



MEMORIA DEL XI FORO NACIONAL DE ESTADISTICA

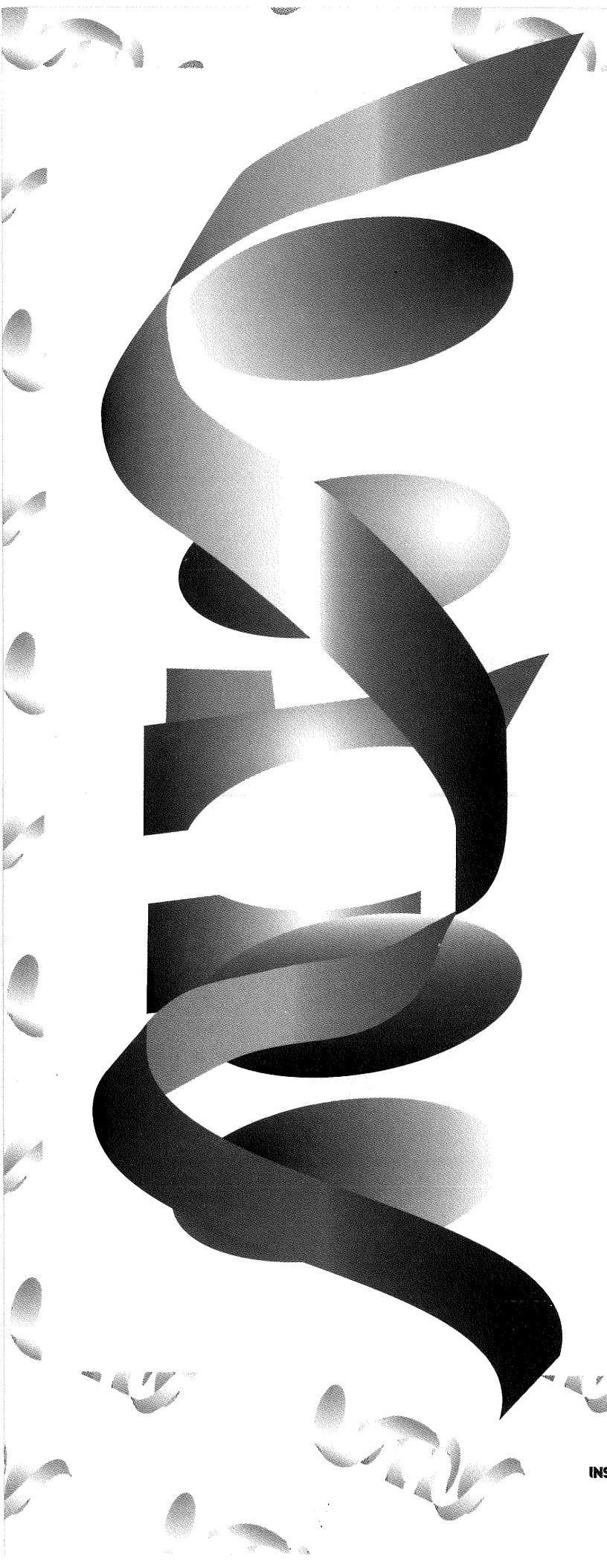
RESUMENES IN EXtenso

UNIVERSIDAD AUTONOMA DE SINALOA
CULIACAN, SINALOA, DEL 21 AL 25 DE OCTUBRE DE 1996



INSTITUTO NACIONAL DE ESTADISTICA
GEOGRAFIA E INFORMATICA





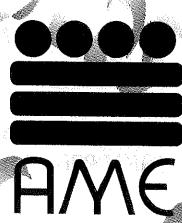
MEMORIA DEL XI FORO NACIONAL DE ESTADÍSTICA

RESUMENES IN EXTOSO

UNIVERSIDAD AUTONOMA DE SINALOA
CULIACAN, SINALOA, DEL 21 AL 25 DE OCTUBRE DE 1996



**INSTITUTO NACIONAL DE ESTADÍSTICA
GEOGRAFÍA E INFORMATICA**



DR © 1997, **Instituto Nacional de Estadística,
Geografía e Informática**
Edificio Sede
Av. Héroe de Nacozari Núm. 2301 Sur
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.

Dirección Internet
<http://www.inegi.gob.mx>

**Memoria del XI Foro Nacional de Estadística
Resúmenes in Extenso**

Impreso en México
ISBN 970-13-1660-6

Presentación

Durante la semana del 21 al 25 de Octubre de 1996 tuvo lugar, en la Cd. de Culiacán, Sinaloa, el XI Foro Nacional de Estadística teniendo como institución sede a la Universidad Autónoma de Sinaloa. En esta ocasión el tema central de dicho evento fue Bioestadística, aunque los trabajos que se presentaron abordaron temas muy diversos sobre el desarrollo y la aplicación de métodos estadísticos.

En esta Memoria se recopilan, en versión resumida, algunos de los trabajos que se presentaron ya sea como Conferencia Invitada, o como Contribución Libre. Cabe mencionar que estos trabajos fueron sometidos a un proceso de revisión pero no de arbitraje.

Agradecemos a Elida Estrada Barragán por la labor tipográfica así como a Francisco Vázquez Jaime por su apoyo computacional, especialmente en el manejo de las figuras. También expresamos nuestro agradecimiento a las diferentes instituciones que brindaron su apoyo e hicieron posible la realización del Foro. Finalmente, la publicación de esta Memoria no hubiera sido posible sin el apoyo del Instituto Nacional de Estadística, Geografía e Informática.

El Comité Editorial

PRESENTACIÓN	iii
CONFERENCIAS INVITADAS.....	1
UTILIZACIÓN DE LA ESTADÍSTICA EN LA MEDICINA VETERINARIA Y EN LA ZOOTECNIA	
<i>Ruiz López, F.J. y Moro Méndez, J.</i>	3
INFERENCE FOR ALMOST PERIODICALLY CORRELATED PROCESSES	
<i>Leskow, J.</i>	7
FISHER AND THE THEORY OF MAXIMUM LIKELIHOOD	
<i>Sprott, D.A.</i>	15
CONTRIBUCIONES LIBRES	25
FUNCIÓN DE VALOR EXTREMO Y SUS APLICACIONES	
<i>Alegria, A. y Soto, H.</i>	27
EL IMPACTO DE LAS CONDICIONES SOCIOECONÓMICAS DE LA POBLACIÓN EN SUS PREFERENCIAS ELECTORALES	
<i>Arroyo Martínez, E. y De la Fuente, L.</i>	32
ESTUDIO MULTICÉNTRICO SOBRE RESISTENCIA A LOS ANTIBIÓTICOS EN HOSPITALES DE TERCER NIVEL EN EL DISTRITO FEDERAL	
<i>Benavides Plascencia, L. y Aldama Ojeda, A.</i>	37
UNA APLICACIÓN DE ESTIMADORES DE MÍNIMOS CUADRADOS EN UN MODELO DE RESPUESTA INMUNOLÓGICA	
<i>Camacho, S., Cervantes, F., Hoyos, L.F. y Romero, J.C.</i>	41
PROGRAMACIÓN ESTOCÁSTICA: UNA ALTERNATIVA AL ESTUDIO DE CONGLOMERADOS	
<i>De los Cobos, S.G., Pérez, B.R. y Gutiérrez, M.A.</i>	45
OBTENCIÓN DE INTERVALOS DE VEROSIMILITUD-CONFIANZA PARA LOS PARÁMETROS DE UNA MEZCLA DE DISTRIBUCIONES WEIBULL	
<i>Díaz-Francés, E. y Villa Diharce, E.</i>	50
VARIABILIDAD NO CONSTANTE EN MODELOS DE REGRESIÓN CON DATOS CENSURADOS	
<i>Domínguez Domínguez, J.</i>	55
ESTIMACIÓN DE LA TASA DE MORTALIDAD INFANTIL EN MÉXICO POR ENTIDAD FEDERATIVA Y MUNICIPIO, CON INFORMACIÓN CENSAL DE 1990	
<i>García, V., Muñiz, O. y Rodríguez, C.A.</i>	60
ALGUNOS MODELOS PARA MEDICIONES REPETIDAS DISCRETAS EN ESTUDIOS LONGITUDINALES	
<i>Gracia-Medrano Valdelamar, L.</i>	65
MÉTODOS MULTIVARIADOS EN LA TAXONOMÍA DEL GRUPO OOCARPA	
<i>Gutiérrez González, P. y Pérez de la Rosa, J.A.</i>	69
CARACTERIZACIÓN DE LA NUEVA CARTA DE CONTROL \hat{p} -V PARA ATRIBUTOS	
<i>Gutiérrez Pulido, H., Camacho Castillo, O. y Hernández Carmona, N.</i>	74
EMPLEO DE TÉCNICAS COMPUTACIONALES MODERNAS PARA IDENTIFICAR DISTRIBUCIONES POSTERIORES	
<i>Hernández-Molinar, R.</i>	80
ANÁLISIS DE UNA SERIE BIOMÉDICA A TRAVÉS DE UN MODELO AUTORREGRESIVO BAYESIANO	
<i>Huerta, G.</i>	85
CADENAS DE MARKOV PARA ANALIZAR EL COMPORTAMIENTO DE LA FLOTA PESQUERA MEXICANA EN LA BÚSQUEDA DE ATÚN	
<i>Lara-Tejeda, J. y Solana-Sansores, R.</i>	90
INTRODUCCIÓN A LOS MODELOS LINEALES BAYESIANOS	
<i>López García, A. y Nuñez Antonio, G.</i>	94
UN MÉTODO PARA MINERÍA DE DATOS CON TRES VARIABLES	
<i>Matuszewski, A.</i>	99
LA INFLUENCIA DE QUETELET EN MÉXICO: UNA POLÉMICA CON UN OCÉANO DE POR MEDIO	
<i>Mayer, L.</i>	105

ESTIMACIÓN BAYESIANA DE PROPORCIONES CONDICIONALES	
<i>Mendoza, M. y Meza, M.</i>	111
MODELACIÓN DE PERFILES DE CRECIMIENTO USANDO MODELOS LINEALES CON COEFICIENTES ALEATORIOS	
<i>Ojeda, M., Sosa-Landa, H. y Nuñez-Antón, V.</i>	116
MODIFICACIÓN DE UN ANÁLISIS BAYESIANO PARA FACTORIALES NO REPLICADOS	
<i>Olgún, J y Romero, P.</i>	121
SINGULARIDADES EN LA DISTRIBUCIÓN FIDUCIAL	
<i>O'Reilly, F. y Rueda, R.</i>	126
ESTIMACIÓN DE COMPONENTES DE VARIANZA DE UN MODELO ESTADÍSTICO PARTICIONADO	
<i>Padrón, E. y Latournerie, L.</i>	131
CLASIFICACIÓN LINEAL BAYESIANA	
<i>Peñaloza Nyssen, B.</i>	135
SOLUCIÓN AL PROBLEMA DE PROGRAMACIÓN CUADRÁTICA USANDO UN MÉTODO HEURÍSTICO	
<i>Pérez Salvador, B.R., De los Cobos Silva, S. y Gutiérrez Andrade, M.A.</i>	140
COMPARACIÓN DE LA CAPACIDAD INFERENCIAL Y PREDICTIVA ENTRE EL ESTIMADOR DE MÁXIMA	
VEROSIMILITUD Y ESTIMADORES ALTERNATIVOS EN PRESENCIA DE MV-COLINEALIDAD	
<i>Ramírez Valverde, G., Rice, J. y Méndez Ramírez, I.</i>	145
VEROSIMILITUD PARA MEZCLAS VÍA APROXIMACIÓN ESTOCÁSTICA	
<i>Rusell Noriega, M.G.</i>	151
EVALUACIÓN DE LA ROBUSTEZ DEL NIVEL DE SIGNIFICANCIA DE LA PRUEBA DE MANN-WHITNEY USANDO	
SIMULACIÓN MONTE CARLO	
<i>Sotres Ramos, D. y Castillo Márquez, L.E.</i>	156
UN SISTEMA DE ANÁLISIS MULTIVARIADO PARA GRANDES VOLÚMENES DE INFORMACIÓN	
<i>Vences Rivera, J., Ramírez Martínez, I. y Flores Nájera, M.A.</i>	161

Conferencias

Invitadas

Utilización de la Estadística en la Medicina Veterinaria y en la Zootecnia

FELIPE DE J. RUIZ LÓPEZ

y

JOSÉ MORO MÉNDEZ

*Centro Nacional de Investigación en
Fisiología Mejoramiento Animal.
INIFAP. SAGAR*

*Facultad de Medicina Veterinaria
y Zootecnia, U.N.A.M.*

La importancia de la estadística dentro de cualquier disciplina científica radica en que permite el análisis e interpretación de datos con el objeto de obtener inferencias sobre la población para posteriormente usarlas como apoyo en la toma de decisiones. Este contexto también aplica para las ciencias agropecuarias que por diversas razones, particularmente en la Medicina Veterinaria y Zootecnia (MVZ), donde estos procesos estadísticos no han sido difundidos o utilizados adecuadamente.

En la aplicación de los métodos estadísticos en MVZ es común la falta de comunicación entre el Médico Veterinario Zootecnista y el Estadístico, lo que frecuentemente es resultado del desconocimiento de las propiedades del fenómeno que se desea estudiar por parte del Médico Veterinario Zootecnista, lo que a su vez trae como consecuencia que no cuente con los elementos necesarios para comunicarse y facilitar la labor de asesoría del Estadístico. Otra causa probable de este distanciamiento es la poca formalidad en el estudio de la estadística dentro de los planes curriculares de las carreras de MVZ en el país. En la mayoría de los planes de estudio la materia de estadística (o bioestadística en algunos casos) es llevada durante los primeros semestres de la carrera, tiempo en el cual el estudiante no tiene una aplicación práctica de la estadística en su quehacer profesional lo que provoca una falta de interés hacia la materia.

Dentro de la MVZ el área que hace un mayor uso de las herramientas estadísticas es la investigación aplicada, sin embargo, en el desarrollo de las diversas actividades en donde se desenvuelven los veterinarios, la estadística podría encontrar aplicación en la comprensión de literatura científica o en la evaluación de situaciones donde se requiera tomar decisiones con base en el análisis de grandes cantidades de datos (Hancock, 1992).

La falta de comunicación con los estadísticos en al área de investigación aplicada resulta en una falta de planeación experimental que podría generar publicaciones con graves deficiencias metodológicas y conclusiones muchas veces arriesgadas.

Esta situación parece no ser exclusiva del área veterinaria y zootecnia del país, también se han reportado aspectos similares en países desarrollados. En un trabajo que tuvo como objetivo analizar la situación de la estadística dentro de la investigación veterinaria basándose en el análisis de los trabajos publicados en el *Journal of American Veterinary Medical Association* (JAVMA) y el *American Journal of Veterinary Research* (AJVR) en el período de 1982 a 1984 (100

artículos de JAVMA y 535 de AJVR) se encontró una frecuencia de errores estadísticos relativamente elevada, 12 y 8%, respectivamente, ocasionados por la utilización de pruebas inadecuadas. Además, aproximadamente 18% de los trabajos publicados no ofrecían información suficiente para que el lector pudiera evaluarlos (Shott, 1985).

Otro trabajo, al analizar los resultados presentados en la revista Veterinary Parasitology de enero de 1992 a enero de 1993, encontró que sólo el 60% contenían algún método estadístico para el análisis de los datos, siendo el análisis de varianza y la Ji cuadrada, las pruebas paramétrica y no paramétrica más utilizadas, respectivamente. Además se detectó que los principales errores cometidos por los autores fueron analizar muestras dependientes como si fueran independientes, no tomar el número de poblaciones bajo estudio para seleccionar la prueba estadística adecuada, no mencionar el cumplimiento de los supuestos de cada prueba y hacer uso de estadísticas inadecuadas para analizar variables discretas y continuas (Sánchez, 1996).

Con el fin de estudiar los principales usos que se le da a la estadística en el área de medicina veterinaria y en la zootecnia en México se analizaron trabajos publicados en 3 de las principales revistas científicas del área en México (Agrociencia, Técnica Pecuaria en México y Veterinaria México) durante los años de 1991 a 1996 y se identificaron los principales errores en la aplicación de métodos estadísticos que resultan en la utilización de técnicas inadecuadas. La clasificación de los trabajos se realizó de acuerdo a los siguientes criterios:

- a) trabajos que tuvieran o no análisis estadístico descrito en la metodología.
- b) tipo de análisis estadístico realizado.
- c) principales errores u omisiones en la aplicación de las técnicas estadísticas que se describieron.

En relación a la presencia o no de análisis estadístico, se encontró que de 113 trabajos estudiados el 20.35% (23 trabajos) no presentaron análisis estadístico, aunque cabe señalar que para efectos de esta revisión, no se consideró como análisis la presentación de estadísticas descriptivas.

Al revisar dos de las revistas para identificar los trabajos que presentan análisis estadístico, se encontró que son las ciencias veterinarias las que mas incurren en falta de análisis (cuadro 1). En el cuadro 2 se muestra la clasificación de las técnicas estadísticas que reportaron utilizar los trabajos con análisis estadístico. El número de técnicas estadísticas reportadas en el cuadro 2 no coincide con el número de trabajos revisados en las tres revistas porque en algunos trabajos se reportó más de un tipo de análisis.

Se puede apreciar que el análisis de varianza fue utilizado en prácticamente en la mitad de los análisis, aunque en más de una ocasión no se cumplió con el supuesto de normalidad de la variable dependiente, como puede ser el caso de los trabajos donde se analizan los conteos de huevos o larvas de parásitos. Además en algunos trabajos se analizaron de manera independiente

variables que podrían estar correlacionadas, como es el caso de la aparición de lesiones en los mismos individuos durante cierto período. También puede apreciarse que la utilización de las técnicas de regresión está poco difundida y lo mismo ocurre con las técnicas no paramétricas, aunque se reconozca que son de gran utilidad en las áreas de medicina veterinaria y zootecnia.

Cuadro 1. *Clasificación de los trabajos de acuerdo a la especialidad*

Especialidad	Con análisis	Sin análisis
Mejoramiento animal	5	0
Nutrición	8	0
Reproducción	15	0
Socioeconomía	1	1
Forrajes y pastizales	1	0
Acuacultura	1	0
Farmacología (incluye Toxicología)	6	0
Inmunología	4	3
Microbiología	3	6
Epidemiología	2	0
Parasitología	5	0
Patología	1	2
Clinica de pequeñas especies	0	1
Anatomía	0	2
TOTAL	52	15

Cuadro 2. *Clasificación de las técnicas estadísticas utilizadas en los trabajo revisados.*

REVISTA\Técnicas	Av	X2	RE	Np	Ts	M M	Mu	Total
Veterinaria México	20	8	4	3	7	1	1	44
Técnica Pecuaria	10	3	1	0	1	0	0	15
Agrociencia	25	3	8	1	1	1	0	39
Total	55	14	13	4	9	2	1	98

Av= Análisis de varianza; X2= Prueba de Ji cuadrada; RE=Regresión lineal; Np= otras pruebas no paramétricas (Kruskal-Wallis, Fisher); Ts= T de student; MM=Modelos mixtos; Mu= Técnicas de muestreo.

En lo que al análisis de experimentos diseñados respecta, se apreció que son pocos los experimentos que no caían dentro de la familia de diseños balanceados y completos. Por ejemplo, se sigue utilizando la aproximación de parcelas divididas para el análisis de medidas repetidas, con la consecuente suposición de correlaciones iguales a través del tiempo. Adicionalmente se encontraron trabajos que eran susceptibles de análisis estadístico pero que no fueron analizados.

Los principales errores de análisis que se encontraron en los trabajos revisados fueron los siguientes: incumplimiento de los supuestos básicos del análisis de varianza, análisis de observaciones no independientes como si lo fueran, distribución no normal de las variables de respuesta, falta de información en las celdas en Ji cuadrada, exclusión de información de unidades experimentales sin notificar causas, utilización de estadísticas descriptivas para hacer comparaciones entre poblaciones y la falta de descripción general del modelo estadístico o la falta de mención de los principales factores y los tratamientos a probar, que afectan la variable estudiada. El último tipo de errores podría ser indicador del desconocimiento del fenómeno que se desea estudiar, ya que no se incluyen todas las posibles fuentes de confusión en el modelo.

Si se estrechara la relación entre el Estadístico y el Médico Veterinario Zootecnista podrían optimizarse los recursos para la investigación a través de la utilización de diseños de experimentos desbalanceados planeados, con el consecuente ahorro en unidades experimentales. Además, la participación de estadísticos en el desarrollo de trabajos de investigación en medicina Veterinaria y Zootecnia es importante dada la existencia frecuente de resultados desbalanceados (aunque el diseño inicial fuera balanceado) tanto en laboratorios como estaciones de campo, donde por diversas razones se pierden unidades experimentales durante en desarrollo de los experimentos.

Con esta colaboración también podrían aplicarse técnicas multivariadas en el análisis de experimentos con mediciones repetidas. Podrían utilizarse técnicas de análisis de supervivencia en el análisis de variables susceptibles. Son muchas las variables de interés que se pueden modelar utilizando este tipo de técnicas y la utilización de modelos de regresión de este tipo permitirían analizar más eficientemente a este tipo de información.

Otro factor que ha modificado de manera importante las técnicas utilizadas en la actualidad es la existencia de paquetes estadísticos, algunos específicos para ciertas especialidades. Sin embargo, existen limitantes de tipo educativo, que se han señalado para el caso del sistema norteamericano de educación veterinaria por Shott (1985) que en ocasiones generan una inadecuada utilización de los mismos. Suponiendo que esta situación también ocurra en México, puede decirse que para los veterinarios zootecnistas que no se dediquen principalmente a la investigación aplicada y se desempeñen en otras actividades, como es la práctica de campo, el conocimiento de conceptos básicos de estadística podrían servirles de apoyo en la elaboración, comprensión y crítica de literatura científica, que sin duda repercutirá en su actualización y mejoramiento de las actividades que desempeñan.

REFERENCIAS

- Shott, S. (1985). Statistics in veterinary research. *Journal of American Veterinary Medical Association*, **187**, 138-141.
Hancock, D. (1992). Critical reading of the scientific literature. *The Bovine Proceedings*, **24**, 29-38.
Sánchez, G. M.G. (1996). Tesis de Licenciatura, Facultad de Medicina Veterinaria y Zootecnia, UNAM.

Inference for Almost Periodically Correlated Processes

JACEK LESKOW

University of California

1. INTRODUCTION

Almost periodically correlated (APC) processes provide an interesting and challenging tool for analysis and statistical inference for time series and stochastic processes. Recent advances in the analysis of autocovariance estimation using APC processes and their applications to statistical inference will be presented. We first survey large-sample results for estimators of covariance for APC processes. Next, we show those results influence statistical inference for nonstationary processes. In particular, the problem of testing stationarity of time series is solved using results established for APC processes. Finally, the large-sample behaviour of the introduced estimators is studied via simulation.

The usefulness of stochastic models that do not require stationarity can be explained by the following examples.

(1) *Signal processing.*

Modulated AM signal can be described as

$$X(t) = P(t) \cdot Z(t),$$

where $Z(t)$ is a wide-sense stationary process and $P(t)$ is a deterministic, periodic function.

Composed signal can be described as

$$X(t) = \sum_{i=1}^n Y_i(t) \sin(2\pi f_i t + \phi_i),$$

where Y_i are independent and stationary, f_i are frequencies and ϕ_i are phases.

Both models are quite popular in signal processing and both require nonstationarity.

(2) *Seismic waves.*

Modeling seismic waves has to take into account that the underlying process cannot be stationary simply because seismic waves rarely have constant mean and variance. When we analyze such waves at some location, the intensity of ground movements intensifies after the arrival of the first wave. Then, it diminishes after reaching a peak. Usually, there is only one realization available at a given location and for such data there is a need to find reasonable model to predict future movements. Grigoriou et al. (1988) have analyzed the 1985 earthquake in Mexico City using nonstationary processes.

(3). *Financial time series*

The appropriate model for volatility is a central theme of many studies in finance. Moreover, recent publications of Pagan and Schwert (1990a,b) show that the assumption of stationarity of financial time series is violated - especially during rapid movements of the market. The popular models used so far - ARMA and GARCH - assume stationarity and there is a need to correct that flaw. We hope that APC process can provide improvement in that situation as well.

