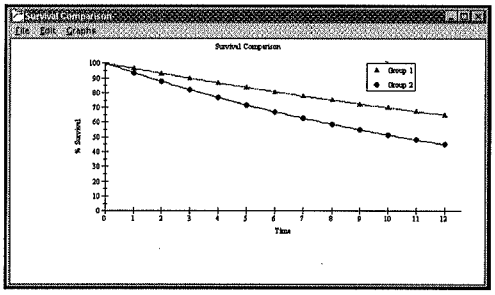
nQuery Advisor Release 2.0 also includes sample size computations when the Wilcoxon/Mann-Whitney Rank-Sum test will be used to analyze results of a two-group study.



Protocol-ready sample size justification statements for the computed results.

Operating in a Windows environment nQuery Advisor® provides:

- A structured index to assist in specifying the research goal
- Spreadsheet-style tabular entry and display of sample size and power
- Guide cards to assist users with statistical and practical questions on input
- Computational aids for specifying effect sizes and estimates of variability
- Plots showing the relationship between power and sample size
- Protocol-ready sample size justification statements for the computed results
- Ease of comparison of results under different conditions or analyses
- Formulas and algorithms chosen on the basis of comparative evaluations
- Access to distribution functions for expert users



nQuery Advisor® Release 2.0 now allows user specification of survival curves.

nQuery Advisor® is developed by Dr. Janet Elashoff Ph.D.of Dixon Statistical Associates, Los Angeles, U.S.A.

Statistical Solutions Ltd.
8 South Bank
Crosse's Green, Cork, Irland

Tel. + 353 21 319629 • Fax + 353 21 319630
Email: sales@statsol.ie
Website: http://www.statsol.ie*

Statistical Solutions
Stonehill Corporate Center Suite 104
999 Bradway, Saugus, MA 01906

Tel. (781) 2317680 • Fax (781) 2317684
Email: info@statsolusa.com
Website: http://www.statsolusa.com

nQuery Advisor® Versió 2.0

Programari de Planificació d'Estudis

Determinació de la dimensió i la potència de la mostra

nQuery Advisor és una eina molt útil per a investigadors i estadístics a l'hora de planificar els estudis d'investigació i de garantir un ús eficaç dels recursos.

La correcta determinació de la dimensió i de la potència de la mostra representa un aspecte important dins de la planificació dels estudis.

nQuery Advisor proporciona uns mètodes simples, fiables i eficaços per determinar aquests valors, a més d'una experta assistència en computació i una àmplia cobertura quant als problemes relacionats amb la dimensió i la potència de les mostres, així com les respostes sobre la dimensió de les mostres per a les anàlisis de l'interval de confiança i d'equivalència. En la major part dels problemes que es presenten permet que els grups posseeixin una dimensió igual o desigual.

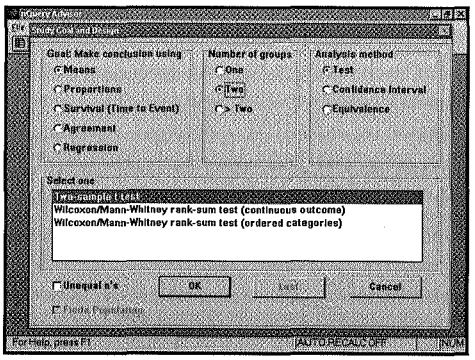
La versió 2.0 de nQuery Advisor compta amb notables millores dissenyades per tal de facilitar, encara més, la planificació dels seus estudis d'investigació. S'han inclòs noves taules amb opcions expandides per a la planificació de les dimensions de les mostres en els estudis de supervivència, el mostreig de poblacions finites, el mostreig per grups i, també, les anàlisis que utilitzen proves no paramètriques. Tant el manual com l'ajuda en pantalla s'han ampliat per proporcionar més detalls sobre l'ús de nQuery Advisor en mesures repetides i dissenys encreuats, demostracions d'equivalència i anàlisis de supervivència.