Definition 1.1. A shifted autocovariance function $B(t, \tau)$ of the process $\{X(t); t \in \mathbb{R}\}$ will be defined as:

$$B(t, \tau) = \text{Cov}(X(t + \tau), X(t)).$$

Definition 1.2. The process $\{X(t); t \in \mathbb{R}\}$ is called **periodically correlated** (PC) if the function $B(t, \tau)$ is periodic in t .

The process $\{X(t); t \in \mathbb{R}\}$ is called **almost periodically correlated** (APC) if the function $B(t, \tau)$ can be represented as a sum of trigonometric polynomials, i.e.

$$B(t, \tau) = \sum_{\lambda \in \Lambda} a(\lambda, \tau) \exp(i\lambda t), \quad (1)$$

where Λ is a countable set of frequencies λ corresponding to $\{X(t); t \in \mathbb{R}\}$ (see Hurd, 1991; Dehay and Leskow, 1996b).

Basic properties of APC processes

(1) The coefficient

$$a(\lambda, \tau) = \lim_{A \rightarrow \infty} \frac{1}{A} \int_0^A B(t, \tau) e^{-i\lambda t} dt$$

exists for each λ and τ if X is APC.

(2) Under mild regularity properties on $B(t, \tau)$ the set

$$\Lambda = \{ \lambda \in \mathbb{R}; a(\lambda, \tau) \neq 0 \text{ for some } \tau \}$$

is countable.

(3) $\{X(t); t \in \mathbb{R}\}$ is stationary iff $a(\lambda, \tau) \neq 0$ only for $\lambda = 0$, i.e. when $\Lambda = \{0\}$.

(4) $\{X(t); t \in \mathbb{R}\}$ is **periodically correlated** if $\Lambda = \{(2\pi k)/T; k \in \mathbb{Z}\}$ and T is the period of $B(t, \tau)$.

The major question that will be addressed in this paper is: how to use the theory of PC and APC processes to asses the stationarity of the process $\{X(t); t \in \mathbb{R}\}$. The key

technique in the will be based on identifying the true frequencies λ and estimating the Fourier coefficients $a(\lambda, \tau)$ of the shifted covariance function $B(t, \tau)$.

2. INFERENCES FOR NONSTATIONARY PROCESSES

In this section we will present major results on large sample estimation of $a(\lambda, \tau)$ for processes $\{X(t); t \in \mathbb{R}\}$ whose shifted covariance function $B(t, \tau)$ can be approximated via Fourier series, or, in other words, for APC processes.

Throughout the study we assume that we have available observations corresponding to one realization of the process $\{X(t); t \in \mathbb{R}\}$ on the increasing time interval $[0, A]$ (see examples in Section 1). Moreover, the covariance $B(t, \tau)$ of $\{X(t); t \in \mathbb{R}\}$ will be assumed to decay in τ with some rate $\varphi(\tau)$ and the process itself is assumed to fulfill additional regularity assumptions.

We can formally express the above conditions and the regularity assumptions in the following form:

(A1) The process $\{X(t); t \in \mathbb{R}\}$ is uniformly mixing with $\varphi(t) = O(t^{-3})$, as $t \rightarrow \infty$ and

$$\exists \delta > 0, \exists c > 0, \forall t \in \mathbb{R}, E(|X(t)|^{4(1+\delta)}) < c$$

For the definition of uniform mixing, see Billingsley (1968).

(A2) The set $\Lambda = \{\lambda; a(\lambda, \tau) \neq 0 \text{ for some } \tau\}$ has the property:

$$\sum_{\lambda \in \Lambda_X \setminus \{0\}} \frac{1}{\lambda^2} < \infty$$

(A3) The process $\{X(t); t \in \mathbb{R}\}$ has the Holder continuity property in L^4 , i.e. there exists a constant c such that for all t_1 and $t_2 \in \mathbb{R}$

$$E(|X(t_1) - X(t_2)|^4) \leq c|t_1 - t_2|^4$$

(A4) Set $Y(t, \tau_1, \tau_2, \tau_3) = X(t)X(t + \tau_1)X(t + \tau_2)X(t + \tau_3)$. Assume that the function $E(Y(t, \tau_1, \tau_2, \tau_3))$ is uniformly almost periodic in t uniformly with respect to τ_1, τ_2, τ_3 varying in \mathbb{R} .

The estimator of $a(\lambda, \tau)$ is defined as follows:

$$\hat{a}_A(\lambda, \tau) = \begin{cases} \frac{1}{A} \int_0^{A-\tau} X(t+\tau) X(t) e^{-i\lambda t} dt, & \text{if } A \geq \tau \geq 0, \\ \frac{1}{A} \int_{-\tau}^A X(t+\tau) X(t) e^{-i\lambda t} dt, & \text{if } -A \leq \tau \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

It is known (see Hurd and Leskow, 1992; and Dehay and Leskow, 1996b) that the estimator $\hat{a}_A(\lambda, \tau)$ is consistent and asymptotically normal under the conditions (A1) to (A4). We can express it simply in the following form:

Theorem 2.1. The estimator $\hat{a}_A(\lambda, \tau)$ is asymptotically normal, i.e.

$$A^{1/2} (\hat{a}_A(\lambda, \tau) - a(\lambda, \tau)) \xrightarrow{d} N_2(0, C(\lambda))$$

The asymptotic variance-covariance matrix $C(\lambda)$ can be estimated via an estimator \hat{C}_A defined and studied in Dehay and Leskow (1996a).

3. APPLICATIONS

3.1 Time series

We can use known results of consistency and asymptotically normality for the estimator $\hat{a}_A(\lambda, \tau)$ to test the time series data for stationarity. Before formal test are introduced, it is worth noting that many data modeled with time series exhibit the covariance structure similar to that introduced in Definition 1.2, for example: earthquake data (Grigoriou et al., 1988), climate data (Coe, 1983), ozone data (Bloomfield et al., 1994). This remark makes us put the problem of testing stationarity within a large framework of nonstationary, APC processes.

We start the technical consideration by requiring that the analyzed time series $\{X(t); t \in \mathbf{Z}\}$, where \mathbf{Z} denotes the set of integers, does not have a covariance beyond the lag M . In applications, this means that the data points separated by more than M time points are considered independent. Moreover, this also means that $a(\lambda, \tau) = 0$ for $|\tau| > M$.

Using the basic property (3) from Section 1 we see that rejecting stationarity of time series $\{X(t); t \in \mathbf{Z}\}$ is equivalent to detecting a frequency $\lambda \in \Lambda$ such that $\lambda \neq 0$. Therefore, consider the following testing problem:

$$\begin{aligned} H_0: \quad & (a(\lambda, 0), \dots, a(\lambda, M-1)) = 0 \\ & \text{versus} \\ H_1: \quad & (a(\lambda, 0), \dots, a(\lambda, M-1)) \neq 0 \end{aligned} \tag{2}$$

If there is a frequency λ such that H_0 can be rejected then the studied time series $\{X(t); t \in \mathbb{Z}\}$ is **not** covariance stationary.

We will use the following statistic:

$$\hat{U}_A(\lambda) = A(\hat{a}_A(\lambda, \underline{\tau})) \left[\hat{C}_A(\lambda, \underline{\tau}) \right]^{-1} (\hat{a}_A(\lambda, \underline{\tau}))^\top, \quad (3)$$

where $\underline{\tau} = (0, \dots, M-1)$, $\hat{a}_A(\lambda, \underline{\tau}) = (\hat{a}_A(\lambda, 0), \dots, \hat{a}_A(\lambda, M-1))$ and the $2M \times 2M$ dimensional covariance matrix estimator $\hat{C}_A(\lambda, \underline{\tau})$ is defined in Dehay and Leskow (1996a).

We have the following

Theorem 3.1. Under H_0 the test statistic $\hat{U}_A(\lambda)$ has asymptotic χ_{2M}^2 distribution.

Rejection of H_0 for a certain $\lambda \neq 0$ is equivalent to detecting nonstationarity in the time series $\{X(t); t \in \mathbb{Z}\}$. Rejection of H_0 in points $\lambda = (2\pi k) / T, k = 1, \dots, T$ suggests that $\{X(t); t \in \mathbb{Z}\}$ is periodically correlated.

3.2 Testing homoscedasticity

Consider the usual one-way MANOVA model

$$\mathbf{Y}_{ij} = \mathbf{m} + \mathbf{b}_i + \mathbf{e}_{ij} \quad (4)$$

where \mathbf{Y}_{ij} are independent T variate vectors with cdf F_i .

We would like to test whether covariance matrices Σ_i of error vectors \mathbf{e}_{ij} are the same for different factor levels i . Therefore, testing homoscedasticity in model (4) is now equivalent to testing whether $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k$.

Using a construction presented in Leskow (1995) we can have a time series $\{X(t); t \in \mathbb{Z}\}$ that is periodically correlated (PC) if and only if $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k$. This makes the problem of testing homoscedasticity of the one-way MANOVA model equivalent to problem of testing whether a corresponding time series $\{X(t); t \in \mathbb{Z}\}$ is periodically correlated.

4. SIMULATIONS

We will show simulations of PC processes and behaviour of the estimator $\hat{a}_A(\lambda, \tau)$. Let us first briefly describe a simple procedure allowing to generate realizations of PC (and also APC) processes.

Algorithm

Step 1. Generate $U(i)$ - i.i.d. uniform on $[-0.5, 0.5]$. Then put

$$Z(t) = \sum_{i=t-p}^{t+p} U(i)$$

We easily see that $Z(t)$ is $(2p+1)$ -dependent and stationary.

Step 2. Put

$$X(t) = f(t) \cdot Z(t)$$

If $f(t)$ is almost periodic then $X(t)$ is APC.

To study the properties of the estimator $\hat{a}_A(\lambda, \tau)$ we will use the above simulation of $\{X(t); t \in \mathbb{R}\}$. Then we graphically compare

$$\hat{B}_A(t, \tau) = \sum_{\lambda \in \Lambda} \hat{a}_A(\lambda, \tau) \exp(i\lambda t)$$

with the theoretical covariance $B(t, \tau)$ that can be calculated for X such as in Step 2 above. In all simulations we took f to be sinusoid with the period $T=160$, the sample size $A=5000$ and $p=2$ (see Step 1).

The pictures below represent results from the simulations.

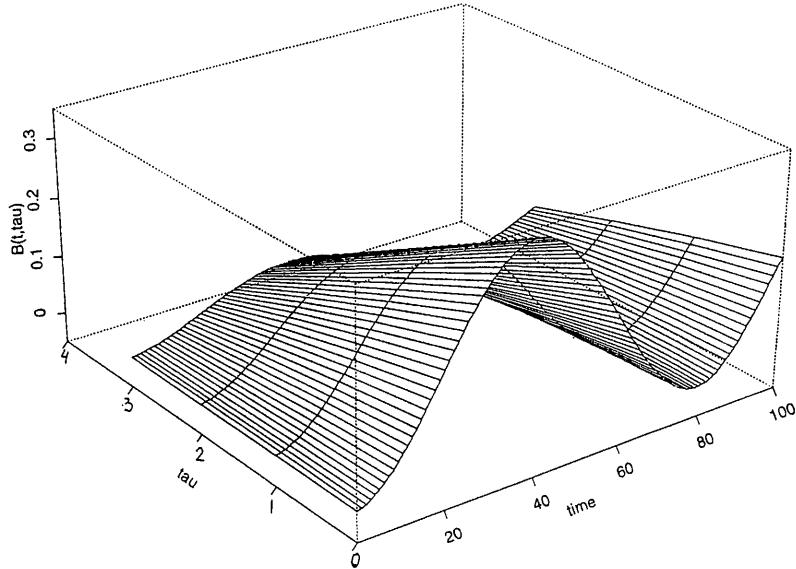


Figure 1. Theoretical shifted covariance $B(t, \tau)$ for the process $X(t)$ when $f(t) = \sin(2\pi t / T)$

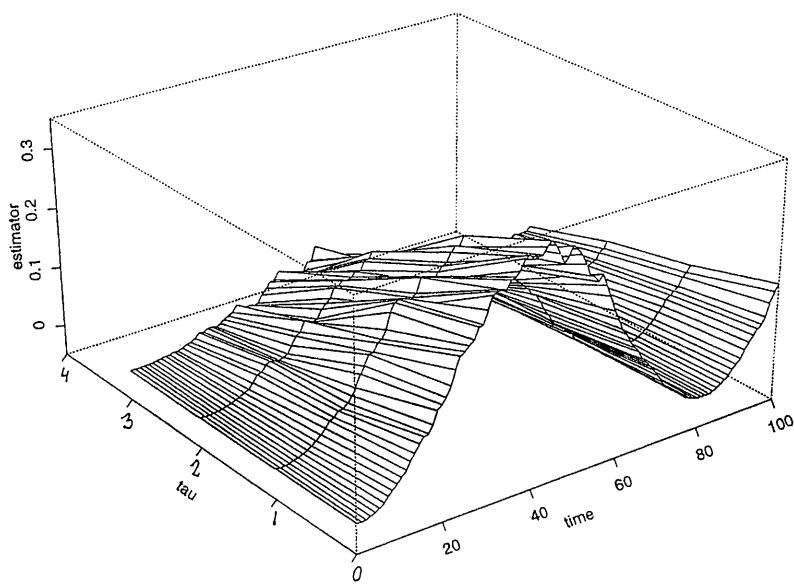


Figure 2. Estimator $\hat{B}_A(t, \tau) = \sum_{\lambda \in \Lambda} \hat{a}_A(\lambda, \tau) \exp(i\lambda t)$

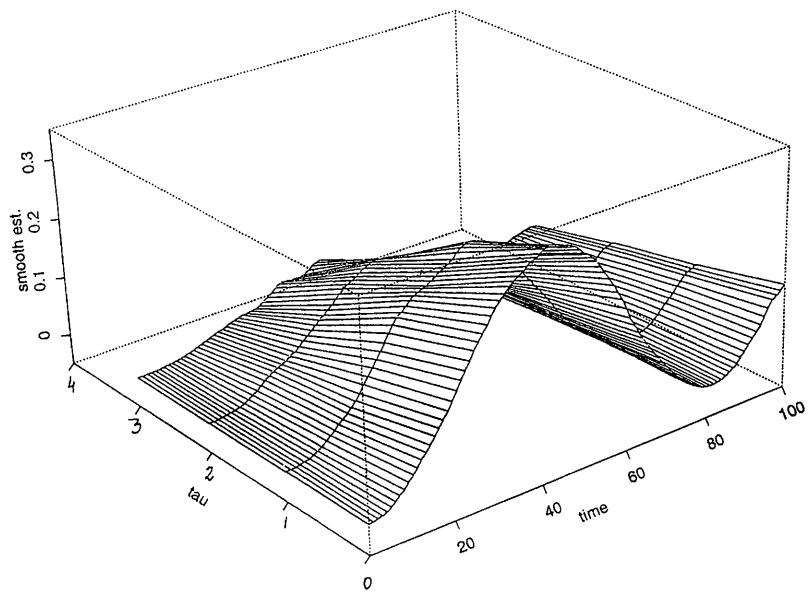


Figure 3. Smoothed version of $\hat{B}_A(t, \tau)$

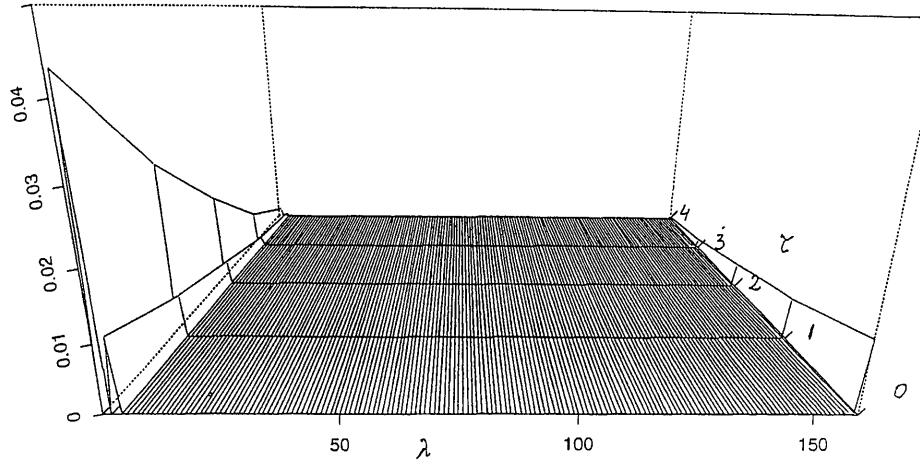


Figure 4. Estimator $\hat{a}_A(\lambda, \tau)$ as a detector of λ . Note the significant peak at the true frequency $2\pi/T$, where $T=160$.

REFERENCES

- Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
- Bloomfield, P., Hurd, H.L. and Lund, R. (1994). *Periodic correlation in Stratospheric Ozone Data*. Journal of Time Series Analysis, **15**, 127-150.
- Coe, R.D. (1983). *Useful models for temperature data*. II International Meeting on Statistical Climatology, Lisboa, Portugal.
- Dehay, D. and Leskow, J. (1996a). *Testing stationarity for stock market data*. Economics Letters, **50**, 205-212.
- Dehay, D. and Leskow, J. (1996b). *Functional limit theory for covariance estimation*, Journal of Applied Probability, **33**, 1077-1092.
- Grigoriou, M., Ruiz, S.E. and Rosenblueth, E. (1988). *The Mexico Earthquake of September 19, 1985 - Nonstationary Model of Seismic Ground Acceleration*. Earthquake Spectra, **4**, 511-568.
- Hurd, H.L. (1991). *Correlation theory for the almost periodically correlated processes with continuous time parameter*. J.Multivariate Anal., **37**, 24-45.
- Hurd, H.L. and Leskow, J. (1992). *Strongly consistent and asymptotically normal estimation of the covariance for almost periodically correlated processes*. Statistics and Decisions, **10**, 201-225.
- Leskow, J. (1995). *Analysis of time series stationarity with applications*. Proceedings of the First Conference on the Applied Statistics. Rieder University, New Jersey, USA.
- Pagan, A.R. and Schwert, A.W. (1990a). *Testing for covariance stationarity in stock market data*. Economics Letters, 165-170.
- Pagan, A.R. and Schwert, A.W. (1990b). *Alternative models for conditional stock volatility*. Journal of Econometrics, **45**, 267-290.

Fisher and the Theory of Maximum Likelihood

D. A. SPROTT

C.I.M.A.T. & University of Waterloo

1. INTRODUCTION

Science is here interpreted to be the study of *repeatable* phenomena. Its purpose is to predict nature. No theory or conclusive inference can be based on a single isolated data set. "In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure", Fisher (1966, p.14).

The data thus are assumed to come from a hypothetical population generated by repetitions of the phenomena. They come in the form of relatively small, but repeatable, samples with probability functions $f(x_i; \theta_i)$, where $\theta_i = (\delta_i, \beta_i)$, reflecting the repeatability of the phenomenon. The parameter of interest is δ_i , assumed to be a scalar. The remaining parameter β_i can be a vector. The repeatable part of the phenomenon is supposed to be embodied in the δ_i 's. Repeatability requires homogeneity of the δ_i 's, $\delta_i = \delta$. The β_i 's represent the non-repeatable or uncontrollable part of the repetitions; they are not assumed to be equal.

The sample information should therefore be divisible into different parts, each part addressing a different issue: the adequacy of the model f , homogeneity $\delta_i = \delta$; the combination of observations to measure the accumulated evidence about δ .

2. DIVISION OF THE SAMPLE INFORMATION

A statistic $T = T(X)$ divides the sample information into two parts

$$f(X; \theta) = g(T; \theta)h(X; \theta|T).$$

An arbitrary division such as this serves no useful purpose and so is not of much interest. The division must have some purpose in mind. It must separate the information in some relevant way. Two such relevant ways are the minimal sufficient and the maximal ancillary divisions. The structure of the minimal sufficient division is

$$\begin{array}{lll} f(X; \theta) & = & g(T; \theta) \quad h(X|T). \\ \text{total} & & \text{minimal} \quad \text{maximal} \end{array} \quad (1)$$

The first factor contains all of the parametric information and as little other information as possible. It yields quantitative inferences about θ . The second factor contains as much of the non-parametric information as possible. It yields shape information useful in testing assumptions about the model f independently of the numerical values of θ . Since h is a logical consequence of f , $f \Rightarrow h$, if the data cast suspicion on h , they equally cast suspicion on f .

The maximal ancillary division is the complementary factoring, where the roles of g and h are interchanged

$$\begin{array}{ccc} f(X; \theta) & = & g(A) \\ \text{total} & & \text{maximal} \\ & & h(X; \theta | A) \\ & & \text{minimal} \end{array}$$

The likelihood function of θ , defined to be proportional to the probability of the observed sample $X = x$ is, from (1),

$$L(\theta; x) \propto f(x; \theta) \propto g(t; \theta).$$

It is completely determined by the minimal sufficient statistic T , and so is itself a minimal sufficient statistic. The likelihood function therefore contains most concisely all of the parametric information. Thus parametric inferences should be based on the observed likelihood function. This implies that estimating intervals should be likelihood intervals.

The above interpretation of the roles of sufficient and ancillary statistics implies that marginal and conditional inferences go hand in hand. The latter is a necessary concomitant of the former. They combine to form all of the information. In the light of this the constant criticism and misunderstanding of “conditional inference” is curious. The only question arising from the above approach is how adequately the sufficient and ancillary statistics succeed in dividing the information. The conditional factors may not contain much information.

3. LIKELIHOOD INTERVALS

The likelihood function ranks the plausibility of individual values of θ by how probable they make the observed value x . Likelihood intervals specify ranges of most plausible values of θ . A level c likelihood interval (or in some cases a union of nonoverlapping intervals) is given by

$$R(\theta; x) = L(\theta; x) / L(\hat{\theta}; x) \geq c, \quad 0 \leq c \leq 1,$$

where $\hat{\theta} = \hat{\theta}(x)$ is the maximum likelihood estimate and R is the relative likelihood function, $0 \leq R \leq 1$. The maximum likelihood estimate $\hat{\theta}$ is contained within all of the likelihood intervals, and therefore serves to specify their location. The complete nested set of likelihood intervals converging to the maximum likelihood estimate $\hat{\theta}$ reproduces the likelihood function.

A single interval is not very informative and so does not suffice. It merely states that values of θ outside the interval have relative plausibilities less than c , while values inside have plausibilities greater than c . But it gives no indication of the behaviour of plausibility within the interval. To do this the interval should at least be supplemented by $\hat{\theta}$ to give some indication of the statistical center of the intervals and hence an idea of the skewness of the likelihood function. It gives some indication of how the plausibility of θ changes within

the interval, whether, for example, values to the right of $\hat{\theta}$ are more plausible than values to the left. Preferably, however, a nested set of likelihood intervals should be given along with $\hat{\theta}$.

Example 3.1 The Poisson dilution series, Fisher (1921), Fisher and Yates (1963, p. 9). Suppose the density of organisms in a given medium is θ per unit volume. To estimate θ the original medium is successively diluted by a dilution factor a to obtain a series of $k+1$ solutions with densities $\theta/a^0, \theta/a, \theta/a^2, \dots, \theta/a^k$. Suppose that a unit volume of the solution with density θ/a^i is injected into each of n_i plates containing a nutrient upon which the organisms multiply, and that only the presence or absence of organisms can be detected. The observations are then x_0, x_1, \dots, x_k , where x_i is the number of sterile plates out of the n_i at dilution level i .

Assuming a Poisson distribution of the organisms in the original medium, the probability of a sterile plate at level i is the probability that a given unit volume at dilution level i contains no organisms, which is

$$p_i = \exp(-\theta/a^i), \quad i = 0, 1, \dots, k.$$

The probability of a fertile plate at level i is $1 - p_i$. Assuming independence of the n_i plates, x_i has the binomial distribution (n_i, p_i) . The likelihood of θ is proportional to the probability of the observations

$$L(\theta; x) \propto \Pr(x_0, \dots, x_k; \theta) = \prod_{i=0}^k \binom{n_i}{x_i} p_i^{x_i} (1-p_i)^{n_i-x_i}. \quad (2)$$

Figure 1 shows the relative likelihood function $R(25\theta)$ for the data, $\{n_i\} = 5, k = 9, a = 2, \{x_i\} = \{0, 0, 0, 0, 1, 2, 3, 3, 5, 5\}$, Fisher and Yates (1963 p. 9). The factor 25 is to convert number of organisms per cc. to number per gm.

The 5% likelihood interval is (358, 1512). By itself this gives little information about the plausibility of $\hat{\theta}$, only that values outside this interval have relative likelihoods less than 5% and values inside greater than 5%. But it gives no hint of behaviour of plausibility within the interval. Merely supplementing this interval by $25\hat{\theta} = 766$ gives considerably more information. This represents the center of likelihood. Its deviation to the left of the geometrical center exhibits the skewness of the likelihood function to the right. This means that $\hat{\theta}$ is more likely to be larger than 766 than smaller. Larger values are more likely than smaller values. Inferences must take this into account or there may be a tendency to underestimate the number of organisms. It is preferable to supplement this still further by giving the 5%, 15% and 25% likelihood intervals, or something similar, as in Figure 1.

4. LIKELIHOOD-CONFIDENCE INTERVALS

The likelihood function is a point function. Thus likelihood intervals are statements of relative plausibility of individual values of θ within the interval. They are not statements of plausibility about the interval itself. It is desirable, if possible, to supplement the likelihood intervals with the probabilities that they actually include the true value of θ . This is a statement of uncertainty about the interval itself.

Relative Likelihood, Dilution Series Data

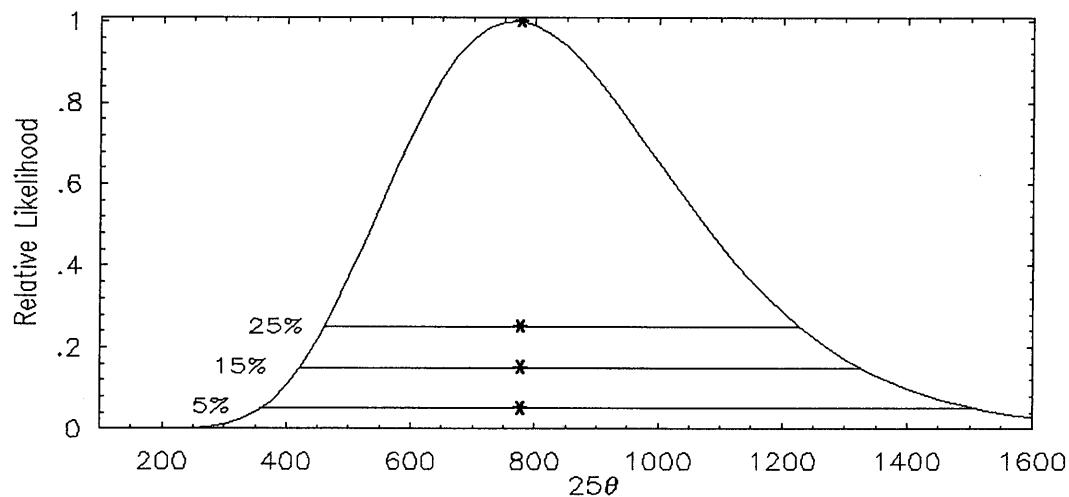


Figure 1. Relative likelihood, $R(25\theta)$, dilution series data

To do this requires associating, if possible, probabilities with the intervals. This requires a pivotal quantity. A pivotal quantity is a function $v(X; \theta)$ that has a completely specified density $f(v)$. For $v(X; \theta)$ to be a pivotal quantity requires that for any $a < b$, the probability $\pi = \Pr(a \leq v \leq b)$ must be numerically calculable. Assuming v is 1-1 in θ for all x , any interval in v is mathematically equivalent to a corresponding interval in θ . Attaching π to this corresponding interval in θ produces a confidence interval having coverage frequency π .

If the relative likelihood can be expressed as a function of a pivotal v , $R(\theta; x) \equiv R(v)$, R is itself a pivotal. Then the nested set of likelihood intervals is also a nested set of confidence intervals. Such nested sets can be called likelihood-confidence sets.

An example is the gamma likelihood arising from n independent exponential variates x_i with mean θ , where $v = t/\theta = \sum x_i/\theta$. The relative likelihood,

$$R(\theta; t|n) = (t/n\theta)^n \exp(n - t/\theta) = (v/n)^n \exp(n - v) = R(v|n),$$

is a pivotal quantity whose distribution can be obtained from the gamma (n) distribution of v

Thus the first requirement of estimating intervals is that they should be a nested set of likelihood intervals. These can then be supplemented, if possible, by coverage frequencies to give a nested set of likelihood-confidence intervals.

5. MAXIMUM LIKELIHOOD ESTIMATION

Example 5.1 Poisson dilution series of Example 3.1. It can be verified by direct substitution into (2) that the parameter $\delta = \log \theta$ has an approximate normal likelihood given by

$$R_N(\delta; x) = \exp\left(-\frac{1}{2} u_\delta^2\right), \quad u_\delta = (\hat{\delta} - \delta) \sqrt{\hat{I}(\hat{\delta}; x)} \quad (3)$$

where

$$\hat{I}(\hat{\delta}; x) = -\left[\partial^2 \log R(\delta; x) / \partial \delta^2\right]_{\delta=\hat{\delta}}$$

is the *observed* Fisher information calculated at $\hat{\delta}$. This yields an approximate complete set of nested likelihood-confidence intervals analytically

$$\delta = \hat{\delta} \pm \frac{u}{\sqrt{\hat{I}(\hat{\delta}; x)}} \Leftrightarrow \theta = \hat{\theta} \exp\left[\pm \frac{u}{\hat{\theta} \sqrt{\hat{I}(\hat{\theta}; x)}}\right]. \quad (4)$$

A c likelihood interval is given by $u = \sqrt{-2 \log c}$. Taking u as a $N(0,1)$ pivotal gives the corresponding confidence level. The accuracy of these confidence levels depend on the extent to which u is a $N(0,1)$ pivotal. This can be checked by simulations if necessary.

For the Fisher and Yates dilution series data, $\hat{\theta} = 30.65$, $I(\hat{\theta}; x) = 0.01225$, giving the complete set of nested approximate likelihood-confidence intervals

$$25\theta = 766 \exp(\pm u / 3.3923), \quad u = \sqrt{-2 \log c}, \quad 0 \leq c \leq 1. \quad (5)$$

These are shown in Figure 2 along with the likelihood function shown in Figure 1. Using the relation $u = \pm \sqrt{-2 \log c}$, the 5%, 15%, and 25% likelihood intervals are approximate 99%, 95%, and 90% confidence intervals. Also shown in Figure 2 is $R_N(25\theta)$. This results in approximate confidence intervals $25\theta = 766 \pm 25u / 0.1107$ based directly on the θ scale instead of on the δ scale. From Figure 2 it can be seen that this ignores the asymmetry in the likelihood and shifts all of the intervals to the left, thus understating the magnitude of θ .

Fisher (1922) gave the complete scientific application of maximum likelihood to the estimation of θ in such a dilution series experiment. It is interesting that he developed this entirely in terms of $\log \theta$, without any comment about the use of the logarithm, implicitly giving (5).

Likelihood-confidence intervals, dilution series data

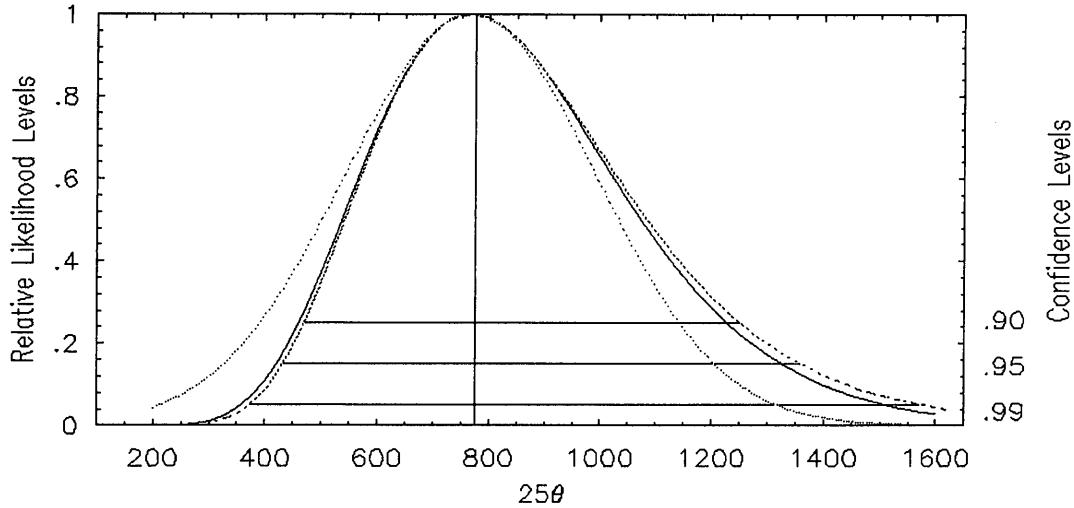


Figure 2: Likelihood-confidence intervals, dilution series data: $R(25\theta)$ -----; approximate likelihood based on $R_N(\delta)$ - - -; approximate normal likelihood $R_N(25\theta)$,

Example 5.2 The difference between two binomial distributions - the 2×2 table. The following data are the results of a double blind randomized trial on the drug ramipril, AIRE Study Group (1993). The purpose was to assess the efficacy of ramipril in enhancing survival after an acute myocardial infarction.

	<i>S</i>	<i>F</i>	Total
Ramipril	834	170	1004
Placebo	760	222	982
Total	1594	392	1986

Let p_1 and p_2 be the probabilities of success on the treatment (ramipril) and the control (placebo), respectively. The difference between ramipril and the control can be measured by the odds ratio $\theta = p_1(1-p_2)/(1-p_1)p_2$, $0 \leq \theta \leq \infty$. The odds of success under ramipril are θ times greater than the corresponding odds under the control.

As in the dilution series case, Example 5.1, to take account of the asymmetry of the relative likelihood $R(\theta)$ it is preferable to apply maximum likelihood to the log odds ratio,

$$\begin{aligned} \delta &= \log \theta = \log[p_1(1-p_2)/(1-p_1)p_2] \\ &= \log[p_1/(1-p_1) - \log[p_2/(1-p_2)]], \quad -\infty \leq \delta \leq \infty. \end{aligned} \tag{6}$$

The log odds ratio measures the difference between these distributions on the logistic scale.

Suppose there are n patients and that r of them are randomized to ramipril. Let x and y be the observed number of successes S and $r - x$, $n - r - y$ the number of failures F , under ramipril and the control, respectively. Assume that x and y are independent binomial (r, p_1) and $(n - r, p_2)$ variates. The maximum likelihood estimate of δ is

$$\hat{\delta} = \log[x(n-r-y)/(r-x)y] = 0.3598.$$

The standard error s , based on the inverse of the *observed* 2×2 Fisher information matrix I , is given by

$$s^2 = I^{\delta\delta} = \frac{1}{x} + \frac{1}{r-x} + \frac{1}{y} + \frac{1}{n-r-y} = 0.0129. \quad (7)$$

The resulting quantity (3) is $u = (\hat{\delta} - \delta)/s$.

Estimation statements take the form

$$\delta = \hat{\delta} \pm su = 0.3598 \pm 0.1136u \Leftrightarrow \theta = \hat{\theta} \exp(\pm su) = 1.4330 \exp(\pm 0.1136u),$$

where u is an approximate $N(0,1)$ pivotal. These are shown in Figure 3 in terms of θ , along with the exact conditional likelihood function $R_c(\theta)$ which is proportional to the conditional probability $\Pr(x=834; \theta|x+y=1594)$. Also shown is the normal approximation $R_N(\theta)$ of $R_c(\theta)$. Again it can be seen that basing intervals on $R_N(\theta)$ results in ignoring the skewness in $R_c(\theta)$ and shifting the intervals to the left, thus understating θ .

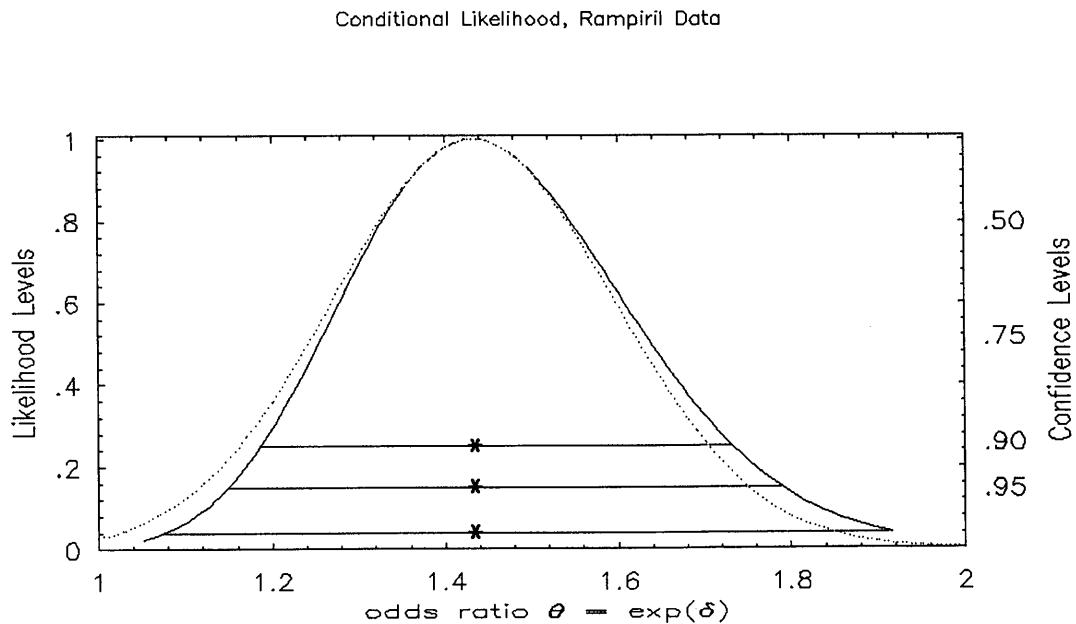


Figure 3: Approximate likelihood-confidence intervals based on $R_N(\delta)$, ramipril data: $R_c(\theta)$ -----; $R_N(\theta)$

The foregoing presents the scientific interpretation of maximum likelihood estimation as a method to obtain approximate likelihood-confidence intervals. This is equivalent to obtaining approximate linear sufficient pivotal quantities, that is, pivots linear in the parameter containing all the sample parametric information.

6. "CORRECTIONS" TO THE MAXIMUM LIKELIHOOD ESTIMATE

Unfortunately, maximum likelihood is usually presented as a method to obtain asymptotically unbiased estimates with minimum variance. This leads to repeated attempts to "correct" $\hat{\theta}$ for bias and to calculate its variance. An example of this is Miyramura (1982). This involved the estimation of the underlying failure rate based on observing the failure times of systems of components connected in series, assuming an exponential failure time at rate $\lambda > 0$ for the individual components. The results for a single component are summarized as follows.

$$\begin{aligned}\tilde{\lambda} &= [1 - (2/\tilde{v})]\hat{\lambda}, \\ \tilde{\sigma}^2 &= \tilde{\lambda} \left[1 - (2/\tilde{v}) + (4\tilde{m}\tilde{\lambda}/\tilde{v}^2) \right] / [\tilde{m}(1 - 4/\tilde{v})], \\ \text{where } \hat{v} &= 2 \left(\sum_{i=1}^n r_i z_i / \tilde{\beta}_i \right)^2 / \left(\sum_{i=1}^n r_i z_i^2 / \tilde{\beta}_i^2 \right), \\ \tilde{m} &= \sum_{i=1}^n r_i z_i / \tilde{\beta}_i, & \tilde{\beta}_i &= (r_i - 1) / t_i.\end{aligned}$$

One of the numerical examples given yielded $\hat{\lambda} = 0.035$, $\tilde{\lambda} = 0.028$, $\tilde{\sigma} = 0.024$ in a sample of size $n = 2$. Viveros (1991) noted that the use of this to produce confidence intervals gives the 95% confidence interval $-0.019 \leq \lambda \leq 0.075$. Values $\lambda < 0$ are of course impossible. Such intervals may be termed "incredible".

Section 5 suggests applying maximum likelihood to the parameter $\delta = \lambda^{1/3}$, the likelihood of which is approximately normal even in samples of size two. The observed information is $I(\hat{\delta}; x) = 9n/\hat{\delta}^2$. The resulting estimating intervals (4) based on the approximate $N(0, 1)$ linear pivotal $u = (\hat{\delta} - \delta)3\sqrt{n}/\hat{\delta}$ take the simple form

$$\delta = \hat{\delta} \left(1 \pm \frac{u}{3\sqrt{n}} \right) \Leftrightarrow \lambda = \hat{\lambda} \left(1 \pm \frac{u}{3\sqrt{n}} \right)^3.$$

The resulting 95% likelihood-confidence interval is $0.005 \leq \lambda \leq 0.109$. Simulations show that the coverage frequency of intervals produced this way are very close to those obtained by assuming $u \sim N(0, 1)$, Viveros (1991). Therefore these are a highly accurate nested set of approximate likelihood-confidence intervals.

This example also illustrates that if simulations are required, the right quantity should be simulated. To set up confidence intervals it is rarely appropriate to simulate the estimate. The estimate by itself rarely determines the confidence intervals. The quantity u should be simulated. This quantity has the form of a Student t pivotal, and so cannot be separated into an estimate and its variance.

A more recent example of this is Mehrabi and Mathews (1995) where a bias correction is applied to the maximum likelihood estimate in the Poisson dilution series model of Example 3.1. For the Fisher and Yates data, the resulting "corrected" estimate is $\tilde{\theta} = 28.666$, with estimated variance 81.1688.

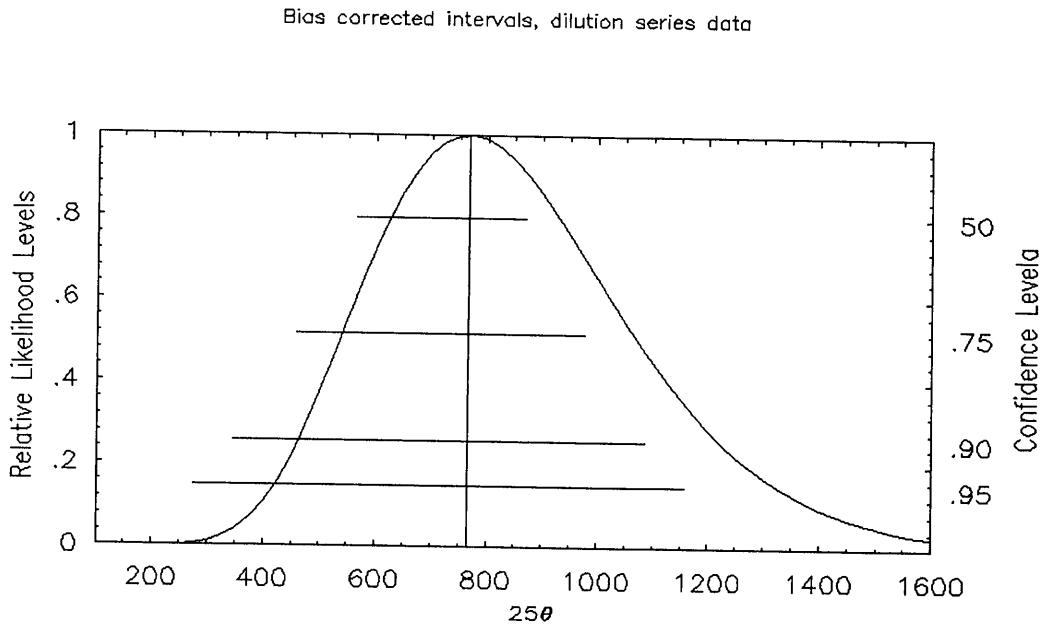


Figure 4: Confidence intervals corrected for bias, dilution series data

The resulting intervals $25\theta = 717 \pm 226u$ are shown in Figure 4 along with the relative likelihood $R(25\theta; x)$ of Example 3.1, Figure 1. The intervals are shifted well to the left of the likelihood function, and so include highly implausible small values of θ and exclude highly plausible large values, again understating the magnitude of θ . Eliminating statistical bias introduces a more important and obvious "scientific" bias. The positive bias of $\hat{\theta}$ is important in forcing attention to values of θ larger than $\hat{\theta}$ reinforcing the message conveyed by the asymmetry of the likelihood. Ignoring these facts results in seriously understating θ , as Figure 4 shows.

REFERENCES

- AIRE Study Group (1993). Effect of ramipril on mortality and morbidity of survivors of acute myocardial infarction with clinical evidence of heart failure. *Lancet*, **342**, 821-828.
 Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, **222**, 309-368.
 Fisher, R.A. and Yates, F. (1963). *Statistical Tables*, 6th edition. Hafner Publishing Company, N. Y.

- Mehrabi, Y. and Mathews, J.N.S. (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics*, **51**, 1543-1549.
- Miyramura, T. (1982). Estimating component failure rates from combined component and systems data: exponentially distributed component lifetimes. *Technometrics*, **24**, 313-318.
- Viveros, R. (1991). Combining series system data to estimate component characteristics. *Technometrics*, **33**, 13-23.

Contribuciones

Libres

Función de Valor Extremo y sus Aplicaciones

ALEJANDRO ALEGRÍA y HUMBERTO SOTO

I.T.A.M.

1. INTRODUCCIÓN

En este trabajo se presenta una aplicación de la distribución de valor extremo en el área deportiva. En primer lugar se dà una breve explicación de cómo surge este tipo de distribución y la forma de la misma. Posteriormente se plantea el problema de saber si un record impuesto en una competencia de 3000 m planos es, o no es, estadísticamente verosímil. Se construyen intervalos que permiten concluir si un nuevo record puede ser considerado poco factible. En las conclusiones se plantean posibles extensiones al modelo utilizado.

2. DISTRIBUCIÓN DE VALOR EXTREMO

El interés por estudiar la distribución de valor extremo surge de la necesidad de modelar fenómenos en donde el interés principal es explicar el comportamiento del valor mínimo o máximo de un conjunto de valores. Las aplicaciones son diversas: resistencia de materiales, inundaciones, sequías, contaminación ambiental, confiabilidad de sistemas, records atléticos, entre otras. Formalmente, la distribución de valor extremo se genera al considerar un conjunto de variables aleatorias independientes X_1, X_2, \dots, X_n , con una distribución común $F(x)$, y para las cuales se desea encontrar la distribución de los valores extremos $M_n = \max\{X_1, X_2, \dots, X_n\}$ y $m_n = \min\{X_1, X_2, \dots, X_n\}$. Se demuestra fácilmente que $G_n(w) = P(m_n \leq w)$ y $H_n(w) = P(M_n \leq w)$ están dadas por

$$G_n(w) = 1 - (1 - F(w))^n \quad , \quad H_n(w) = (F(w))^n.$$

Como en muchos casos las expresiones de $G_n(w)$ y de $H_n(w)$ no son fáciles de manejar, y además en muchas aplicaciones se cuenta con un valor de n grande, la teoría de valores extremos se ha desarrollado usando argumentos asintóticos. Para una elección apropiada de sucesiones $\{a_n\}$, $\{b_n\}$, $\{A_n\}$, $\{B_n\}$, las distribuciones límite de $(m_n - a_n)/b_n$ y de $(M_n - A_n)/B_n$, pertenecen a la llamada familia de Valor Extremo Generalizada (VEG). En esta familia, las funciones de distribución respectivas del mínimo y el máximo, tienen la siguiente forma,

$$G(w) = 1 - \exp \left\{ - \left[1 - \xi \left(\frac{w - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\} \quad , \quad H(w) = \exp \left\{ - \left[1 + \xi \left(\frac{w - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\}$$

en donde, $[y]_+ = \max\{y, 0\}$. Los parámetros de localización y escala son μ y σ respectivamente, mientras que ξ es el parámetro de forma, el cual determina el peso de la cola de la distribución. Sobre el tema de valores extremos algunas referencias son Smith (1992) y Leadbetter *et al.* (1983).