Per aquells investigadors que estan planificant assajos clínics o altres estudis amb supervivència del temps a un esdeveniment com a punt final, nQuery Advisor versió 2.0 introdueix una nova taula de dimensions de mostres per a les anàlisis de supervivència. Aquesta taula permet a l'usuari especificar unes corbes de supervivència que posseeixen una forma no exponencial o bé ràtios de risc que no es mantenen sempre constants, a més d'especificar uns patrons complexos d'acumulació o d'abandonament.

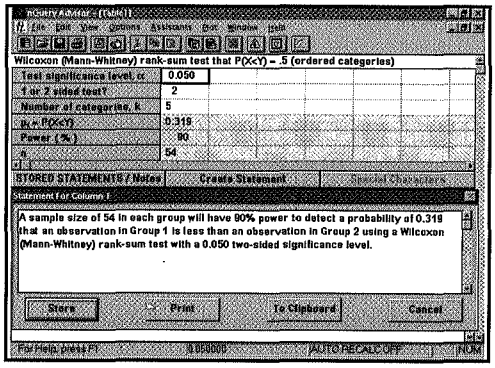
Els investigadors de ciències socials sabran, sens dubte, apreciar les noves opcions per a la planificació dels estudis basats en una població finita. La versió 2.0 de nQuery Advisor ajuda a estimar la desviació estàndard de grups per poder determinar el nombre de grups sobre els quals cal fer el mostreig.

La versió 2.0 de nQuery Advisor incorpora a més computacions de dimensions de mostra en pretendre utilitzar la proba de rang-suma de Wilcoxon/Mann-Whitney per analitzar els resultats d'un estudi de dos grups.

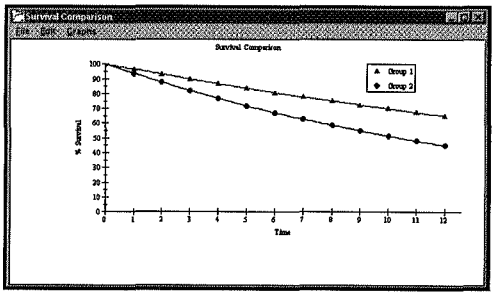
Quan s'utilitza en un entorn Windows nQuery Advisor, proporciona: Un índex estructurat per tal de facilitar l'especificació de l'objectiu de la investigació; Entrades i visualització de la dimensió i potència de mostres realitzades similars a les dels fulls de càlcul; Fitxes guia per ajudar els usuaris amb les seves preguntes estadístiques i pràctiques sobre les entrades; Eines de computació per especificar les dimensions d'efecte i les estimacions de variabilitat; Traces que mostren la relació entre la potència i la dimensió de les mostres. Quant als resultats computats proporciona informes de justificació de les dimensions de la mostra preparats per a la seva protocol_lització; facilitat en comparar els resultats sota diferents condicions o anàlisis; fórmules i algoritmes seleccionats d'acord amb les avaluacions comparatives i accés a les funcions de distribució per als usuaris experts.



Les noves prestacions introduïdes en la versió 2.0 inclouen proves no paramètriques i mostreigs en funció de les poblacions finites.



Per als resultats computats, s'han preparat informes de justificació de la dimensió de la mostra per a la seva protocol_lització.



nQuery Advisor versió 2.0 permet a l'usuari especificar ja les corbes de supervivència.

nQuery Advisor és un producte desenvolupat per la Dra. Janet Elashoff de Dixon Statistical Associates, Los Angeles, EE.UU.

Statistical Solutions Ltd.
8 South Bank
Crosse's Green, Cork, Irland

Tel. + 353 21 319629 • Fax + 353 21 319630
Email: sales@statsol.ie
Website: http://www.statsol.ie*

Statistical Solutions
Stonehill Corporate Center Suite 104
999 Bradway, Saugus, MA 01906

Tel. (781) 2317680 • Fax (781) 2317684
Email: info@statsolusa.com
Website: http://www.statsolusa.com

solas *For Missing Data Analysis 1.0*

HOW DO YOU HANDLE MISSING DATA?