La aplicación que se presenta en este trabajo esta enfocada al estudio de valores mínimos, así que se hará uso de la distribución $G(\cdot)$.

3. APLICACIÓN: RECORDS EN ATLETISMO

En las competencias atléticas la constante mejora en equipo y en técnicas de entrenamiento, así como un mayor conocimiento de la dieta adecuada para cada competencia, ha favorecido que los records se hayan roto con más frecuencia y en forma más notoria. Cuando el margen por el cual se rompe un record es sorprendentemente mayor que en otras ocasiones, es natural especular sobre el posible uso de sustancias prohibidas.

Si en la prueba antidoping un atleta presenta un resultado negativo, la única evidencia que se tiene del posible uso de drogas es la evidencia muestral contenida en los datos de distancia, tiempo, evento y posición final del atleta en cuestión y de otros atletas en competencias similares. Con esta información se podría saber si un nuevo record es considerado dentro del conjunto de valores más probables.

El caso específico que aquí se presenta es el de la competidora china Wang Junxia. Durante el campeonato nacional en Beijing, en septiembre 13 de 1993, esta atleta corrió los 3000 m. planos en un tiempo de 8 m 6.11 seg (486 seg), mejorando el último record en 6.08 seg, habiendo roto el día anterior dicho record por 10.43 seg. Nunca se pudo comprobar el uso de drogas. A partir de los datos no se puede saber si ella usó drogas o si proviene de una población diferente a la de los atletas que usualmente compiten. El modelo a usar en este trabajo pretende incorporar la evidencia pasada para decidir si el record es consistente con la información histórica.

La información que se tiene se muestra en la figura 1, y corresponde a los mejores cinco tiempos en la competencia de 3000 m planos para mujeres, desde 1972 hasta 1992 (Track & Field News, 1972 a 1992). Para 1993 se puede ver el tiempo de la china, y se aprecia como redujo el tiempo en forma sustancial. Se supondrá que el mínimo tiempo para esta prueba es una variable aleatoria X con distribución en la familia de VEG, es decir, la distribución de X es la función $G(\cdot)$ presentada en la sección anterior. En la figura 1 se puede apreciar una tendencia en los tiempos la cual es conveniente incorporar en nuestro modelo. Esto se puede hacer permitiendo que μ varie con el tiempo t , de tal forma que ahora se tiene

$$G_t(x) = 1 - \exp \left\{ - \left[1 - \xi \left(\frac{x - \mu_t}{\sigma} \right) \right]_+^{-1/\xi} \right\}$$

El decaimiento que se muestra en los datos será modelado suponiendo un decamiento exponencial,

$$\mu_t = \alpha - \beta [1 - \exp(-\gamma t)],$$

con $\beta > 0$ y $\gamma > 0$. Tanto σ como ξ se supone que no dependen del tiempo, lo cual no parece contrario a lo se ve en la figura 1.

Sea ahora $x_{p,t}$ el tiempo para el cual $G_t(x_{p,t}) = p$. Es fácil comprobar que

$$x_{p,t} = \mu_t + \sigma [1 - \{-\log(1-p)\}^{-\xi}] / \xi$$

De esta última expresión y de la relación de μ_t con el tiempo se obtiene lo que se llama el **tiempo último** en el se puede correr la prueba, x_{ult} , el cual se define como el valor de $x_{p,t}$ cuando $p = 0$ y $t \rightarrow \infty$. El valor de x_{ult} esta dado por

$$x_{ult} = \begin{cases} \alpha - \beta + \sigma / \xi & , \text{ si } \xi < 0 \\ -\infty & , \text{ si } \xi \geq 0 \end{cases}$$

Como no hay tiempos negativos, $-\infty$ se interpreta como 0.

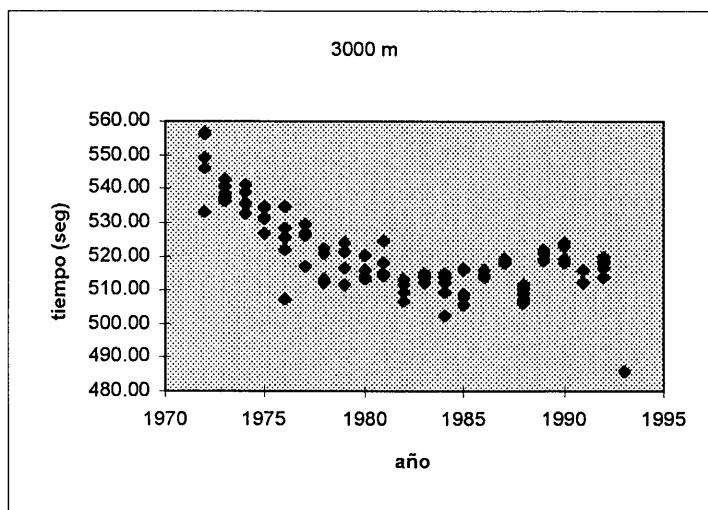


Figura 1

Para la estimación de los parámetros α , β , γ , σ y ξ , se requiere construir la densidad conjunta de las primeras $r = 5$ estadísticas de orden para cada año. Esta densidad conjunta esta dada, de acuerdo a Smith (1986), por

$$(-1)^r \left[\prod_{i=1}^r \frac{d}{dx^{(i)}} \ln(1 - G_t(x^{(i)})) \right] (1 - G_t(x^{(i)}))$$

donde $x_t^{(1)} \leq x_t^{(2)} \leq \dots \leq x_t^{(r)}$ son los r mejores tiempos en al año t . La obtención de esta última densidad supone que r es un valor fijo y que asintóticamente los valores extremos son

independientes. Si además se supone que hay independencia entre los años, los estimadores máximo verosímiles de los parámetros, y las desviaciones estándar respectivas son,

$$\begin{aligned}\hat{\alpha} &= 558 & \hat{\beta} &= 48.7 & \hat{\gamma} &= 0.272 & \hat{\sigma} &= 4.85 & \hat{\xi} &= -0.197 \\ (4) & & (3.8) & & (0.034) & & (0.40) & & (0.068)\end{aligned}$$

Con estos valores se puede estimar la función de regresión $E(X_t)$, como

$$\hat{E}(X_t) = \hat{\mu}_t + \hat{\sigma}[1 - \Gamma(1 - \hat{\xi})]/\hat{\xi}$$

Esta función se muestra en la figura 2

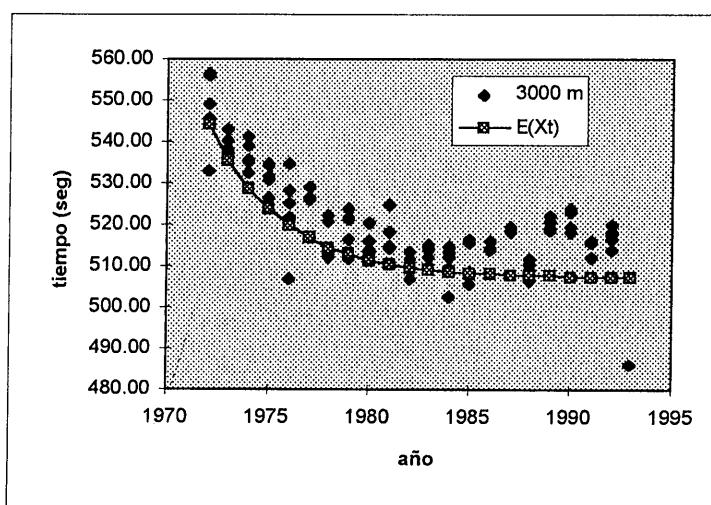


Figura 2

El valor de \hat{X}_{ult} es 484.68, el cual resulta ser menor que el tiempo de la atleta china (486.11 seg). Para valorar el tiempo de Wang Junxia es de más utilidad construir un intervalo de confianza para x_{ult} . Usando resultados asintóticos (profile log-likelihood), el intervalo de 90% de confianza que se obtiene para x_{ult} es (430.1, 493.8). Con respecto a este intervalo conviene mencionar que si se hubieran utilizado solamente los tiempos mínimos de cada año ($r = 1$), el intervalo hubiera sido (277.9, 656.9). Es notorio como disminuye el valor superior del intervalo cuando se utiliza mas información que la dada por el mínimo de cada año. En cualquier caso, es importante observar que el record de Junxia resulta ser consistente con la experiencia previa.

4. CONCLUSIONES

En este trabajo se ha presentado una aplicación de la distribución de valor extremo generalizada en donde el interés era determinar si un valor mínimo puede considerarse un valor verosímil dentro de los posibles mínimos. El tiempo realizado por la competidora china en 1993, resultó ser consistente con la información contenida en los mejores cinco tiempo anuales, de 1972 a 1992. El modelo utilizado se puede extender considerando información

de competencias similares e incluyendo una variable indicadora en el análisis para distinguir entre Olimpiadas y Campeonatos Mundiales de otro tipo de competencias. En este sentido, el trabajo de Robinson y Tawn (1995) incorpora dichos elementos al modelo, y llegan a la misma conclusión: el tiempo de Junxia es consistente. Los supuestos que se han utilizado para la obtención de los estimadores son: independencia entre los tiempos mínimos de cada año, independencia entre los años, variabilidad constante en el tiempo. El modificar los supuestos con las implicaciones que ésto conlleva, es algo que vale la pena estudiar. La utilización de métodos asintóticos en la estimación de los parámetros es algo que se puede criticar, pues no es claro que con los tamaños de muestra que se utilizaron se obtengan estimadores con buena eficiencia. Lo que se puede hacer en este caso es aplicar técnicas bayesianas, las cuales no se ven afectadas, en principio, por una verosimilitud no acotada; sólo se requiere que la verosimilitud sea integrable con respecto a la distribución inicial. La distribución final a la que se llega es complicada analíticamente, pero este problema se puede resolver con procedimientos numéricos (Gibbs Sampler). Se está trabajando con los mismos datos utilizados en el presente modelo pero aplicando procedimientos bayesianos. Los resultados se reportarán en un trabajo futuro.

REFERENCIAS

- Leadbetter, M.R. Lindgren, G. and Rootzen, H. (1983). *Extreme and Related Properties of Random Sequences and Series*. New York: Springer.
- Robinson, M.E. and Tawn, J. (1995). Statistics for Exceptional Athletics Records, *Appl. Statist.*, **44**, 499-511.
- Smith, R.L. (1986). Extreme value theory based on the r largest annual events. *J. Hydrol.*, **86**, 27-43.
- Smith, R.L. (1992). Extreme Value Theory. In *Handbook of Applicable Mathematics*, 7, 437-471. John Wiley.
- Track & Field News. 1972-1992.

El Impacto de las Condiciones Socioeconómicas de la Población en sus Preferencias Electorales

ELOÍSA ARROYO MARTÍNEZ y LOURDES DE LA FUENTE

I.T.A.M.

1. MARCO TEÓRICO

Existen diversos estudios que analizan, tanto de manera teórica como empírica, la existencia de una relación real entre las condiciones económicas y sociales de la gente y sus decisiones de orden político¹. Sin embargo el empleo de técnicas inferenciales de análisis estadístico es un asunto reciente en este tipo de investigación, que hasta no hace mucho utilizaba técnicas de análisis cualitativas o elementos estadísticos descriptivos básicos. Si bien la existencia de dicha relación no es nada nuevo, el presente trabajo pretende demostrarla utilizando modelos lineales generalizados de elección cualitativa. El objetivo principal, es encontrar evidencia empírica sobre el impacto que las condiciones de marginalidad social pudieran llegar a ejercer sobre las decisiones electorales de la gente.

2. ESPECIFICACIÓN DEL MODELO

El modelo considera los cuatro estados con gobierno de oposición en el país: Baja California, Chihuahua, Guanajuato y Jalisco. La justificación de la elección de los estados que configuran al modelo es simplemente metodológica: para aplicar la metodología elegida, es necesario contar con datos que reflejen niveles de competencia “real” entre los diferentes partidos². El modelo emplea información a nivel municipal de las variables que en México se consideran como determinantes de la marginalidad (dichas variables abarcan cuestiones referentes a educación, vivienda e ingresos) como variables independientes o explicativas, por un lado; y variables de carácter político, como variables dependientes, por el otro.

La razón de utilizar modelos de elección cualitativa es simple: son aquellos que mejor se adecuan a situaciones en que la variable respuesta es de tipo dicotómica. En este estudio, la variable dependiente del modelo se codifica como uno si ganó la oposición en el municipio, y como cero en el caso contrario. Las variables independientes del modelo estarán determinadas por variables de carácter socioeconómico que abarcan tres áreas principales: educación; condiciones de la vivienda, y niveles de ingresos. Se utilizan las variables que conforman el índice de marginalidad de CONAPO, pues lo que se busca es analizar si realmente existe una relación entre las condiciones de pobreza y las preferencias electorales que pudiera manipularse con fines electorales.

¹ Véase Molinar Horcasitas, Juan y Rafael Vergara, *Los estudios sobre el elector mexicano: cuatro enfoques de análisis electoral en México*, mimeo, el Colegio de México, 1996.

²En este sentido es importante resaltar que no se está suponiendo que en el resto del país no exista competencia política real, sino que, en muchos casos, ésta no se ve reflejada en las urnas electorales.

El modelo original se plantea de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 ANALF + \beta_2 SPRI + \beta_3 SAGUA + \beta_4 SDRE + \beta_5 SENERG + \beta_6 PTIE + \beta_7 HAC + \beta_8 LOC + \beta_9 SALMIN$$

donde:

$$y_i = \begin{cases} 1 & \text{si en el municipio } i \text{ ganó la oposición} \\ 0 & \text{en otro caso} \end{cases}$$

ANALF indica el porcentaje de la población de 15 años y más analfabeta.

SPRI indica el porcentaje de la población de 15 años y más sin primaria completa.

SAGUA indica el porcentaje de ocupantes en viviendas sin agua entubada.

SDRE indica el porcentaje de ocupantes en viviendas sin drenaje ni excusado.

SENERG indica el porcentaje de ocupantes en viviendas sin energía eléctrica.

PTIE indica el porcentaje de ocupantes en viviendas con piso de tierra.

HAC indica el porcentaje de viviendas con algún nivel de hacinamiento.

LOC indica el porcentaje de población en localidades con menos de 5000 habitantes.

SALMIN indica el porcentaje de población ocupada con ingresos de hasta dos salarios mínimos.

3. MODELO FINAL: ANÁLISIS DE RESULTADOS

Antes de proceder a la estimación del modelo, se analizaron las correlaciones entre las variables, las cuales se presentan en la siguiente tabla:

	ANALF	SPRI	SAGUA	SDRE	SENERG	PTIE	HAC	LOC	SALMIN
ANALF	1	0.8035	0.7503	0.7666	0.6822	0.7837	0.6956	0.3630	0.4270
SPRI		1	0.7008	0.7986	0.6569	0.6809	0.4865	0.6477	0.6019
SAGUA			1	0.7934	0.8369	0.6823	0.4884	0.4146	0.3804
SDRE				1	0.7086	0.6357	0.5323	0.5800	0.6278
SENERG					1	0.7193	0.4214	0.4768	0.4227
PTIE						1	0.5714	0.4284	0.3102
HAC							1	0.0820	0.2245
LOC								1	0.5908
SALMIN									1

Como puede observarse se tiene una alta correlación entre las variables, por lo cual no fue posible incluirlas todas, sin embargo, el modelo final incluye por lo menos una variable de cada uno de los tres grupos originales.

Se aplicaron tres procedimientos para determinar las variables definitivas:

1. Se aplicó el procedimiento stepwise³ para una regresión por Mínimos Cuadrados Ponderados (MCP)⁴ para un modelo Logit. El procedimiento *stepwise* toma en cuenta la correlación entre todas las variables y va incluyendo en el modelo las que guardan la mayor relación con la variable endógena:

$$\frac{Y_i}{\sqrt{w_i}} = \beta_0 \frac{1}{\sqrt{w_i}} + \beta_1 \frac{ANALF}{\sqrt{w_i}} + \beta_2 \frac{SPRI}{\sqrt{w_i}} + \beta_3 \frac{SAGUA}{\sqrt{w_i}} + \beta_4 \frac{SDRE}{\sqrt{w_i}} + \beta_5 \frac{SENERG}{\sqrt{w_i}} + \beta_6 \frac{PTIE}{\sqrt{w_i}} + \beta_7 \frac{HAC}{\sqrt{w_i}} + \beta_8 \frac{LOC}{\sqrt{w_i}} + \beta_9 \frac{SALMIN}{\sqrt{w_i}} + \varepsilon_i$$

³Del paquete estadístico Statgraphics.

⁴Al usar ponderadores, se está corrigiendo el problema de heterocedasticidad inherente al modelo.

2. El modelo Probit:

$$Z_i = \beta_0 + \beta_1 ANALF + \beta_2 SPRI + \beta_3 SAGUA + \beta_4 SDRE + \beta_5 SENERG + \beta_6 PTIE + \beta_7 HAC + \beta_8 LOC + \beta_9 SALMIN + \varepsilon_i$$

3. El modelo Logit.

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 ANALF + \beta_2 SPRI + \beta_3 SAGUA + \beta_4 SDRE + \beta_5 SENERG + \beta_6 PTIE + \beta_7 HAC + \beta_8 LOC + \beta_9 SALMIN + \varepsilon_i$$

El resultado de la aplicación de dichos procedimientos fue el siguiente: de los indicadores sobre hacinamiento y vivienda, entraron al modelo SAGUA y PTIE; de los indicadores de educación, ANALF; y finalmente, como indicador del ingreso, SALMIN.

Los signos esperados de los coeficientes son negativos, ya que al tratarse de variables que están indicando déficits en las diferentes áreas, se esperaría que un incremento en cualquiera de ellos, repercutiera de manera negativa en el porcentaje de votos para la oposición.⁵ De manera que los modelos estimados quedaron como:

Mínimos cuadrados ponderados:⁶

$$\frac{Y_i}{\sqrt{w_i}} = 0.86394 + 0.034798 ANALF - 0.011378 SAGUA - 0.018952 PTIE - 0.008449 SALMIN$$

(0.1224)	(0.0062)	(0.0022)	(0.0035)	(0.0027)
(7.0567)	(5.5579)	(-5.0721)	(-5.4020)	(-3.1063)

Modelo Probit:

$$Z_i = 2.0891 + 0.0722 ANALF - 0.0211 SAGUA - 0.0305 PTIE - 0.0322 SALMIN$$

(0.5456)	(0.209)	(0.0078)	(0.0082)	(0.0089)
(3.8289)	(3.4517)	(-2.6723)	(-3.6789)	(-3.5913)

Modelo Logit:

$$\ln\left(\frac{P_i}{1-P_i}\right) = 3.397868 + 0.121126 ANALF - 0.036083 SAGUA - 0.050326 PTIE - 0.052557 SALMIN$$

(0.914255)	(0.035383)	(0.013669)	(0.014102)	(0.015078)
(3.716542)	(3.423253)	(-2.639861)	(-3.568670)	(-3.485584)

Los resultados fueron consistentes, ya que en todos los casos los signos coinciden y las variaciones en los parámetros no son muy grandes. El valor de la estadística T indica que todas las variables son significativas al 95%, es decir, todas aportan información para cuantificar las probabilidades de votar a favor del partido. Un nivel de significancia del 95% se interpreta como que con una confiabilidad del 95%, el valor del parámetro es diferente de

⁵ Estamos considerando, en base a lo que otros autores han encontrado, que la relación entre los votos por el PRI y el índice de marginalidad es positiva.

⁶ Los valores entre paréntesis del primer renglón (en todos los modelos) corresponden al error estándar del coeficiente, mientras que los del segundo renglón corresponden a la estadística T.

cero. Para probar la bondad de ajuste del modelo, se calculó la R ajustada para el modelo estimado por mínimos cuadrados ponderados, resultando de 0.6141; y el índice de cocientes de verosimilitudes para el Logit y el Probit, obteniéndose para el modelo Logit $r = 0.8505$ y para el modelo Probit $r = 0.8497$

El ajuste del modelo es mucho mejor cuando se emplean los modelos Logit o Probit. En ambos casos, los modelos explican gran parte de la variabilidad de la probabilidad de votar por la oposición, de manera que pueden emplearse para predecir probabilidades esperando sean correctas con un alto grado de confiabilidad. Como se observa, la mayoría de los signos resultaron igual a los que se esperaban, sin embargo, para la variable que indica el porcentaje de la población analfabeta de 15 años o más, el signo resultó positivo.

Buscando encontrar la razón de esta relación, se buscó diferenciar el efecto de las variables sobre la probabilidad de votar por la oposición en los diferentes estados (para lo que se utilizaron variables ficticias), encontrándose diferencias significativas para las variables ANALF, SALMIN Y PTIE.

En el caso del porcentaje de población de 15 años y más analfabeta, los resultados indicaron que existe un efecto diferencial significativo (ver la estadística T) para Guanajuato y Jalisco.

Variable	Coeficiente	Error Std.	Estadística T	Probabilidad
C	4.726114	1.196747	3.949133	0.0001
ANALF	-0.504634	0.254116	-1.985843	0.0483
SAGUA	-0.008979	0.014635	-0.613534	0.5401
PTIE	-0.036007	0.015465	-2.328300	0.0208
SALMIN	-0.044639	0.019484	-2.291015	0.0229
CHAN	0.177158	0.232183	0.763011	0.4463
GAN	0.526667	0.239974	2.194681	0.0292
JAN	0.434985	0.231794	1.876598	0.0619

Como se puede observar, el efecto diferencial para Guanajuato, resultó significativo y positivo; su magnitud fue tal, que la relación entre el porcentaje de población analfabeta y la probabilidad de votar por la oposición dejó de ser negativa. Esto indica que, en el estado de Guanajuato, un aumento en el porcentaje de la población analfabeta incrementa la probabilidad de votar por la oposición.

En el caso de Jalisco, se obtuvo una relación negativa y significativa, al igual que en Baja California y Chihuahua, lo que se interpreta como que un incremento en la población analfabeta implicaría disminuciones en la probabilidad de votar por la oposición. La relación entre el analfabetismo y la probabilidad de votar por la oposición, cuando se hace la diferenciación por estado, es congruente con lo que se esperaría *a priori*, excepto en el caso de Guanajuato, por lo que podríamos suponer que el signo positivo en el modelo general se puede atribuir al hecho de que en este modelo hay un factor de confusión debido a que se

están mezclando los efectos de todos los estados y domina la fuerza que ejerce Guanajuato sobre el resto.

Al considerar la variable pisos de tierra, se encontró que existía un efecto diferencial significativo y positivo para el estado de Guanajuato.

Variable	Coeficiente	Error Std.	Estadística T	Probabilidad
C	4.176634	1.133407	3.685025	0.0003
ANALF	0.019501	0.047503	0.410519	0.6818
SAGUA	-0.018967	0.014077	-1.347440	0.1792
PTIE	-0.232104	0.131852	-1.760339	0.0797
SALMIN	-0.048149	0.018559	-2.594317	0.0101
CHPT	0.001215	0.143458	0.008469	0.9932
GPT	0.228998	0.134415	1.703665	0.0898
JPT	0.186083	0.130234	1.428830	0.1544

Donde la probabilidad de votar por la oposición se relaciona de manera negativa con incrementos en los déficits en viviendas con piso de tierra y el porcentaje de la población que gana hasta dos salarios mínimos. Es decir, en todos los estados, si las condiciones referentes a ambas variables empeoran, la probabilidad de votar por la oposición cae.

La siguiente variable para la cuál se encontró un efecto diferencial significativo por estado, fue el porcentaje de la población ocupada que percibe hasta dos salarios mínimos, donde nuevamente las diferencias se observan para los estados de Guanajuato y Jalisco, que tienen un efecto diferencial significativo y positivo respecto a la variable salario mínimo.

Variable	Coeficiente	Error Std.	Estadística T	Probabilidad
C	4.884397	1.188634	4.109251	0.0001
ANALF	-0.084148	0.057499	-1.463463	0.1447
SAGUA	-0.002446	0.015053	-0.162483	0.8711
PTIE	-0.032700	0.015626	-2.092669	0.0375
SALMIN	-0.100132	0.037747	-2.652727	0.0086
CHPT	0.029070	0.027318	1.064164	0.2884
GPT	0.084365	0.030363	2.778570	0.0059
JPT	0.051529	0.027385	1.881682	0.0612

La relación entre las variables y la probabilidad de votar por la oposición sigue siendo negativa, es decir, si disminuye el porcentaje de la población que gana hasta dos salarios mínimos, aumentará la probabilidad de votar por la oposición.

4. CONCLUSIONES

1. Se demostró, con evidencia empírica, que existe una fuerte relación entre las condiciones socioeconómicas de la población y sus preferencias electorales.
- 2.- Se ratificó que la relación entre las variables socioeconómicas y el voto opositor es negativa. Es decir, conforme aumenta el déficit en alguna de las variables consideradas, disminuye la probabilidad de votar por la oposición y, por consiguiente, aumenta la de votar por el PRI.
- 3.- Se corroboraron las conclusiones de otros estudios que sostienen la existencia de una relación positiva entre el analfabetismo e ignorancia con votos a favor del PRI (excepto para el estado de Guanajuato).

Estudio Multicéntrico sobre Resistencia a los Antibióticos en Hospitales de Tercer Nivel en el Distrito Federal

LILIA BENAVIDES PLASCENCIA y ALEJANDRO ALDAMA OJEDA
UAM-Xochimilco *UAM-Azcapotzalco*

1. INTRODUCCIÓN

El problema del aumento en la resistencia a los antibióticos observado en las bacterias en los últimos años es de naturaleza compleja. Se ha atribuido en parte, a la poderosa presión selectiva que ejerce el uso de los antibióticos sobre las poblaciones bacterianas, pero para comprender el comportamiento de la resistencia bacteriana es necesario establecer parámetros específicos para cada microorganismo, cada fármaco y cada medio hospitalario (Fuchs, 1994; ASM, 1995).

Desde los años setenta se manifestó la preocupación de los Comités de Salud Pública Internacionales por el notorio incremento del uso de antibióticos tanto en hospitales como en las comunidades y se intuía que esta práctica podría tener consecuencias negativas en el fenómeno de resistencia a los antimicrobianos. Debido a que las formas de utilización de antibióticos varían de hospital a hospital y dentro de un mismo hospital, de tiempo en tiempo, se ha sugerido que cada hospital debe identificar sus modelos de prescripción y sus problemas de prescripción de antibióticos. Este sería el primer paso a seguir en cualquier programa dirigido a hacer más racional la terapia antimicrobiana y disminuir con esto la resistencia bacteriana a los antibióticos (Kupersztoch, 1994; Murray, 1994; Wolff, 1993).

El presente estudio pretende aportar información confiable sobre el estado de la resistencia a los antibióticos en los Hospitales de III nivel del D. F. así como de la situación en la que se encuentran nuestras instituciones hospitalarias, tanto públicas como privadas, en su esfuerzo por controlar el desarrollo de la resistencia a los antibióticos. Se propone desarrollar una metodología que por su rigor y efectividad en la captación de datos, pudiera ser aplicable a todo el sistema hospitalario nacional. De lograrse el objetivo, se tendrá un panorama real del problema de la resistencia a los antibióticos que presentan las bacterias productoras de infecciones nosocomiales en la Ciudad de México.

2. OBJETIVOS

Conocer los perfiles de susceptibilidad a los antibióticos de las cepas aisladas de infecciones nosocomiales en algunos hospitales de III nivel del D. F.

Identificar las tendencias del uso de antibióticos en cada una de las instituciones hospitalarias en estudio.

Conocer el tipo de medidas que han tomado las instituciones hospitalarias en su afán de modular el problema de la resistencia a los antimicrobianos.

3. HIPÓTESIS

Es alto el porcentaje de resistencias a los antibióticos comúnmente usados en los hospitales del D. F. para tratar las infecciones.

En el D. F., la mayoría de los hospitales no disponen de un sistema adecuado de control de infecciones ni de utilización de antibióticos.

4. METODOLOGÍA

El presente estudio es una encuesta descriptiva, observacional y transversal. Tiene como población objetivo a los Hospitales de III nivel del D. F. Para el cálculo del tamaño de muestra se tomaron en cuenta las recomendaciones sugeridas por la OMS para estudios de uso de medicamentos, la cual considera adecuada una muestra de 20 X 100, al menos 20 instituciones hospitalarias diferentes con 100 aislamientos en cada una de ellas. Debido a que en la Ciudad de México se tienen seis tipos de instituciones hospitalarias y a que en una de ellas, el ISSSTE, sólo existen cuatro Hospitales de III nivel, se tomaron seis grupos con cuatro hospitales cada uno, esto es, 24 Hospitales de III nivel y 100 aislamientos por hospital, para un total de 2400 aislamientos (OPS/HSS, 1994; HSD/SILOS, 1991:178; SSA, 1981).

Grupos de hospitales y claves asignadas a cada uno:	Número de Hospitales de III Nivel	
Institutos Nacionales de Salud	IN	7
Secretaría de Salud	SS	5
IMSS	IM	12
ISSSTE	IS	4
Privados	PR	9
Otros	O	6
POBLACIÓN TOTAL		43
Población Muestreada		24

La selección de las instituciones hospitalarias se hizo con base en el DIRECTORIO DE HOSPITALES LATINOAMERICANO Y DEL CARIBE, OMS-OPS. Los cuatro hospitales de cada grupo se seleccionaron aleatoriamente con excepción de los del ISSSTE que sólo son cuatro, formando seis estratos representativos homogeneizados. Representativos porque se incluyen en la muestra todos los tipos de Hospitales de III nivel existentes en el D. F., homogeneizados porque se consideraron en la muestra hospitales de III nivel con mínimo de 100 camas y que cumplieran con los siguientes criterios de inclusión:

- con dos especialidades médicas cuando menos;

- que ofrezcan como mínimo dos servicios de alta tecnología;
- que imparten al menos un curso en áreas médicas; y
- con un área de investigación como mínimo (no para los privados).

En cada hospital se colectarán en el mismo período de tiempo, los datos de 100 aislamientos provenientes de igual número de casos de infecciones nosocomiales y de éstos se registrarán las cepas resistentes a los antibióticos probados, esto es, los 2,400 aislamientos bacterianos.

Se levantará información en las instancias siguientes dentro de cada hospital:

- Administración del hospital.
- Laboratorio de diagnóstico microbiológico.
- Comité de Control de Infecciones.
- Farmacia del hospital.

Se verificará la confiabilidad de los laboratorios de diagnóstico de los hospitales en estudio, por un control realizado externamente.

Se analizarán los datos relativos a:

- b) Aislamientos bacterianos procedentes de infecciones nosocomiales.
- c) Registros del laboratorio y de la Farmacia del Hospital.

Se revisarán:

Los reglamentos relativos a utilización de antibióticos y al control de infecciones en cada Hospital.

5. VARIABLES

La **resistencia bacteriana** registrada como diámetro de inhibición en mm o como concentración mínima inhibitoria en ug/ml, será considerada como la variable respuesta a ser explicada por variables endógenas y exógenas. Serán consideradas como variables exógenas: el tipo de hospital, y la especialidad médica del mismo. Como variables endógenas se tienen: la reglamentación, la infraestructura, las cepas nosocomiales aisladas y los antibióticos consumidos.

6. FORMAS DE CAPTACIÓN DE INFORMACIÓN

Los datos se captarán en formas únicas y específicas para cada instancia en las modalidades de cuestionarios y registros de laboratorio.

Se elaborará una base de datos específica para cada forma.

7. LOGÍSTICA

El estudio comprende tres etapas:

- I. Validación de instrumentos (cuestionarios y formas de registros).
- II. Levantamiento de información y entrenamiento de personal de laboratorio para el registro de la información.
- III. Procesamiento de datos, análisis y presentación de resultados.

8. ANÁLISIS ESTADÍSTICO

El análisis estadístico se hará utilizando técnicas multivariadas y tendrá como objetivo, encontrar que variables se relacionan directa o indirectamente con los niveles de resistencia a los diferentes antibióticos de las diferentes cepas resistentes.

REFERENCIAS

- Desarrollo y Fortalecimiento de los Sistemas Locales de Salud. El Control de Infecciones Hospitalarias, *Organización Panamericana de la Salud*, HSD/SILOS-12; 1991:178.
- Directorio de Hospitales Latinoamericanos y del Caribe, *Organización Mundial de la Salud*, OPS/HSS, 94.08, 1994.
- Fuchs, Y. (1994). Mecanismos moleculares de la resistencia bacteriana. *Salud Pública de México*, **36**, 428-438.
- Kupersztoch-Portnoy, Y. M. (1981). Antibiotic resistance of Gram negative bacteria in Mexico: relation to drug consumption, In *Molecular Biology, Pathogenicity, and Ecology of Bacterial Plasmids* (S. B. Levy, C. Royston, C. Clowes y E. L. Koenig Eds.) Plenum Publishing, pp. 529-537.
- Murray, B. E. (1994). Can antibiotic resistance be controled? *New England Journal of Medicine*, **330**, 1229-1230.
- Report of the ASM Task Force on Antibiotic Resistance (1995). Suppl. to *Antibacterial Agents and Chemotherapy*
- Sistemas de Servicios de Salud (1981). *Secretaría de Salubridad y Asistencia*, México, **21**.
- Wolff, M. J. (1993). Use and misuse of antibiotics in Latin America. *Clinical Infectious Diseases*, **17** (Suppl 2), 5346-5351.

Una Aplicación de Estimadores de Mínimos Cuadrados en un Modelo de Respuesta Inmunológica

SARA CAMACHO, F. CERVANTES, LUIS F. HOYOS y JOSÉ C. ROMERO

UAM-Azcapotzalco

1. INTRODUCCIÓN

El término inmunidad, según Bach (1990), se puede emplear abarcando todos los aspectos de la defensa contra factores ambientales externos e internos, pero en general su significado se encuentra restringido a las reacciones específicas que resultan de la penetración de materiales biológicos extraños, llamados antígenos.

Un individuo o animal es inmune cuando resiste a un determinado agente patógeno o sus toxinas, aunque la protección puede ser contrarrestada por una dosis excesiva del agente patógeno.

Stauffer y Pandey (1992) propusieron un modelo dinámico discreto, en el tiempo y en el espacio, para simular el comportamiento de un sistema inmunológico, a nivel microscópico, ante la presencia de un virus o antígeno.

Como el modelo se basa en una serie de reglas locales de transición que actúan a nivel microscópico, resulta natural implementarlo mediante autómatas celulares.

El objetivo de este trabajo consiste en explorar las posibilidades que tienen los estimadores de mínimos cuadrados para construir una ecuación que relacione la concentración de virus en el sistema inmunológico, en términos de las concentraciones medias de células H, C, M y del tiempo de evolución del sistema.

Para evitar problemas de multicolinealidad aplicamos el método paso a paso para calcular los estimadores de mínimos cuadrados.

2. SIMULACIÓN DE UN SISTEMA DE RESPUESTA INMUNOLÓGICA

El objetivo de la respuesta inmune es el bloqueo específico, la neutralización o destrucción de los antígenos que han estimulado el sistema inmunológico.

Stauffer y Pandey (1992) y Celada y Seiden (1992) han propuesto independientemente diversos modelos de respuesta inmunológica, este trabajo está basado en la propuesta de los primeros:

El modelo contempla los siguientes elementos:

V = virus o antígeno
M = células macrófagos
H = células asistentes
C = células citotóxicas

La concentración de cada célula puede ser:

$$\text{alta} = 1 \quad \text{o} \quad \text{baja} = 0$$

Las reglas de interacción entre los elementos V, M, C y H son:

- I. En ausencia de células citotóxicas, el virus crece si presenta alta concentración o existan células macrófagicas o asistentes.
- II. Las citotóxicas crecen sólo si existen macrófagos y asistentes.
- III. En ausencia del virus, las asistentes crecen con la presencia de macrófagos y asistentes.
- IV. Las macrófagos crecen por sí mismas o en presencia de virus

A partir de los supuestos I a IV se construyó un autómata celular, es decir, un sistema dinámico discreto en el tiempo y en el espacio donde las reglas de evolución están definidas de forma local (Macintosh, 1990).

En nuestro caso, el espacio celular unidimensional está constituido por 960 unidades que contienen diferentes concentraciones de los cuatro elementos del sistema inmunológico a nivel celular, V, M, C y H.

La evolución temporal sincrónica está definida por las ecuaciones:

$$\text{I. } V = V * (1 - C) + M + H$$

$$\text{II. } C = M * H$$

$$\text{III. } H = (1 - V) * (M * H)$$

$$\text{IV. } M = M + V$$

Cada ecuación establece la evolución de la concentración de cada elemento en el sistema inmunológico.

La adición (+) y multiplicación (*) se implementan mediante operadores booleanos OR y AND respectivamente, con la regla especial $1 + 1 = 1$.

De esta forma, si partimos de:

$$C = H = M = 0, \quad V = 1$$

y consideramos el vector (C,H,M,V):

$$(0, 0, 0, 1) \rightarrow (0, 0, 1, 0) \rightarrow (0, 0, 1, 0)$$

llegamos a un punto fijo y el sistema inmunológico gana al destruir el virus.

El modelo propuesto es fenomenológico, es decir, describe el comportamiento cualitativo de un sistema dinámico, sin embargo resulta interesante poder caracterizar las reglas empíricas microscópicas de evolución a través de cantidades macroscópicas observables, es decir, obtener una ecuación de concentración media del virus en el sistema en términos de las concentraciones medias de C, H, y M y del tiempo transcurrido al momento de la observación.

3. ESTIMADORES DE MÍNIMOS CUADRADOS

Se establecieron configuraciones aleatorias iniciales para los vectores (C,H,M,V), y se observaron las concentraciones medias, es decir, el número de células de alta concentración entre el número total de células durante diez unidades de tiempo (tiempo promedio de convergencia a un estado estable). Se efectuaron cinco simulaciones, obteniendo un total de 50 observaciones. Todos los cálculos se hicieron empleando SAS.

Definimos:

$$Y = \bar{V} \quad X_1 = \bar{C} \quad X_3 = \bar{M}$$

$$X_2 = \bar{H} \quad X_4 = T$$

Variable	Parametro estimado	Error estandar
Intersec.	- 0.037601	0.03592207
X ₁	- 0.193675	0.23891351
X ₂	- 0.247264	0.30459055
X ₃	0.391963	0.7458454
X ₄	0.022318	0.01006498

El modelo tiene un coeficiente de determinación múltiple de solamente 0.6715

Para evitar problemas de multicolinealidad se aplicará el método de paso a paso:

Variable	Parametro estimado	Error estandar
Intersec.	- 0.03765781	0.03578861
X ₁	- 0.37069864	0.09724391
X ₃	0.42077677	0.06535420
X ₄	0.01733740	0.00794949

Variable	R**2 Parcial	R**2 Acumulada	F	Step Prob>F
X ₃	0.5598	0.5598	61.0397	0.0001
X ₁	0.0724	0.6322	9.2489	0.0038
X ₄	0.0345	0.6666	4.7565	0.0343

Al analizar el comportamiento de los residuales y por construcción de las ecuaciones de transición del modelo de respuesta inmunológica, removemos heterocedasticidad al definir:

$$Y = \bar{V} * \bar{C} \quad X_1 = \bar{C}^2 \quad X_3 = \bar{M} * \bar{C}$$

$$X_2 = \bar{H} * \bar{C} \quad X_4 = T * \bar{C}$$

y el coeficiente de determinación múltiple mejora de 0.666 a 0.915, además de que ahora todas las variables explicativas participan en el modelo lineal.

Variable	Parámetro estimado	Error estándar
Intersec.	- 0.00678104	0.00557560
X ₁	- 0.62508637	0.07340728
X ₃	0.59443365	0.04903589
X ₄	0.00845221	0.00423644

Variable	R**2 Parcial	R**2 Acumulado	F	Step Prob>F
X ₃	0.7794	0.7794	169.5998	0.0001
X ₁	0.1288	0.9083	66.0080	0.0001
X ₄	0.0073	0.9156	3.9805	0.0520

4. CONCLUSIONES

Los estimadores de mínimos cuadrados nos permiten construir una ecuación en términos de medidas macroscópicas de un modelo inmunológico definido a partir de interacciones microscópicas.

Esto no excluye la futura aplicación de diversas técnicas de estadísticas, (modelos no lineales por ejemplo) que complementen o mejoren el modelo predictivo.

Por otra parte existe un interesante horizonte teórico en la aplicación de métodos estadísticos que nos aproximen a la caracterización de autómatas celulares.

REFERENCIAS

- Bach, T. F. (1990). *Immunología*, 1, Editorial Ciencia y Técnica.
 Celada, F. and Seiden, P. E. (1992). A computer model of cellular interactions in the immune systems. *Immunology Today*, 13, 56-62.
 Macintosh, H. V. (1990). *Linear cellular automata*. Universidad Autónoma de Puebla.
 Stauffer, Dietrich and Pandey, R.B. (1992), Immunologically motivated simulations of cellular automata. *Computer in Physics*, 6, 404 -410.

Programación Estocástica: una Alternativa al Estudio de Conglomerados

SERGIO G. DE LOS COBOS S.

BLANCA R. PÉREZ S.

U.A.M-Iztapalapa

y

MIGUEL A. GUTIÉRREZ A.

U.A.M.-Azcapotzalco

1. INTRODUCCIÓN

En situaciones prácticas, uno de los problemas computacionales es *la explosión combinatoria*. Cabe mencionar que, para instancias grandes, no es posible encontrar solución óptimal en tiempos de cómputo razonables, incluso, para problemas de programación lineal, se han encontrado instancias “patológicas” donde se observa que el método simplex no tiene la propiedad de ser un algoritmo de tiempo polinomial.

En la actualidad existe un gran esfuerzo por parte de la comunidad investigadora para el diseño de buenas heurísticas¹, i.e., algoritmos eficientes con respecto al tiempo de cómputo y al espacio de memoria, y con cierta verosimilitud de entregar una solución “buena” (relativamente cercana a la óptima) mediante el examinar sólo un pequeño subconjunto del número total de soluciones posibles. La característica sobresaliente de los métodos de programación estocástica presentadas en este trabajo es su aplicación general.

La búsqueda tabú, junto con la técnica del recocido simulado y los algoritmos genéticos, han sido singularmente calificados por el Committee on Next Decade of Operations Research (1988) como “extremadamente promisorios” para el tratamiento futuro de aplicaciones prácticas.

2. BÚSQUEDA TABÚ

La Búsqueda Tabú (BT) es un procedimiento heurístico de “alto nivel” introducido y desarrollado en su forma actual por Fred Glover (1989) y (1990). En términos generales, el método BT puede esbozarse consistente en:

1. Se desea moverse paso a paso desde una solución factible inicial de un problema de optimización combinatoria hacia una solución que proporcione el valor mínimo de la función objetivo C . Para esto, se puede representar a cada solución por medio de un punto s (en algún espacio) y se define una vecindad $N(s)$ de cada punto s .

2. El paso básico del procedimiento consiste en empezar desde un punto factible s y generar un conjunto de soluciones en $N(s)$; entonces se escoge al mejor vecino generado s^* y se posiciona en ese nuevo punto ya sea que $C(s^*)$ tenga o no mejor valor que $C(s)$.

3. La característica importante de la búsqueda tabú es precisamente la construcción de una lista tabú T de movimientos: aquellos movimientos que no son permitidos

¹La palabra heurística proviene de la palabra griega *heuriskein* que significa encontrar o descubrir.

(movimientos tabú) en la iteración presente. Las condiciones tabú tienen la meta de prevenir ciclos e inducir la exploración de nuevas regiones.

4. Las restricciones tabú no son inviolables bajo toda circunstancia. Cuando un movimiento tabú proporciona una solución mejor que cualquier otra encontrada, su clasificación tabú puede eliminarse. La condición que permite dicha eliminación se llama *criterio de aspiración*.

5. Ahora bien, conforme la búsqueda progresá, la forma de la evaluación empleada por la búsqueda tabú llega a ser más adaptativa, incorporando referencias concernientes para la *intensificación* y la *diversificación* regional de búsqueda.

3. RECOCIDO SIMULADO

3.1 *El proceso de recocido de un sólido*

El algoritmo de recocido simulado está basado en una analogía entre la simulación de recocido de sólidos y la problemática de resolver problemas de optimización combinatoria de gran escala. Por esta razón el algoritmo se conoce como **recocido simulado**. Recocido denota un proceso de calentamiento de un sólido a una temperatura en la que sus granos deformados recristalizan para producir nuevos granos. La temperatura de Seguida a la fase de calentamiento, viene un proceso de enfriamiento en donde la temperatura se baja poco a poco. De esta manera, las partículas se reacomodan en estados de más baja energía hasta que se obtiene un sólido con sus partículas acomodadas conforme a una estructura de cristal.

El equilibrio térmico está caracterizado por la distribución de Boltzmann (Toda et al., 1983).

3.2 *Aspectos generales*

Las soluciones se generan continuamente tratando de transformar la solución actual en una subsecuente por medio de aplicar los mecanismos de generación y el criterio de aceptación. Las aplicaciones del algoritmo de recocido simulado requieren de la especificación de los siguientes puntos:

1. Una descripción concisa de la **representación del problema** consiste de una representación del espacio de soluciones y una expresión de la función de costo.
2. La generación de ensayos para transformar la solución actual en una subsecuente consiste de tres pasos. Primero, se debe generar una nueva solución aplicando un mecanismo de generación. Enseguida se debe calcular la diferencia de costo de las dos soluciones, por último, se hace una decisión de aceptar o no, la nueva solución.
3. Ejecutar el proceso de recocido, requiere de la especificación de los parámetros que determinan el programa de enfriamiento. Estos parámetros son el valor inicial del parámetro de control, una función que especifique el decremento del parámetro de

control, la longitud de cada bloque donde permanece constante el parámetro de control y el criterio de paro.

4. ALGORITMOS GENÉTICOS

Los algoritmos genéticos fueron desarrollados por Holland en la Universidad de Michigan por los 60's. El nombre de algoritmos genéticos se origina de la analogía entre la estructura genética de los cromosomas y de la representación de las soluciones de problemas por medio de anillos o vectores.

4.1 Componentes básicos de los AG

Un algoritmo genético básico tiene tres operadores: *reproducción, cruzamiento y mutación*. La reproducción es un apareamiento aleatorio de individuos (soluciones muestra) de una población para crear una o más descendencias. El cruzamiento define el resultado como un cambio de gene (plan reproductivo), cuyo valor específico es conocido como *allele*, i.e., los alleles pueden concebirse como instancias en el sentido de sistemas expertos. El cambio de genes (tipo de información y sus atributos) siguen las reglas posicionales tradicionalmente modelada después de la reproducción biológica. Finalmente, una mutación es simplemente el introducir un elemento aleatorio. La mutación diversifica el espacio de búsqueda y protege de la pérdida de material genético que puede darse en la reproducción y el cruzamiento.

Existen muchas variaciones de los algoritmos genéticos debido al uso de diferentes operadores de reproducción, cruzamiento y mutación. En general, un algoritmo genético para resolver un problema combinatorio debe contener cinco componentes principales (Davis y Streenstrup, 1987).

1. Una representación cromosómica de las configuraciones del espacio de búsqueda del problema.
2. Una forma de generar la población inicial de soluciones.
3. Una función de evaluación que juega el papel del medio ambiente, calificando las configuraciones en términos de su “valor de ajuste”.
4. Una descripción precisa de las operaciones genéticas que alteran la composición de los “hijos” durante la reproducción.
5. Los valores de los parámetros que el algoritmo genético utiliza (tamaño de la población, probabilidades de aplicar las operaciones genéticas, etc.)

Holland y Golberg, desarrollan el concepto de *esquema* o patrón de similaridad para analizar el problema de convergencia para algoritmos que trabajan con representaciones de cadenas de bits. Un esquema es un patrón en el que ciertas posiciones del cromosoma son

fijas. A través de la noción de esquemas, el teorema fundamental de los algoritmos genéticos, proporciona una cota inferior a la evaluación para un tipo de esquema en una generación, por lo que, se pueden mantener “buenos” esquemas después de un cierto número de generaciones.