Many advanced studies have suffered from the use of ad-hoc methods of handling missing data such as "complete-case-analysis" or " available-case analysis". These methods are becoming increasingly viewed as biased by regulatory agencies. According to the types of missing data (bounded, univariate or multivariate) in your dataset SOLAS allows you to choose the most appropriate techniques to apply to a subset or all of your data.

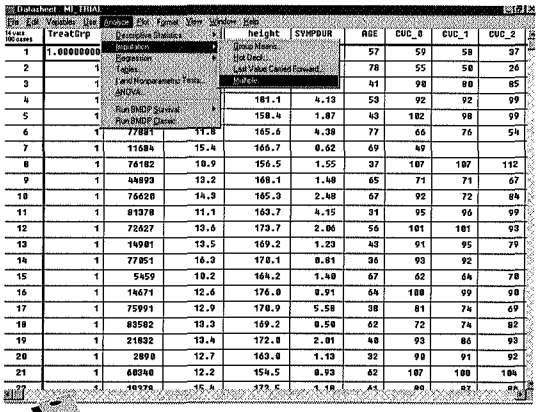
solas *For Missing Data Analysis* , a fully Windows based statistical software package, incorporates the latest industry accepted techniques for treating missing values. Imputation techniques include :

Multiple Imputation

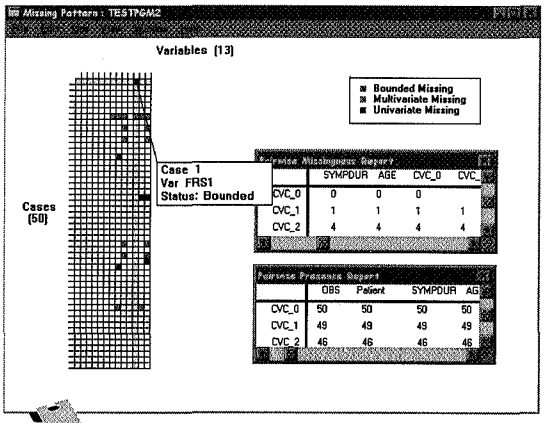
- Based on techniques developed by Rubin et Al.
- Generates multiple imputations using propensity scores and the approximate Bayesian bootstrap.
- Allows for analysis of longitudinal data.

Single Imputation

- Last Value Carried Forward - allows the user to impute Missing Value based on last value observed
- Group Means - Imputation of group mean (or mode in the case of categorical data)
- Hot Decking - Allows the user to specify and prioritize all combinations of variables for 'closest match' imputation



Select one of the four imputation methods available in SOLAS for Missing Data Analysis



Identify the type of missing values using the Missing Pattern Report

Solas Missing Data Pattern

The unique missing data pattern in Solas provides the user with a clear over-view of the quantity, positioning and types of missing values in each cell. This feature allows you to further isolate particular cells in a data set and identify observation and status defaults. Following imputation this colour coded graphics feature easily identifies the treatment method selected.

Statistical Solutions Ltd.
8 South Bank
Crosse's Green, Cork, Irland

Tel. + 353 21 319629 • Fax + 353 21 319630
Email: sales@statsol.ie
Website: http://www.statsol.ie*

Statistical Solutions
Stonehill Corporate Center Suite 104
999 Bradway, Saugus, MA 01906

Tel. (781) 2317680 • Fax (781) 2317684
Email: info@statsolusa.com
Website: http://www.statsolusa.com

solas *l'anàlisi de les dades no disponibles*

COM TRACTAR LES DADES NO DISPONIBLES?

Són molts els estudis avançats que han estat perjudicats per l'ús de mètodes puntuals en el tractament de dades no disponibles, com ara per exemple les "anàlisis de cas global" o les "anàlisis de cas disponible". Els organismes de control consideren aquests mètodes cada cop més esbiaixats. D'acord amb l'índole de dades que no es troben disponibles (ja siguin acotades, univariants o multivariants), dins d'aquest conjunt de dades, Solas permet escollir entre les tècniques més adients per aplicar a un subconjunt de les dades o a totes en conjunt.