5. ANÁLISIS DE CONGLOMERADOS

Básicamente el objetivo del Análisis de Conglomerados(AC), es el de obtener una partición de un conjunto de objetos basada ésta en las similitudes, o en las “distancias” entre los objetos (conjuntos independientes), de forma que, los objetos agrupados (conglomerados) en la misma clase (grupo, o conglomerado) sean similares o cercanos entre sí.

5.1 Conglomerados Jerárquicos

En el análisis de conglomerados jerárquicos, se tienen O objetos inicialmente considerados como O conglomerados y el análisis procede secuencialmente a agrupar éstos dentro de conglomerados mayores hasta que, todos los objetos formen un solo conglomerado.

Las tres elecciones más comunes para definir la distancia entre conglomerados son las referentes a la optimización de las distancias entre éstos, y son: conglomerados de liga simple, conglomerados por diámetro y conglomerados de liga promedio.

5.2 Conglomerados No-jerárquicos

En el análisis de conglomerados no-jerárquicos, la atención se enfoca a un número específico de conglomerados, por lo que, se desea satisfacer el criterio de “menos conglomerados” en base de: a) compactación intra-conglomerados y b) separación entre-conglomerados.

En este caso, los conglomerados inician desde una partición de los objetos de manera iterativa a través de un agrupamiento en conglomerados de mayor agrupación (menor disimilitud o distancia). En general se transfiere un objeto de un conglomerado a otro en una iteración para producir un incremento máximo en la optimización.

En el caso particular de considerar un conjunto de objetos con masas en el espacio Euclídeo ponderado, la mayor agrupación (conglomerado) está definido cuando se minimiza la inercia entre-conglomerados, lo que es equivalente a minimizar la inercia intra-conglomerados. En este caso, la ganancia respecto a la inercia entre-conglomerados es con respecto a transferir uno de los objetos evaluados respecto de su distancia al centroide, dicho ajuste depende de las masas de los conglomerados y del objeto. Los centroides de los nuevos conglomerados son recalculados después de cada iteración y el proceso continúa hasta que ningún objeto pueda cambiarse de conglomerado.

REFERENCIAS

- Committee on the Next Decade of Operations Research (1988). *Operations Research: The Next Decade*. Ops. Res. **36**.
- Davis L. and Streenstrup, M. (1987). *Genetic Algorithms and Simulated Annealing*, Davis (ed.), Pittman, London.
- De los Cobos Silva S. (1994). La Técnica de la Búsqueda Tabú y sus Aplicaciones, *Tesis doctoral*, DEP-FI, UNAM.
- Glover F. (1989). Tabu Search, Part I, ORSA *Journal on Computing* **1**, 190-206.
- Glover F. (1990). Tabu Search, Part II, ORSA *Journal on Computing* **2**, 4-31.
- Goldberg D. E., (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, Mass.
- Grenacree M. J., (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, Inc.
- Gutiérrez-Andrade M. (1991). La Técnica del Recocido Simulado y sus Aplicaciones, *Tesis de Doctorado*, D.E.P.F.I.-U.N.A.M., México.
- Holland, J. (1975). *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI.
- Metropolis, N., Rosenbluth M., Rosenbluth A., Teller A., and Teller E., (1953), Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physcis*, **21**, 1087-1092.
- Murillo F. A. (1996). Particionamiento usando Búsqueda Tabú, *IV Encuentro Centroamericano de Investigación Matemática*, Antigua, Guatemala.
- Piza V. E. y Trejos Z. J. (1996). Clasificación Automática; Particionamiento mediante Sobrecalentamiento Simulado, *IV Encuentro Centroamericano de Investigación Matemática*, Antigua, Guatemala.
- Trejos Z. J., Piza V. E. y Figueroa M. G. (1996). Clasificación automática mediante un algoritmo genético: resultados numéricos, *IV Encuentro Centroamericano de Investigación Matemática*, Antigua, Guatemala.
- Toda, M., Kubo, R. and Saito, M. (1983). *Stastistical Physcis*, Springer-Verlang.

Obtención de Intervalos de Verosimilitud-Confianza para los Parámetros de una Mezcla de Distribuciones Weibull

ELOÍSA DÍAZ-FRANCÉS M. y ENRIQUE VILLA DIHARCE

CIMAT

1. INTRODUCCIÓN

La presencia de la distribución Weibull es muy común en el análisis de datos de confiabilidad, donde también lo es el tener tiempos de falla provenientes de poblaciones mezcladas, ya sea porque tenemos un sistema compuesto por varios componentes o porque consideramos tiempos de falla de un lote de componentes que provienen de diferentes fabricantes.

Consideremos una mezcla de dos distribuciones Weibull, cuya función de densidad es:

$$f(t, \varphi) = p \left(\frac{\beta_1}{\alpha_1} \right) \left(\frac{t}{\alpha_1} \right)^{\beta_1-1} \exp \left[- \left(\frac{t}{\alpha_1} \right)^{\beta_1} \right] + (1-p) \left(\frac{\beta_2}{\alpha_2} \right) \left(\frac{t}{\alpha_2} \right)^{\beta_2-1} \exp \left[- \left(\frac{t}{\alpha_2} \right)^{\beta_2} \right]$$

donde $\varphi = (\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5) = (p, \alpha_1, \beta_1, \alpha_2, \beta_2)$ y $\varphi, t > 0$.

El problema aquí consiste en hacer inferencia sobre un parámetro de interés, tomando a los demás, como de estorbo. Si se requieren afirmaciones de estimación por intervalo sobre φ_i , el procedimiento usual es considerar la distribución asintótica normal del estimador de máxima verosimilitud de φ_i para construir la siguiente cantidad pivotal que también converge a una variable aleatoria ε con distribución normal estándar. Esto es,

$$u_{\varphi_i} = u(\varphi_i, X) = \left(\varphi_i - \hat{\varphi}_i \right) \sqrt{I_{\hat{\varphi}_i}} \xrightarrow{d} \varepsilon \sim N(0,1),$$

donde u_{φ_i} es una cantidad pivotal (i.e. una variable aleatoria que es función de los parámetros pero que tiene una distribución conocida e independiente de ellos), X es la muestra observada, $\hat{\varphi}_i$ es el estimador de máxima verosimilitud del parámetro de interés φ_i , y $I_{\hat{\varphi}_i}$ es la información observada. Entonces las afirmaciones de estimación para φ_i en su forma más simple serían

$$\varphi_i = \hat{\varphi}_i \pm \sqrt{I_{\hat{\varphi}_i}} \varepsilon, \quad \text{donde } \varepsilon \sim N(0,1). \quad (1)$$

Sin embargo, puede ocurrir que para muestras finitas el pivotal no tenga distribución normal; aún mas, la distribución asociada podría ser asimétrica, en tales casos, las afirmaciones de estimación (1) serían incorrectas.

La información sobre el parámetro de interés está contenida en la verosimilitud perfil o maximizada de dicho parámetro, considerando los restantes en la mezcla como parámetros de estorbo. El objetivo es obtener una cantidad pivotal lineal aproximada u_{φ_i} , con una densidad $\log F$ adecuada, con aproximación suficientemente buena a la verosimilitud perfil observada de φ_i . Esta aproximación pivotal puede llevar a inferencias sobre φ_i simples y eficientes de la forma

$$\hat{\varphi}_i = \hat{\varphi}_i - \sqrt{I_{\hat{\varphi}_i}} \varepsilon, \quad \text{donde } \varepsilon \sim \log F(m, n),$$

donde m y n son los grados de libertad que se deben determinar con base en la muestra observada. El procedimiento es similar al descrito en Viveros y Sprott (1987), donde estiman los parámetros m y n analíticamente al tomar la aproximación de Taylor conservando los términos de cuarto orden, mientras que aquí, se hace approximando por mínimos cuadrados la log-verosimilitud relativa observada, por la log-verosimilitud de una distribución $\log F$. Este método se explica ampliamente en Diaz-Francés y Villa (1995).

2. EJEMPLO

Mendenhall y Hader, (1958), analizaron los tiempos de falla de 369 transmisores de comunicación de una aerolínea comercial. Al fallar, las unidades se enviaban a mantenimiento. Sin embargo, había casos en los que los transmisores que habían fallado, funcionaban satisfactoriamente al llegar al taller. Es por esto que los transmisores se podían clasificar en dos grupos: los que tuvieron falla confirmada en el taller, y los que no se les confirmó la falla. La muestra se censuró a las 630 horas, pues por política de la aerolínea, se renovaban todos los transmisores que hubiesen funcionado este tiempo. A la aerolínea le interesa hacer inferencias sobre los tiempos medios de vida tanto de los transmisores con falla confirmada (Grupo I), como los de fallas no confirmadas (Grupo II). Este es un modelo especial de mezclas, pues de las fallas que fueron observadas, se conoce de cual población provinieron. Sin embargo, de las observaciones censuradas, no se sabe a cual población pertenecían y por tanto se les puede modelar a través de una mezcla de las distribuciones de los Grupos I y II.

Sean f_1 la densidad de las fallas en el Grupo I, f_2 la del Grupo II, p la proporción de observaciones censuradas que provienen del Grupo I, n el tamaño de muestra, r el número de observaciones censuradas (en este caso $n = 369$ y $r = 44$), r_1 el número de observaciones del Grupo I, r_2 las del Grupo II, X_{1i} son las fallas observadas del Grupo I, X_{2j} son las del Grupo II y T es el límite de detección por la derecha. La verosimilitud correspondiente L bajo este modelo es la siguiente:

$$L \propto p^r \prod_{i=1}^{r_1} f_1(x_{1i})(1-p)^{n-r} \prod_{i=1}^{r_2} f_2(x_{2i}) \{1 - pF_1(T) - (1-p)F_2(T)\}^{n-r}$$

Se ajustó este modelo a los datos suponiendo que f_1 y f_2 eran Weibulls para mínimos y las pruebas de bondad de ajuste aplicadas no dieron evidencia en contra de este supuesto. Se calcularon los estimadores de máxima verosimilitud y la información de Fisher observada para los cinco parámetros de este modelo, β_1, β_2 , los parámetros de escala de las dos densidades, α_1, α_2 , los parámetros de forma, y p la proporción de observaciones censuradas que pertenecen al Grupo I. Posteriormente se calculó la verosimilitud perfil de cada uno de los parámetros y se observó que para cuatro de ellos, éstas eran simétricas, mientras que la verosimilitud perfil de β_2 era asimétrica, como se muestra en la Figura 1. Se observó que la aproximación asintótica normal habitual reproducía satisfactoriamente a las verosimilitudes perfiles para estos cuatro parámetros; sin embargo, en el caso de β_2 , claramente se alejaba de la verosimilitud, mientras que la aproximación $\log F$ aquí propuesta la describía satisfactoriamente.

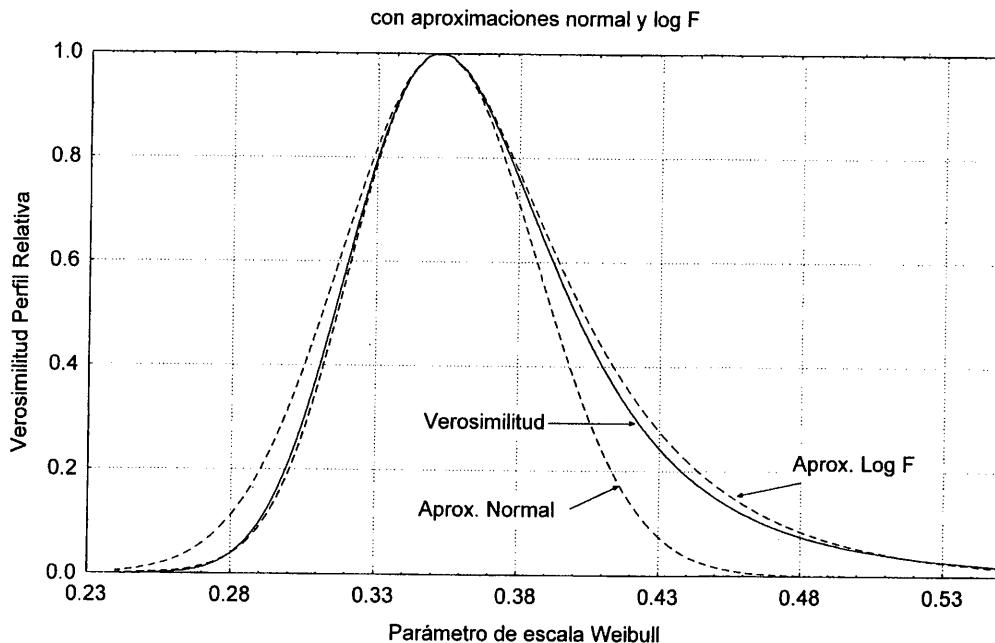


Figura 1. Verosimilitud perfil asimétrica con aproximaciones normal y $\log F$

En la Tabla 1 se muestran los estimadores de máxima verosimilitud, así como los intervalos de verosimilitud-confianza del 95% que se obtienen con la aproximación normal habitual y que en estos casos es satisfactoria. En la Tabla 2 se contrastan los intervalos de distintos niveles de confianza que se obtendrían para β_2 usando la aproximación $\log F$ propuesta y los obtenidos con la aproximación normal habitual. Obsérvese cómo los intervalos normales excluyen a la derecha valores del parámetro que son plausibles, mientras

que por la izquierda incluyen valores con muy baja plausibilidad. En cambio, los intervalos $\log F$ consideran la asimetría de la verosimilitud perfil e incluyen a todos los valores del parámetro que tienen más plausibilidad que los valores excluidos.

TABLA 1
Estimadores máximo verosímiles e intervalos normales del 95 % de confianza.

Parámetro	Estimador	Intervalo del 95 %
β_1	0.561	(.4940, .6289)
α_1	1.125	(.999, 1.251)
β_2	0.3517	(.2845, .4188)
α_2	1.266	(1.034, 1.498)
p	.7025	(.6538, .751)

TABLA 2
Intervalos $\log F$ y normales para β_2 .

Nivel	LogF	Normal
50 %	(206.19, 241.35)	(206.82, 236.25)
90 %	(187.36, 283.94)	(185.97, 257.04)
95 %	(181.37, 302.46)	(179.24, 263.84)
99 %	(169.22, 345.93)	(165.82, 277.26)

Este es un ejemplo en donde la aproximación asintótica normal usual hubiese llevado a inferencias incorrectas para uno de los parámetros debido a la asimetría de la verosimilitud perfil correspondiente.

3. SIMULACIONES

Para ratificar las coberturas de los intervalos $\log F$ de verosimilitud-confianza propuestos, se simularon 6000 muestras de datos con las mismas características de los de Mendenhall y Hader, obteniendo los resultados presentados en la Tabla 3. Se observa que las coberturas de los intervalos $\log F$ y de los normales son muy buenas para todos los niveles de confianza considerados. Sin embargo, los intervalos normales dejan más veces afuera por la derecha al parámetro verdadero que por la izquierda; esto indica como la aproximación normal no reproduce bien a la verosimilitud, que es asimétrica.

La propiedad de cobertura para intervalos es necesaria y relativamente fácil de cumplir, aunque no es la característica más importante que debe satisfacer un conjunto de intervalos. La propiedad de fidelidad, de reproducción de la verosimilitud es mucho más importante y, de la Figura 1, queda claro que los intervalos $\log F$ sí la cumplen mientras que

los normales no. El hecho de que los intervalos normales muestren buena cobertura tal vez se deba a que se compense el error de estos intervalos de incluir a más valores por la izquierda mientras que excluyen a valores plausibles del parámetro por la derecha.

TABLA 3
*Porcentajes de Coberturas de los intervalos log F y normales
 para 6000 muestras, con la proporción de exclusiones
 por la izquierda y la derecha.*

Nivel	Cobertura	LogF		Cobertura	Normal	
		Cola Izq.	Cola Der.		Cola Izq.	Cola Der.
50 %	50.63	24.82	24.55	49.25	24.83	25.92
90 %	90.72	4.83	4.45	89.93	4	6.07
95 %	95.37	2.52	2.12	94.78	1.78	3.43
99 %	99.18	0.42	98.97	0.4	0.2	0.83

4. CONCLUSIONES

Al analizar un conjunto de datos de la naturaleza aquí descrita se recomienda graficar y estudiar las verosimilitudes perfiles de los parámetros de interés. Así se puede observar gráficamente si la aproximación normal reproduce bien o no a la verosimilitud. En caso negativo, conviene explorar la posibilidad de aproximar a la verosimilitud con una densidad $\log F$ y en caso de que ésta tampoco la describa bien, existen otras posibilidades como ajustar una densidad del Nilo (ver Chamberlin, 1989) o tal vez una hiperbólica. Lo que se observó es que para el caso de tiempos de falla, parece ser que la $\log F$ da en general una buena aproximación pues es una familia muy rica de distribuciones. Finalmente, los intervalos que se obtengan mediante la aproximación $\log F$ tendrán la doble ventaja de tener las propiedades de cobertura y de fidelidad.

REFERENCIAS

- Chamberlin, S.R. (1989). Logical Foundation and Application of Inferential Estimation. *PhD Thesis* of the University of Waterloo.
- Díaz-Francés, E. y Villa, E. (1995). Approximations to the Profile Likelihood of a Location Parameter of a Mixture of Gumbel Distributions. *Comunicaciones Técnicas del CIMAT*, No. I-95-16 (PE/CIMAT).
- Mendenhall, W. y Hader, R.J. (1958). Estimation of Parameters of Mixed Exponentially Distributed Failure Time Distributions from Censored Life Test Data. *Biometrika*, **45**, 504-520.
- Viveros, R. y Sprott, D.A. (1987). Allowance for Skewness in Maximum-Likelihood Estimation with Application to the Location-Scale Model. *The Canadian Journal of Statistics*, **15**, 349-361.

Variabilidad no Constante en Modelos de Regresión con Datos Censurados

JORGE DOMÍNGUEZ DOMÍNGUEZ

CIMAT

1. INTRODUCCIÓN

Estimar el tiempo de falla de componentes mecánicos y eléctricos en la manufactura de un artículo, o el tiempo de degradación de un producto son objetivos relevantes dentro del contexto industrial. En ocasiones, es de interés estudiar el tiempo de mantenimiento de un equipo, es importante conocer el tiempo de cambio entre piezas en un equipo. Frecuentemente en estas situaciones existen variables tales como el esfuerzo, temperatura o carga a las que se someten los componentes o productos, estas influyen en la determinación de la respuesta. Para evaluar el desempeño de un producto, existen otras variables de respuesta, como el número de kilómetros recorridos por un automóvil, número de veces que un sistema ha prestado un servicio. En el desarrollo de una estrategia experimental se consideran variables y valores de estas que tienen impacto en aumentar durabilidad de un producto.

La existencia de covariables permiten considerar el estudio de modelos de regresión, con el tiempo de falla como variable de respuesta, los modelos son de utilidad para poder predecir la durabilidad de un equipo. El interés está en establecer la relación entre las variables T : tiempo de falla ($T > 0$) y las covariables X . Para ello, frecuentemente se usa la transformación $Y = \log T$ para representar el tiempo de falla y describir las distribuciones probabilísticas tanto en términos de Y como de T , donde la distribución Y es usualmente más simple. Así el modelo de regresión paramétrico es:

$$Y = \log T = X\beta + \sigma_x \varepsilon \quad (1)$$

β_{px1} , y $\sigma_x = \sigma(x)$ son los parámetros, en general la distribución de probabilidad Y está representada por parámetros de posición y escala, se puede hacer notar que estos guardan una relación con los parámetros de la distribución T .

El objetivo de este trabajo es evaluar la estimación de los parámetros del modelo 1, la idea es estudiar un conjunto de factores y observar como afectan la estimación, para alcanzar el objetivo se plantea un diseño factorial 3^4 , los factores considerados son el tamaño de muestra, los niveles de censura, el grado de variabilidad $\sigma_x = \sigma(x)$ y diferentes procedimientos de estimación. La generación de los datos en este diseño experimental se obtienen mediante un estudio Monte Carlo.

La estructura de los apartados de este trabajo es como sigue, en la segunda parte se describirán las características de los datos. Los procedimientos para estimar los parámetros se muestran en 3, la estrategia experimental se presenta en el apartado 4. Finalmente, en el último apartado se describen los resultados.

2. CARACTERÍSTICA DE LOS DATOS

Dentro del marco de este tema, los métodos estándar para estimar los parámetros β_{px1} , y $\sigma_x = \sigma(x)$, en el modelo 1 se basa en procedimientos de máxima verosimilitud y en torno a este se derivan las propiedades estadísticas, cuyos resultados generalmente son asintóticos. Sin embargo, es de interés en este estudio considerar el comportamiento empírico de estos parámetros en **muestras pequeñas**.

Muestra con censura. Una característica anexa al análisis de este tipo de datos surge porque algunas unidades pueden no fallar antes de un período establecido o en una fase de experimentación. Esta circunstancia da lugar a un conjunto de datos incompletos para el análisis estadístico, esta situación es común en estudios de tiempo de falla y se denomina censura. La censura da lugar a crear situaciones especiales e interesantes dentro del análisis estadístico, la censura aparece de distintas maneras, en Lawless (1982), pag. 31 ss., se realiza una discusión sobre estos conceptos con detalle.

El tipo de censura que se considera para los datos generados por el modelo 1, es el denominado censura tipo I, éste se define como sigue, considere un período de observación C , entonces una unidad observa un tiempo de falla T_i . Si T_i es menor o igual que C entonces la unidad no se censura, en caso contrario la unidad se censura y el tiempo de observación es C . Así, si T_o es un tiempo de observación de una unidad, se tiene que, $T_o = \min(T_i, C)$. En una situación como la planteada, se dice que hay censura por la derecha.

Distribuciones. El interés principal en el análisis de datos de vida, es modelar mediante las distribuciones de probabilidad la variable aleatoria: tiempo de falla, en este trabajo se consideraran las distribuciones Weibull y Valores Extremos (DVE) (Kalbfleisch-Prentice, 1980).

Existe una estrecha relación entre las distribuciones Weibull y DVE. Esta se puede establecer mediante el siguiente procedimiento. La variable T sigue una distribución Weibull $W(\lambda(x), \gamma(x))$ cuya función densidad es:

$$f(t) = \lambda(x)\gamma(x)(\lambda(x)t)^{\gamma(x)-1} \exp(-(\lambda(x)t)^{\gamma(x)}),$$

la función de confiabilidad $S(t)$ para la distribución Weibull toma la forma:

$$S(t) = \exp((-\lambda(x)t)^{\gamma(x)})$$

Si en lugar de trabajar con la variable T , se propone $Y = \log T$, entonces $Y \sim \text{DVE}$, por el método de transformación se sigue que la función de densidad de Y es:

$$f(y) = \exp\left(\frac{y - \mu(x)}{\sigma(x)} - \exp\left(\frac{y - \mu(\mu)}{\sigma(x)}\right)\right) \quad (2)$$

con $y \in (-\infty, \infty)$. La relación de los parámetros en ambas densidades está dada por $\mu(x) = \log \lambda(x)$ y $\sigma(x) = \gamma^{-1}(x)$. La función $S(t)$ para la DVE es:

$$S(t) = \exp\left(-\exp\left(\frac{y - \mu(x)}{\sigma(x)}\right)\right) \quad (3)$$

Es importante notar que la relación entre las variables T y Y en el modelo 1 siguen respectivamente una distribución Weibull y de DVE, esto indica que $\varepsilon \sim \text{DVE}$.

Variabilidad. El análisis de datos del tiempo de falla considerando el modelo 1 supone que la varianza es constante, esto es, $\sigma(x) = cte$. Esta suposición permite asegurar la estimación de los parámetros β , σ_x del modelo por máxima verosimilitud con mayor precisión. En varias ocasiones no se satisface el supuesto de que $\sigma(x) = cte$, dada esta situación surge el interés en estudiar las propiedades estadísticas de los estimadores, en este trabajo se considera el caso de que $\sigma(x) = \exp(\alpha_0 + X\alpha_1)$, entonces se dice que existe efecto de dispersión. El impacto sobre el efecto de dispersión se debe principalmente al parámetro α_1 , es importante evaluar el comportamiento estadístico de su estimador, en este trabajo se hará globalmente mediante la respuesta estimada \hat{Y} . En este planteamiento se considera que la covariable X tiene efecto tanto en la media como en la variabilidad. Las ideas generales sobre este tema se pueden consultar en Carroll-Ruppert(1988), Davidian-Carroll(1987) y McCullagh-Nelder(1989).

3. PROCEDIMIENTOS DE ESTIMACIÓN

El objetivo es estimar los parámetros $\theta = (\beta_0, \beta_1, \alpha_0, \alpha_1) = (\beta, \alpha)$, en el modelo 1. Para alcanzar tal fin se proponen tres procedimientos relacionados con la función de verosimilitud y considerando el caso de censura tipo I, el primero consiste en la estimación conjunta del parámetro θ , éste se denota por PBA, el planteamiento es como sigue: se escribe la función densidad conjunta de la variable tiempo de falla t_i y de la indicadora δ_i , como $L_i(\beta, \alpha) = f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$, donde δ_i es:

$$\delta_i = \begin{cases} 1 & \text{si } t_i \leq C & \text{no hay censura} \\ 0 & \text{si } t_i > C & \text{existe censura} \end{cases}$$

entonces el logaritmo de la función de verosimilitud es:

$$l(\theta) = \log \prod_i^n L_i(\beta, \alpha) = \log \prod_i^n (\delta_i \log f(t_i) + (1 - \delta_i) \log S(t_i)) \quad (4)$$

La variable aleatoria ε en el modelo 1 sigue una DVE, entonces $f(t)$ y $S(t)$ para esta distribución están dadas por las expresiones 2 y 3. Finalmente la función $l(\theta)$ se expresa por:

$$l(\theta) = \log(\exp(x\alpha)) + \sum_i^r \frac{Y_i - x_i\beta}{\exp(x\alpha)} - \sum_i^n \exp \frac{Y_i - x_i\beta}{\exp(x\alpha)}$$

por método el Newton-Raphson se optimiza la función $l(\theta)$, para encontrar los valores estimados de θ .

Aparte del procedimiento de estimación conjunta se plantean otros dos, el primero de ellos se le conoce como seudoverosimilitud y se denotará por PPV, en éste, inicialmente se estima el parámetro β por máxima verosimilitud, el valor estimado de β se sustituye en la expresión 4, luego por máxima verosimilitud se estima el parámetro α , el proceso continúa hasta la convergencia de ambos parámetros. El otro procedimiento consiste en hacer una estimación ponderada del parámetro β por mínimos cuadrados, el siguiente paso es sustituir este valor estimado en la expresión 4 para estimar α por máxima verosimilitud, el proceso termina hasta alcanzar la convergencia, éste se denota por PON.

4. DESCRIPCIÓN DEL DISEÑO EXPERIMENTAL

Con la descripción anterior se plantean una estrategia experimental para evaluar el desempeño estadístico de los estimadores, los factores que se consideran en este estudio se derivan de las características de los datos y de los procedimientos de estimación, esto da lugar a tener cuatro factores cada uno con tres niveles, es decir un diseño factorial 3^4 los factores son:

A: el tamaño de muestra, los niveles que se consideran son 25, 50 y 100.

B: el porcentaje de censura, los niveles planteados son 0%, 25% y 40%.

C: la variabilidad, los niveles dependen del valor de α_1 , y son 0, 1.25, 2.3, α_1 se obtiene por la relación:

$$\alpha_1 = \frac{1}{x_{\max} - x_{\min}} \log\left(\frac{\sigma(x_{\max})}{\sigma(x_{\min})}\right)$$

entonces α_1 depende de la razón $\sigma(x_{\max}) / \sigma(x_{\min})$ y del rango $x_{\max} - x_{\min}$. El primer valor de α_1 propuesto considera el caso de $\sigma(x)$ es una constante, luego para los otros dos valores se tiene que $\sigma(x)$ varía entre (1.41,3.35) y (1.09,5.39) respectivamente.

D: el procedimiento de estimación, los niveles son PBA, PPV y PON.

Los datos que permiten modelar la respuesta se obtienen mediante la siguiente expresión

$$Y = 21 - 6X + \exp(\alpha_0 + X\alpha_1)z,$$

donde los valores de α_0 y α_1 son los que corresponden al planteamiento de los tres niveles del factor C, es decir (0,0), (-3.4,1.25) y (-6.8,2.3).

Es común en estudios de confiabilidad usar una trasformación de las covariables X, en este caso se usó el logaritmo natural, los valores que toma X están comprendidos entre el log20, log40, es decir $3.12 \leq X \leq 3.25$. La variable explicativa X toma cinco valores igualmente espaciados, en cada uno de ellos se realiza el mismo número de replicaciones hasta alcanzar los tamaños de muestra considerados en el estudio. $z \sim DVE$, y esta se genera mediante $\log(-\log(U(0,1)))$.

La simulación

Cada Ensayo Monte Carlo consiste en:

1. Generar los datos de Y , para los 3^4 tratamientos.
2. Obtener el estimador de $\theta = (\beta_0, \beta_1, \alpha_0, \alpha_1)$ en cada tratamiento se realizaron 500 réplicas.

3. Calcular el modelo estimado \hat{Y} , y la varianza de \hat{Y} . Para un valor fijo de X y una probabilidad de z.

5. ANÁLISIS DE RESULTADOS

En un breve resumen, se describe en la tabla 1 el análisis de la varianza realizado para la respuesta $(Y_{obs} - \hat{Y}) \sim N(0, \text{Var}(Y - \hat{Y}))$. En este trabajo no se dan todos los detalles estadísticos de \hat{Y} , ni de los estimadores de θ .

Tabla 1
Resultados del experimento

variación	SC	gl	CM	F	p
A	15.070	2	7.535	18.133	0.000
B	0.533	2	0.266	0.642	0.053
C	4.536	2	2.268	5.460	0.010
D	1.070	2	0.535	1.290	0.300
AC	4.348	4	1.087	2.62	0.040
error	28.258	68	0.416		
total	53.815	80			

Existe un efecto de interacción entre la censura y variabilidad, cabe mencionar que cuando no hay censura (nivel bajo de censura) la diferencia $Y_{obs} - \hat{Y}$ es muy cercana a cero en todos los niveles de variabilidad. Se da una diferencia grande cuando hay mayor censura y mayor variabilidad. El factor D que representa a los tratamientos resultó no significativo, sin embargo, se puede resaltar que el procedimiento PBA tiene un mejor comportamiento, el procedimiento PPV tiene una tendencia positiva en la diferencia y contrariamente PON tiene una tendencia negativa en $Y_{obs} - \hat{Y}$.

La programación para estimar los parámetros del modelo y el desarrollo de la estrategia de simulación se realizó usando el sistema GAUSS versión 3.14, cuyo conjunto de procedimientos, funciones y generador de números aleatorios están altamente probados para asegurar eficiencia y exactitud en los resultados numéricos.

REFERENCIAS

- Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.
- Davidan, M. and Carroll, R.J. (1987). Variance Function Estimation. *JASA*, **82**, 1079-1091.
- Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley and Sons.
- Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley and Sons.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. 2nd. ed. New York: Chapman and Hall.

Estimación de la Tasa de Mortalidad Infantil en México por Entidad Federativa y Municipio, con Información Censal de 1990

VÍCTOR GARCÍA, OSCAR MUÑIZ y CRUZ ALBERTO RODRÍGUEZ

INEGI

1. INTRODUCCIÓN

El nivel de la mortalidad infantil está asociado con las condiciones de vida de la población, en virtud de que los menores de un año son altamente sensibles a las características ambientales, tanto climáticas como de alimentación, higiene y sanidad general, entre otras. Estas condiciones de vida reflejan el grado en que se satisfacen necesidades sociales básicas, lo cual influye en la tendencia de la mortalidad infantil. Para su estudio se toma como fuente directa el registro de los nacimientos y las defunciones de menores de un año, sin embargo este registro puede ser afectado por situaciones de diversa índole en las que operan los organismos coordinadores del registro civil en los países, que impactan en la cobertura tanto de las defunciones como de los nacimientos, lo cual provoca que el conocimiento que se tenga de la mortalidad infantil pueda ser poco confiable, más aún, si se pretende detallar el desglose por subpoblaciones.

Ante esta situación se han realizado esfuerzos para obtener datos complementarios de la mortalidad infantil, como son la realización de encuestas y el uso de información censal; de éstos, las encuestas permiten un mejor acercamiento por su mayor nivel de detalle temático, sin embargo su representatividad generalmente dificulta obtener estimaciones con amplios detalles geográficos; en el caso de la información censal, se tiene la ventaja de su universalidad, lo cual permite obtener estimaciones indirectas de la mortalidad infantil para subpoblaciones con pequeños volúmenes de habitantes.

En esta última línea de investigación se encuentra el trabajo que aquí se presenta, donde se aplica un procedimiento que propone William Brass (1974), con el cual es posible estimar la tasa de mortalidad infantil (TMI). Para estimar la TMI en México, el procedimiento se aplicó con una variante desarrollada por Michael Hartmann (1991), en el cual se introducen modelos paramétricos para la fecundidad y la mortalidad en los primeros años de vida, lo que permite ajustar las curvas de mortalidad a un patrón seleccionado, como tablas modelo, tablas con datos de la población del país o región, de la población urbana, etc.

Además del ajuste de las curvas de mortalidad a un patrón seleccionado, uno de los aspectos destacados del procedimiento utilizado es su aplicabilidad a poblaciones de tamaño reducido, que permite superar algunas limitaciones propias del sistema de estadísticas vitales en México (y del mismo procedimiento de Brass y sus variantes más utilizadas) y efectuar estimaciones de la mortalidad infantil a nivel municipal, lo cual facilita la identificación de zonas geográficas con distinto nivel de mortalidad infantil y en consecuencia, con diferentes condiciones de vida.

Los datos básicos utilizados son el total de mujeres de 15 a 29 años por edad desplegada y el total de nacidos vivos y sobrevivientes declarados por las mismas en el Censo de 1990. Las estimaciones se obtuvieron a nivel nacional y por entidad federativa y municipio (INEGI, 1996).

2. DESCRIPCIÓN DEL PROCEDIMIENTO

Este procedimiento se basa en la introducción de modelos paramétricos para la fecundidad y mortalidad. Hartmann propone para la mortalidad la función $q(z;a,b)=1-1/(1+z^{2b}e^{2a})$, donde a y b son los parámetros del modelo y $q(z;a,b)$ representa la probabilidad que tiene un recién nacido de no llegar con vida a la edad z , y para la fecundidad, la función $f(x;u)=c^{18}x^{17}e^{-cx}/17!$, en donde $c=18/u$, u es la edad media de la fecundidad y $f(x;u)\Delta x$ es la probabilidad que tiene una mujer de dar a luz a un niño en el intervalo de tiempo x a $x+\Delta x$ por unidad de longitud Δx .

Con estos dos modelos se obtienen una expresión para el número total esperado de hijos fallecidos de las mujeres de edad x , en el municipio j de la entidad i , $W_{ij}(x)R\int_{\alpha}^x f(y;u_i)q(x-y;a,b)dy$, y del mismo modo, una expresión para el número total de hijos nacidos vivos de las mujeres de edad x , $W_{ij}(x)R\int_{\alpha}^x f(y;u_i)dy$, donde α es la edad inicio del periodo reproductivo, R es la tasa global de fecundidad y $W_{ij}(x)$ el número de mujeres el municipio j de la entidad i .

Así, la proporción derivada de la estructura de la mortalidad y del patrón de fecundidad está dada por $H_i(x;u_i,a,b)=(\int_{\alpha}^x f(y;u_i)q(x-y;a,b)dy)/(\int_{\alpha}^x f(y;u_i)dy)$. Estas proporciones ($H_i(x;u_i,a,b)$) corresponden a una población con la fecundidad y la estructura de la mortalidad de los modelos.

Por consiguiente, para determinar la relación entre el nivel de la mortalidad del modelo y el de la población, asumiendo que la estructura de la mortalidad y el patrón de la fecundidad son, respectivamente, los determinados por el modelo oeste de Coale y Demeny y la derivada a nivel entidad federativa con estadísticas vitales, se calculó un factor de ajuste mínimo cuadrático, $F_{ij}=\sum_{x=\alpha}^{29} ((H_i(x;u_i,a,b)Q_{ij}(x))/(H_i^2(x;u_i,a,b)))$, donde $H_i(x;u_i,a,b)$ es la proporción esperada de hijos fallecidos obtenida con el modelo oeste de Coale y Demeny ($a=-1.505046$ y $b=0.075653$) y la estructura de la fecundidad de la entidad i y $Q_{ij}(x)$ es la proporción observada de hijos fallecidos en las mujeres de edad x , en el municipio j de la entidad i .

Con el factor de ajuste y las proporciones esperadas, $H_i(x;u_i,a,b)$, se obtienen las proporciones suavizadas de hijos fallecidos en las mujeres de edad x para el municipio j de la entidad i . Así entonces, estas proporciones suavizadas reflejan la estructura de los modelos de mortalidad y fecundidad y el nivel de la mortalidad de las proporciones observadas.

El teorema generalizado del valor medio para las integrales justifica que $H_i(x;u_i,a,b)=q(z(x);a,b)$; esto es, que la proporción esperada de hijos fallecidos en las

mujeres de edad x es igual a la probabilidad que tiene un recién nacido de morir antes de cumplir una cierta edad “ $z(x)$ ”.

Como consecuencia, se tiene que, $\hat{Q}_{ij}(x_0; u_i, a, b) = F_{ij}H_i(x_0; u_i, a, b) = F_{ij}q(z(x_0); a, b)$, para $z(x_0)=1$, es la estimación de la Tasa de Mortalidad infantil para el municipio j de la entidad i . Esta estimación es el resultado de la estructura de los modelos de fecundidad (de la entidad i) y mortalidad (Tabla modelo oeste de Coale y Demeny, nivel 20), así como el nivel de la mortalidad de las proporciones observadas del municipio j de la entidad y .

Este procedimiento tiene dos características especiales: uno, al parametrizar las funciones de fecundidad y mortalidad, se suaviza el comportamiento de las proporciones de hijos fallecidos observadas en la población, lo cual ofrece mayor confiabilidad en las estimaciones para las poblaciones en las que, por su tamaño, los valores observados presentan fluctuaciones de cierta importancia; dos, para ajustar los datos a funciones específicas, es posible utilizar como patrones de mortalidad y fecundidad diversas opciones, tales como: tablas modelo, tablas con datos de la población del país o región, de la población urbana, etc.

3. RESULTADOS

Hacia 1990 la TMI estimada para México es de 40 defunciones de menores de un año por cada mil nacidos vivos, en tanto que en el mundo variaba entre cuatro defunciones por cada mil nacidos vivos en Japón y Finlandia y 134 en Malawi (Population Reference Bureau, 1994). En América Latina, los valores extremos se presentaban en Cuba (11 por mil) y Haití (101) (CELADE, 1993).

Por entidad federativa la mortalidad infantil varía desde una tasa de 24 por mil, hasta una de 56. Con las TMI obtenidas, las entidades federativas se agruparon en siete estratos y tres grupos de riesgo para la sobrevida infantil (ver Tabla 1), con el fin de visualizar las diferencias por regiones, donde destaca el estrato uno, formado por Chiapas, Puebla, Guerrero y Oaxaca por presentar las TMI más altas, además de compartir fronteras estatales, de tal manera que estos dos factores la convierten en la región del país que concentra los niveles más altos de mortalidad infantil. Por otro lado destacan Nuevo León, Baja California, Baja California Sur, Distrito Federal y Tamaulipas, por representar el estrato con las más bajas TMI.

Por municipio la mortalidad infantil varía desde 11 defunciones por mil nacidos vivos, hasta 152. Al igual que la estratificación de las entidades federativas, también se agrupó a los municipios de acuerdo a las TMI obtenidas y se clasificaron en siete estratos, donde se distinguen tres grupos de riesgo para la sobrevida infantil (ver Figura 1): el primero es el grupo de riesgo alto e incluye los estratos del 1 al 5; el segundo es el grupo de riesgo medio y lo integra el estrato 6; y el tercero es el grupo de riesgo bajo conformado por el estrato 7, el cual concentra la información de los menores de un año en los municipios con las TMI más bajas del país. De la población menor de un año, el 70% se encuentra en municipios con condiciones de riesgo bajo y medio para la sobrevida infantil, donde la TMI se presenta en un intervalo de 11.45 a 43.84 defunciones por mil. El resto de menores de un año, el 30%, se ubica en municipios de riesgo alto, donde la TMI va de 43.85 a 152.

TABLA 1
ESTADOS UNIDOS MEXICANOS. RESULTADOS POR ENTIDAD FEDERATIVA

*TASA DE MORTALIDAD INFONATAL, TOTAL DE NACIMIENTOS, SOBREMORTALIDAD, ESTRATO, RIESGO DE MUERTE Y
 LUGAR NACIONAL SEGUN ENTIDAD FEDERATIVA
 (Ordenadas de mayor a menor según la TMI)*

ENTIDAD FEDERATIVA	TASA DE MORTALIDAD INFANTIL	TOTAL DE NACIMIENTOS 1/	SOBRE- MORTALIDAD 2/	ESTRATO 3/	RIESGO DE MUERTE 4/	LUGAR NACIONAL 5/
ESTADOS UNIDOS MEXICANOS	40	2,393,540	-	-	-	-
CHIAPAS	56	130,218	2.34	1	ALTO	21
PUEBLA	56	128,296	2.32	1	ALTO	21
GUERRERO	55	89,428	2.30	1	ALTO	20
OAXACA	52	97,071	2.15	1	ALTO	19
ZACATECAS	48	36,027	2.01	2	ALTO	18
VERACRUZ	46	196,538	1.91	3	ALTO	17
DURANGO	45	41,457	1.87	3	ALTO	16
TLAXCALA	45	23,441	1.86	3	ALTO	16
GUANAJUATO	44	113,766	1.85	3	ALTO	15
MICHOACÁN	44	102,015	1.83	3	ALTO	15
HIDALGO	43	61,392	1.80	3	ALTO	14
NAYARIT	43	23,720	1.77	3	ALTO	14
QUERÉTARO	43	32,453	1.78	3	ALTO	14
SAN LUIS POTOSÍ	43	59,075	1.79	3	ALTO	14
TABASCO	40	56,190	1.68	4	MEDIO	13
AGUASCALIENTES	38	19,960	1.59	4	MEDIO	12
MÉXICO	38	298,867	1.59	4	MEDIO	12
CHIHUAHUA	37	71,286	1.56	4	MEDIO	11
JALISCO	36	135,899	1.49	5	BAJO	10
COLIMA	35	12,128	1.47	5	BAJO	9
YUCATÁN	35	38,163	1.44	5	BAJO	9
CAMPECHE	34	18,359	1.42	5	BAJO	8
MORELOS	33	34,898	1.39	5	BAJO	7
SINALOA	31	62,236	1.30	6	BAJO	6
QUINTANA ROO	30	18,423	1.25	6	BAJO	5
COAHUILA	29	56,955	1.22	6	BAJO	4
SONORA	29	49,286	1.21	6	BAJO	4
TAMAULIPAS	27	61,271	1.11	7	BAJO	3
BAJA CALIFORNIA SUR	26	9,517	1.07	7	BAJO	2
DISTRITO FEDERAL	26	190,992	1.09	7	BAJO	2
BAJA CALIFORNIA	24	48,712	1.02	7	BAJO	1
NUEVO LEÓN	24	75,501	1.00	7	BAJO	1

1/ Datos estimados.

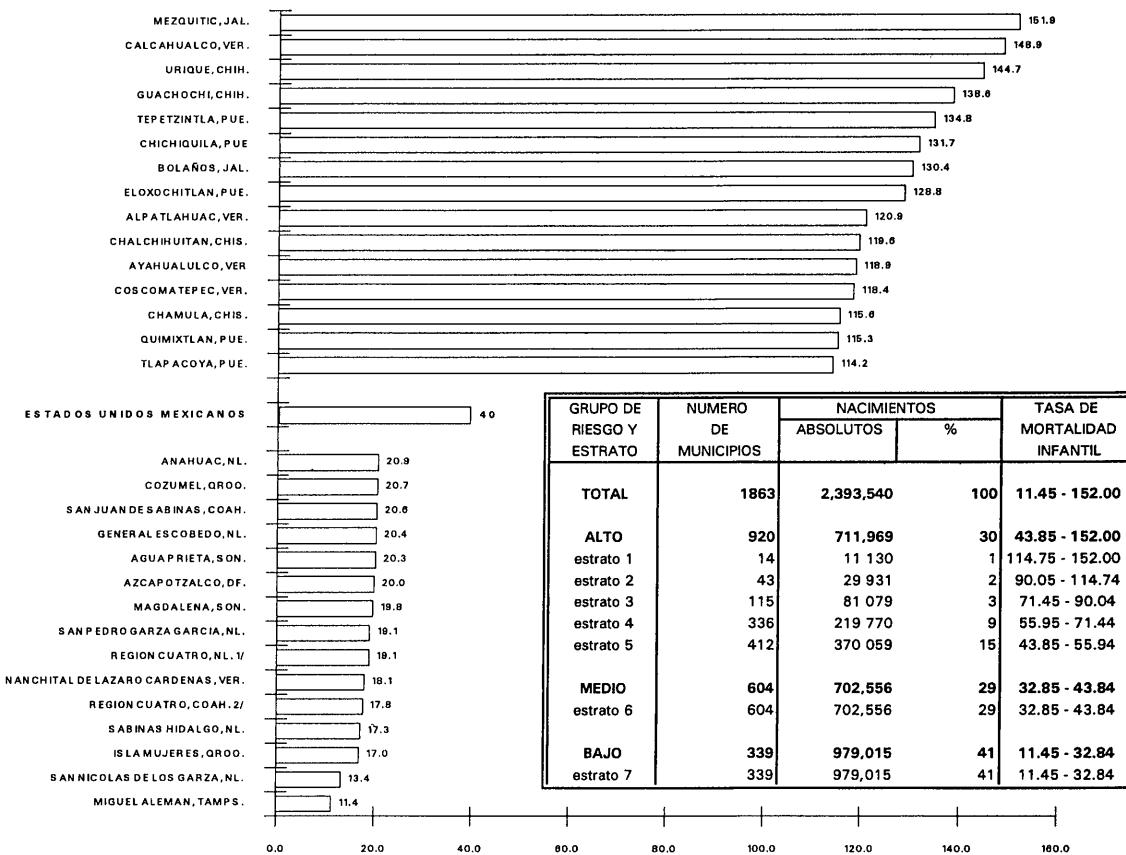
2/ Exceso proporcional de la mortalidad infantil en cada entidad federativa, respecto a la de más baja mortalidad en el país, en este caso Nuevo León.

3/ Con base en siete estratos en que se agruparon las entidades federativas, donde el uno es el de más alta mortalidad y el siete el de más baja.

4/ Al que se encuentra expuesta la población infantil, con base en la clasificación de los estratos en tres grupos.

5/ Ordenados de menor a mayor, donde la entidad federativa con menor mortalidad ocupa el primer lugar.

Figura 1. ESTADOS UNIDOS MEXICANOS. RESULTADOS POR MUNICIPIO
MUNICIPIOS DEL PAÍS CLASIFICADOS EN GRUPOS DE RIESGO Y ESTRATO SEGÚN
SU TMI Y MUNICIPIOS CON LOS 15 VALORES MÁXIMOS Y MÍNIMOS



1/ Incluye los municipios de Cerralvo, Ciénega de Flores, China e Hidalgo.

2/ Incluye los municipios de Morelos y Zaragoza.

NOTA: Para 1990 el país contaba con 2403 municipios, de los cuales 570 corresponden a la entidad federativa de Oaxaca. En el presente trabajo las TMI en la entidad se estimaron para sus 30 distritos que agrupan esos 570 municipios. Por ello el total nacional suma 1863 y no 2403.

REFERENCIAS

- Brass, William (1974). *Métodos para estimar la fecundidad y la mortalidad en poblaciones con datos limitados*. Santiago de Chile: CELADE.
- CELADE-UNICEF. *Mortalidad en la Niñez. Una base de datos desde 1960. América Latina*. Agosto de 1993.
- Hartmann, Michael (1991). A Parametric Method for Census Based Estimation of Child Mortality. *Journal of Official Statistics. Statistics Sweden*, 7, 45-55.
- INEGI (1996). *La mortalidad infantil en México: Estimaciones por entidad federativa y municipio*. México: INEGI.
- Population Reference Bureau, Inc. Cuadro de la Población Mundial. Abril de 1994.