SOLAS, l'anàlisi de les dades no disponibles, és un paquet de programari estadístic completament integrat dins l'entorn Windows i inclou les últimes tècniques reconegudes per la indústria en els tractaments dels valors no disponibles. Les tècniques d'imputació inclouen:

Imputació múltiple

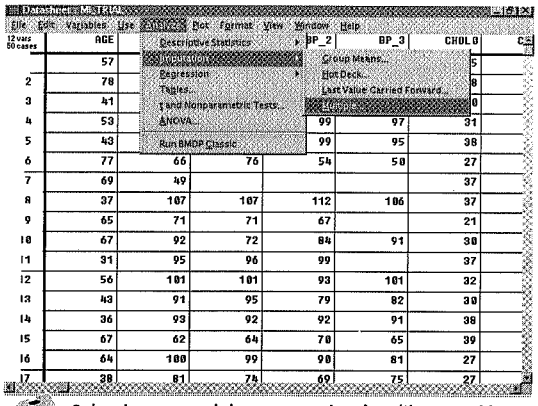
- Basada en tècniques desenvolupades per Rubin et altri.
- Genera imputacions múltiples en utilitzar unes puntuacions de propensió, així com la rutina bayesiana d'inicialització més adient.
- Permet l'anàlisi de les dades longitudinals.

Imputació senzilla

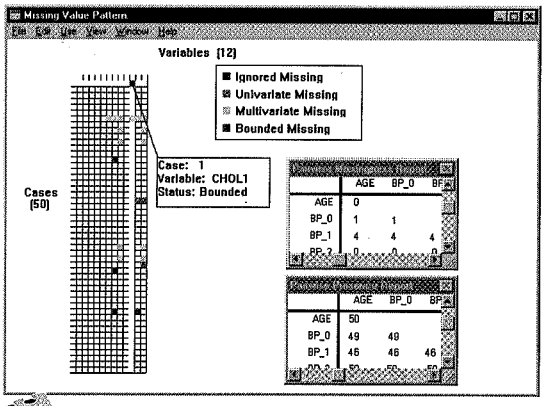
- Últim valor transferit: permet a l'usuari imputar el valor que falta en funció de l'últim valor observat.
- Mitjana de grup: permet la imputació d'una mesura de grup (o d'una modalitat en cas que les dades siguin categòriques).
- Hot decking: permet a l'usuari especificar i prioritzar totes les combinacions de variable per a una imputació del tipus "correspondència més propera".

Patró SOLAS de dades no disponibles

El patró de dades no disponibles és una exclusiva de Solas i ofereix a l'usuari una visualització clara i global de la quantitat, la localització i els tipus de dades no disponibles a cadascuna de les cel_les. Aquesta prestació ofereix la possibilitat d'aïllar, encara més, les cel_les individuals d'un conjunt de dades i identificar-ne les configuracions per defecte d'observació i d'estat. Un cop acabada la imputació aquest dispositiu gràfic, codificat amb colors, identifica amb facilitat el mètode de tractament seleccionat.



Seleccioneu un dels quatre mètodes d'imputació disponibles amb SOLAS per a l'anàlisi de les dades no disponibles.



Determineu a quin tipus pertanyen els valors no disponibles utilitzant l'Informe de Patró de Dades No Disponibles.

Statistical Solutions Ltd.
8 South Bank
Crosse's Green, Cork, Irland

Tel. + 353 21 319629 • Fax + 353 21 319630
Email: sales@statsol.ie
Website: http://www.statsol.ie*

Statistical Solutions
Stonehill Corporate Center Suite 104
999 Bradway, Saugus, MA 01906

Tel. (781) 2317680 • Fax (781) 2317684
Email: info@statsolusa.com
Website: http://www.statsolusa.com