Algunos Modelos para Mediciones Repetidas Discretas en Estudios Longitudinales

LETICIA GRACIA MEDRANO VALDELAMAR

IIMAS -UNAM

1. INTRODUCCIÓN

Las mediciones repetidas son observaciones de una misma característica que se hacen varias veces, distinguiéndolas de otras observaciones: 1) el que la misma variable se observa en una unidad más de una vez; 2) que éstas no son independientes como en el análisis de regresión; y 3) el que más de una unidad está involucrada en el estudio, formando algo más complejo que una serie de tiempo.

Los datos categóricos o de conteo registrados a lo largo del tiempo pueden ser modelados de muchas formas . En este documento sólo se habla de tres modelos que ayudan a analizar este tipo de datos. Estos modelos se eligieron porque son muy ilustrativos y porque puede utilizarse para su desarrollo paquetería convencional como lo son GLIM y S-Plus.

2. CADENAS DE MARKOV

Suponiendo que la variable respuesta (categórica) de un individuo en un momento dado depende sólo del evento inmediato anterior, es decir que estamos en los supuestos de una cadena de Markov de primer orden, se tiene entonces una matriz cuadrada T_t de probabilidades condicionales de eventos que pueden estar variando con el tiempo. Si los renglones son los diferentes estados (categorías) al tiempo t y las columnas son los estados al tiempo $t+1$, entonces los renglones suman 1. Al premultiplicar el vector n_t de frecuencias de unidades en los diferentes estados al tiempo t por la matrix T_t transpuesta se tendrá el vector para el siguiente periodo $t+1$, esto es : $n_{t+1} = T^* n_t$, donde T representa el patrón de cambio.

Teniendo datos categóricos longitudinales este modelo se puede construir utilizando modelos loglineales o logísticos para estimar esas probabilidades condicionales de la matriz T . En la mayoría de las situaciones interesa conocer algunas características de la cadena como son el orden de la cadena, la existencia de estacionariedad, de reversibilidad o de equilibrio.

Por ejemplo, el orden de la cadena estará dado por el orden de interacción entre los factores tiempo del modelo loglineal que se ajuste mejor a los datos. La estacionariedad, se verifica a través de una prueba de independencia entre el estado observado al final del periodo y el factor que caracteriza al tiempo. La reversibilidad se

refiere a que la probabilidad condicional entre eventos es la misma en ambas direcciones, es decir, si existe cuasisimetría en la tabla. Si las marginales no cambian con el tiempo, se dice que el proceso descrito por la tabla de contingencia está en equilibrio, es decir, la distribución marginal es estacionaria. En tablas de contingencia esto corresponde a probar homogeneidad marginal.

Ejemplo. La tabla siguiente fue tomada de Lindsey (1993).

Lluvia en el mes de junio en Madison Wisconsin año

1961	10000	01101	01100	00010	01010	00000
1962	00110	00101	10000	01100	01000	00000
1963	00001	01110	00100	00010	00000	11000
1964	01000	00000	11011	01000	11000	00000
1965	10001	10000	00000	00001	01100	01000
1966	01100	11010	11001	00001	00000	11100
1967	00000	11011	11101	11010	00010	00110
1968	10000	00011	10011	00100	10111	11011
1969	11010	11000	11000	01100	00001	11010
1970	11000	00000	01000	11001	00000	10000
1971	10000	01000	10000	00111	01010	00000

Se codificó 1 si llovía y 0 si no, y las unidades son los meses de junio. Para encontrar las probabilidades condicionales de T se ajustaron algunos modelos logísticos con S-Plus.

Si p es la probabilidad de que llueva, para empezar se ajustó el siguiente modelo:

$$\log \left\{ \frac{p}{1-p} \right\} = -0.74055,$$

el cual tiene una devianza de 401.3175 con 318 grados de libertad.

A continuación se ajustó un modelo que permite modelar una cadena de Markov de primer orden. Si p_1 es la probabilidad de que llueva dado que llovió el día anterior y p_2 la probabilidad de que llueva dado que el día anterior no llovió, ajustando el siguiente modelo se pueden encontrar esas probabilidades,

$$\log \left\{ \frac{p_i}{1-p_i} \right\} = 0.9463 + 0.5716 x_i,$$

donde x_i es un factor que indica si el día anterior llovió o no. Este modelo tiene una devianza de 396.0697 con 317 grados de libertad; la diferencia en devianza respecto al modelo anterior es 5.25. Usando este modelo se tiene que $p_1=0.4074$ y $p_2=0.2796$.

Entonces la matriz de transición es:

$$T = \begin{bmatrix} .721 & .279 \\ .603 & .407 \end{bmatrix}$$

Para saber si la cadena es o no estacionaria se introduce el factor año en el modelo logístico anterior, para que indique si esas probabilidades varían con los años. Los coeficientes para este modelo son:

Intersepto	x_i	año1	año2	año3	año4
-0.9225672	0.417281	-0.1755184	0.00304676	0.0468616	0.02811695

año5	año6	año7	año8	año9	año10
0.01874463	-0.03889833	0.08699759	0.0958042	0.07664336	0.03836212

Este modelo tiene una devianza de 384.245 con 307 grados de libertad. La diferencia de devianzas es 11.82464 con 10 grados de libertad, por lo que no se rechaza el que los coeficientes asociados a año sean cero y entonces se considera que la cadena es estacionaria. Los mismos resultados hubiesen sido obtenidos si las observaciones se clasificaran en una tabla de $2 \times 2 \times 11$.

3. MODELO PERMANENCIA Y CAMBIO

Este modelo es una modificación a la cadena de Markov, y se aplica a tablas de movilidad que son tablas cuadradas o de más dimensiones que tienen a la misma variable categórica como variable observada en dos o más puntos en el tiempo. Ejemplo de esto son: comportamiento de migrantes, movimientos de clases sociales, comportamiento de los votantes a través de elecciones, etc. Se desea saber si la posición en un momento depende de la posición previa. Para verificarlo se hace una prueba de independencia entre columnas y renglones de la tabla de movilidad. En algunos casos hay muchas unidades que no cambian de posición, de manera que tal independencia no parece aceptable, es decir, que hay demasiadas observaciones en la diagonal. Una manera de solucionar esto es eliminar la diagonal y crear un modelo de cuasi-independencia para las observaciones que están fuera de la diagonal. Teóricamente este modelo está suponiendo que la diagonal contiene dos tipos de unidades: "las que cambian" pero que no lo hicieron en el tiempo observado y "los que permanecen"; de aquí el nombre del modelo. Con el modelo de cuasi-independencia se estiman las unidades de la diagonal que potencialmente cambiarán de posición.

4. MODELOS DE APRENDIZAJE

En los modelo de cadenas de Markov los eventos dependen estocásticamente del evento ocurrido un paso o más atrás. En algunas ocasiones el evento puede modelarse mejor en función del número total de eventos de los diferentes tipos de eventos ocurridos en el pasado. Esta acumulación de eventos constituye precisamente un modelo de aprendizaje. En estos modelos se deja ver que tipo de experiencias tienen mayor influencia sobre el comportamiento actual. Si sólo se registra un tipo de experiencia se tiene un modelo de nacimientos o de contagios. Ejemplos acerca de estos dos modelos pueden verse en Lindsey (1993).

REFERENCIA

Lindsey, J. K. (1993). *Models for repeated measurements*. Oxford Science Publications. Oxford.

Métodos Multivariados en la Taxonomía del Grupo *Oocarpa*

PORFIRIO GUTIÉRREZ GONZÁLEZ y JORGE A. PÉREZ DE LA ROSA

Universidad de Guadalajara

1. INTRODUCCIÓN

México es considerado como un centro secundario de origen y dispersión del género *Pinus*, debido a la gran diversidad que presentan estos árboles en nuestro país. Casi la mitad de las especies que se reconocen para este género se localizan en México. Esta variación es debida a la diversidad de climas, suelos y aislamiento geográfico. Todos estos factores interactúan y generan especies, variedades y formas, que en muchos casos constituyen grandes problemas taxonómicos. Uno de estos problemas es el que constituye *Pinus oocarpa* Schiede ex Schltdl, y sus variedades.

El *P. oocarpa* fue descrito por Schiede en el trabajo de Schlechtendal y fue hasta 1909 cuando se describió la primera variedad *P. oocarpa* var. *microphylla* Shaw. Posteriormente en la década de los cuarenta, Maximino Martínez auxiliado de más material proveniente de toda el área de distribución, describió tres variedades más: *P. oocarpa* var. *trifoliata* (publicada en 1945 como forma *trifoliata*), *P. oocarpa* var. *ochoterenae*. Mientras que Martínez (1948) define el Grupo *Oocarpa* como pinos que tienen cono simétrico u ovoide, con escamas liradas, de color ocre o rojizo. El pedúnculo es débil, largo y delgado. Posteriormente, Styles (1976), menciona que le nombre *ochoterenae* esta incorrectamente utilizado como variedad de *P. oocarpa*, ya que en base a los resultados de un muestreo hecho en Chiapas, debería ser considerado como variedad de *P. patula*. En seguida, Styles y McVaugh (1990), afirman que *P. oocarpa* var. *microphylla*, es tan diferente a la variedad *oocarpa* que lo describen como *Pinus praetermissa*. McVaugh (1992) menciona que la variedad *trifoliata* es muy escasa y guarda gran similitud con la variedad *oocarpa*.

Por lo anterior, esta taxa forma un complejo el cual no se encuentra aún bien delimitado y crea confusión al tratar de comprender sus variaciones morfológicas y geográficas, con lo que genera la necesidad de realizar un estudio detallado en toda su área de distribución en México con el auxilio de las técnicas multivariadas como son: Componentes Principales, Análisis Cluster, y Análisis Discriminante.

2. MATERIALES Y MÉTODOS

Con el propósito de comprobar la naturalidad del Grupo *Oocarpa* (*fide* Martínez, 1948), se incluyeron en el análisis muestras del *P. oocarpa* var. *oocarpa*, *P. oocarpa* var. *ochoterenae*, *P. oocarpa* var. *trifoliata*, y *P. oocarpa* var. *microphylla*.

Trabajo de Campo

Se realizaron excusiones de colecta y toma de datos de campo en algunos sitios seleccionados de la distribución conocida del Grupo *Oocarpa* en México. En cada localidad seleccionada se colectaron muestras de 25 árboles.

Cada muestra consiste en una ramilla de aproximadamente 30 cm de longitud, que contiene acículas, conos maduros y microsporófilas masculinas. Además se incluyeron las características convencionales de rama, acículas, cono y semillas, así como la de las microsporófilas masculinas, ya que las estructuras fundamentales de los conos masculinos tienen una forma de organización más simple que los conos femeninos, por lo que se consideran más primitivos.

Trabajo de laboratorio

Las variables seleccionadas y analizadas fueron las siguientes:

R1: Ramilla lisa ó escamosa; R2: Diámetro de la ramilla en la inserción del cono (mm); A1: Número de acículas por fascículo; A2: Largo de bráctea axilar del fascículo (mm); A3: Ancho de la bráctea axilar del fascículo (mm); A5: Largo de las acículas (mm); A6: Ancho de las acículas (mm); A8: Largo de la vaina en fascículos jóvenes (mm); A9: Largo de la vaina en fascículos maduros (mm); A10: Número de dientes marginales en 5 mm de la parte media de las acículas; A11: Número de hileras de estomas en la cara dorsal de la acícula; A12: Número de hileras de estomas en las caras internas de las acículas; A14: Penetración de hipodormo en el mesófilo:

1. delgado y uniforme; 2. con entrantes en el clorénquima; 3. con entrantes hasta la endodermis; Cantidad de canales resiníferos: A15: Externos, A16: Internos, A17: Medios; A18: Septales; C1: Longitud del pedúnculo (cm); C2: Diámetro del pedúnculo (cm); C4: Longitud del cono cerrado (cm); C5: Diámetro del cono cerrado (cm); C6: Índice; longitud / diámetro del cono cerrado; E1: Escama basal caediza ó persistente; E2: Longitud de las escamas de la parte media del cono (mm); E3: Longitud de la apófisis (mm); E4: Ancho de la apófisis (mm); E5: Índice; longitud de la escama/longitud de la apófisis; S1: Longitud de la semilla (mm); S2: Ancho de la semilla (mm); S3: Índice; longitud/ancho de la semilla; S4: Grueso de la semilla (mm); S5: Longitud del ala de la semilla (mm); S6: Ancho de la semilla (mm); S8: Índice; longitud del ala / longitud de la semilla; M1: Longitud de la microsporófila (mm); M2: Diámetro de la microsporófila (mm); M4: Cantidad de brácteas báslas de la microsporófila; M5: Longitud de la mayor bráctea de la microsporófila (mm); M6: Ancho de la mayor bráctea de la microsporófila (mm).

3. RESULTADOS

Análisis de Componentes Principales

Primero, se inició un Análisis de Componentes Principales con el objetivo de identificar las variables de mayor importancia entre las cuatro especies. En este estudio se analizaron 38 variables en 393 árboles de los cuales 224 corresponden al *P. oocarpa* var. *oocarpa*, 43 del *P. oocarpa* var. *ochoterenae*, 50 del *P. oocarpa* var. *microphylla*, y 50 del *P. oocarpa* var. *trifoliata*.

Del análisis se obtuvieron 5 componentes principales, con los cuales se explica el 75.18% de la variabilidad total de los datos originales (tabla 1).

TABLA 1. Valores propios y porcentaje de varianza de los primeros cinco componentes principales.

Componente Principal	Valor Propio	% de varianza	% de varianza acumulada
1	5.31	29.50	29.50
2	3.86	18.81	48.13
3	2.19	12.18	60.49
4	1.54	8.56	69.09
5	1.10	6.12	75.18

Las variables de mayor significancia de los tres primeros componentes principales son:

R1: Ramilla lisa ó escamosa; A1: Número de acículas por fascículo; A2: Largo de bráctea axilar del fascículo (mm); A14: Penetración de hipodermo en el mesófilo: (1) delgado y uniforme, (2). con entrantes en el clorénquima, (3). con entrantes hasta la endodermis; Cantidad de canales resiníferos: A15: Externos, A17: Medios; C6: Índice; longitud/diámetro de cono cerrado; E1: Escama basal caediza ó persistente; S8: Índice; longitud del ala/longitud. de la semilla; M1: Longitud de la microsporófila (mm); M4: Cantidad de brácteas basales de la microsporófila.

En la figura 1 se observan los tres primeros componentes principales. Nótese la identificación de cuatro grupos de especies.

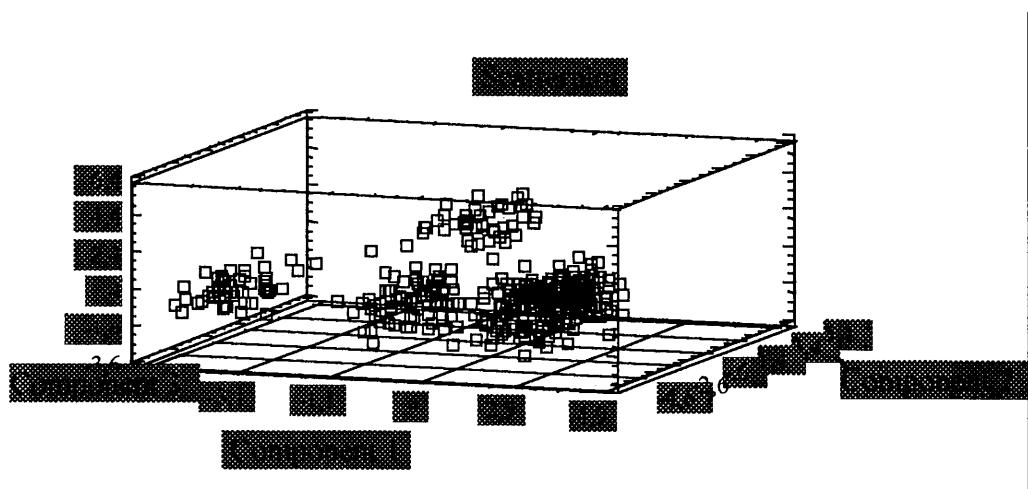


Fig. 1. Los tres primeros componentes principales

Análisis de Cluster

Se hizo un Análisis de Cluster (método de encadenamiento completo) con el objetivo de encontrar una clasificación que muestre la relación de las cuatro especies.

El resultado se puede resumir con el Dendograma de la figura 2. Nótese y compare con la figura 1 en la identificación de los grupos, se puede ver que en el primer grupo los más parecidos son el *P. oocarpa* var. *oocarpa* y el *P. oocarpa* var. *trifoliata*; un segundo grupo lo forma el *P. oocarpa* var. *ochoterenae*; y como tercer grupo se tiene al *P. oocarpa* var. *microphylla*.

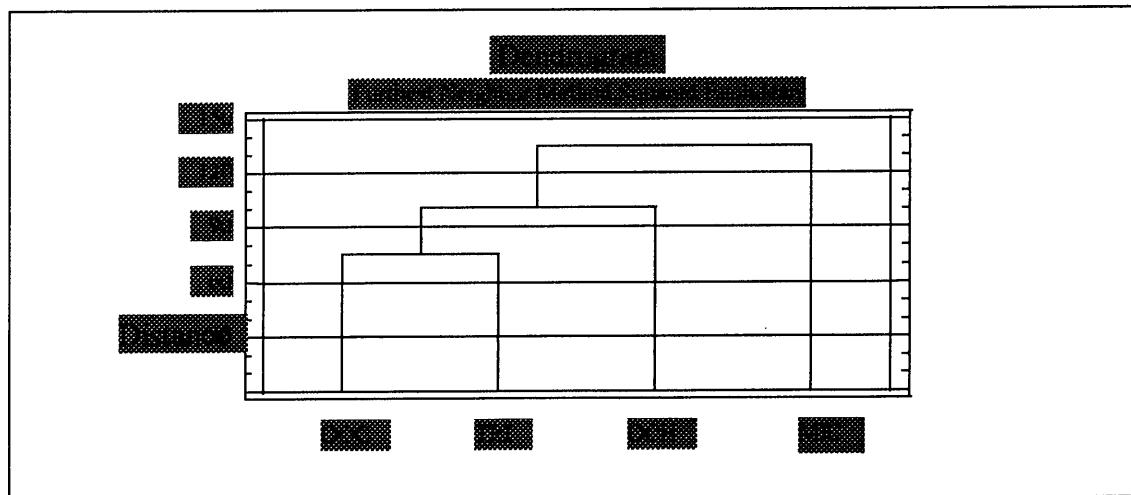


Fig. 2. Dendograma

Análisis Discriminante

Con el propósito de confirmar lo anterior; se hizo un Análisis Discriminante para los cuatro grupos. Considerando como primer grupo al *P. oocarpa* var. *oocarpa*, como segundo grupo al *P. oocarpa* var. *trifoliata*, como un tercer grupo al *P. oocarpa* var. *ochoterenae* y un cuarto grupo al *P. oocarpa* var. *microphylla*.

De acuerdo al análisis se encontró dos funciones discriminantes significativas con una correlación de 0.99 con las cuales se explica el 100% de la variabilidad de los datos. En seguida con el procedimiento stepwise (método de forward), encontramos que las variables significativas para la separación de los grupos son:

A1: Número de acículas por fascículo; A3: Ancho (en la base) de la bráctea axilar del fascículo (mm); A4: Índice; largo/ancho de la bráctea. A5: Largo de las acículas; A9: Largo de la vaina en acículas maduras; A13: Índice; hileras de estomas en la cara dorsal/hileras de estomas en las caras internas; A14: Penetración de hipodormo en el mesófilo: (1) delgado y uniforme, (2). con entrantes en el clorénquima, (3). con entrantes hasta la endodermis; Cantidad de canales resiníferos: A15: Externos; A16: Internos; A17: Medios; C4: Longitud del cono cerrado; C5: Diámetro del cono cerrado; C6: Índice; longitud/diámetro del cono cerrado; E2: Longitud de la escama de la parte media del cono; E9: Índice; longitud del cono/longitud de la escama; S4: Grueso de la semilla; M1: Longitud de la microsporófila (mm); M4: Cantidad de brácteas basales

de la microsporófila; M6: Ancho de la mayor bráctea basal de la microsporófila; M7: Índice; longitud/ancho de la mayor bráctea basal de la microsporófila.

La gráfica de las dos funciones discriminantes se puede observar en la figura 3. Nótese la identificación de los tres grupos. El grupo marcado como No. 1 corresponde al *P. oocarpa* var. *oocarpa* y al *P. oocarpa* var. *trifoliata*; el grupo marcado como No. 2 corresponde al *P. oocarpa* var. *ochoterenae*; el grupo marcado como No. 3 corresponde al *P. oocarpa* var. *microphylla*.

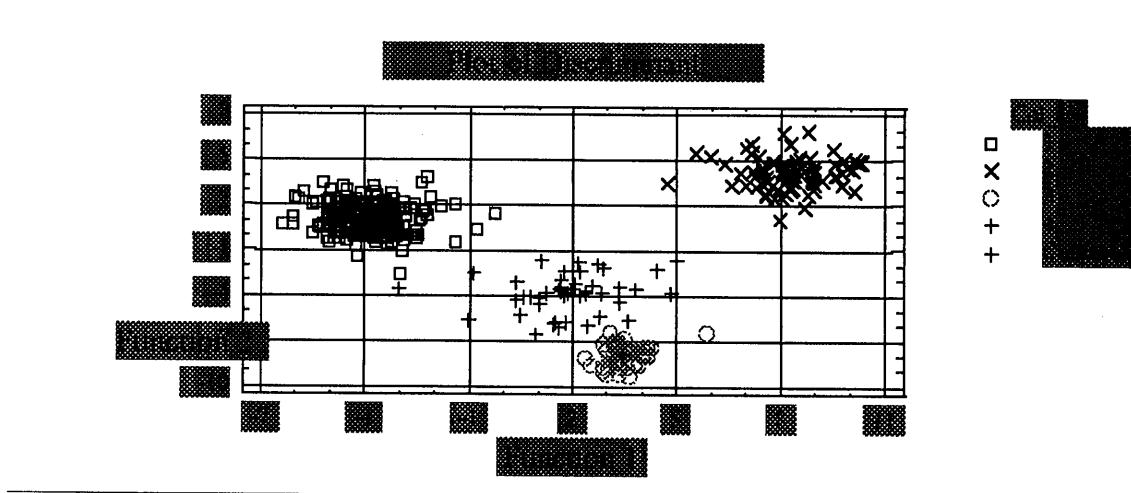


Fig. 3. Funciones discriminantes

4. CONCLUSIONES

De acuerdo a los resultados obtenidos es posible concluir que las variedades *ochoterenae* y *microphylla* son muy diferentes a la variedad *oocarpa* por lo que son consideradas como especies diferentes en la actualidad: *Pinus tecumumanii* y *Pinus praetermissa* respectivamente. La variedad *trifoliata* es la más semejante a la *oocarpa*, por lo que se sugiere hacer más estudios entre estos dos últimos taxa incluyendo más especies afines y/o nuevas variables.

Los procedimientos estadísticos multivariados utilizados demuestran las bondades que tiene su uso en la solución de problemas biológicos de clasificación.

REFERENCIAS

- Martínez, M. (1948). *Los pinos mexicanos* (2da. ed.). México: Botas.
McVaugh, R. (1992). In *Flora Novo-Galiciano* (W. Anderson, Ed.), 17. Ann. Arbor.
Michigan: Univ. of Michigan Press.
Styles, B. T. (1976). Studies of variation in Central American Pines Y. The identity of *Pinus oocarpa* var. *ochoterenai* Martinez. *Silva Genet.* 25, 109-118.

Caracterización de la Nueva Carta de Control \hat{p} -V para Atributos

HUMBERTO GUTIÉRREZ PULIDO OSVALDO CAMACHO CASTILLO

U. de Guadalajara

y

NANCY HERNÁNDEZ CARMONA

IBM de México, Jal.

1. INTRODUCCIÓN

A través de las tradicionales cartas de atributos (Gutiérrez, 1997), se puede detectar cambios en el promedio del proceso (variación entre subgrupos); pero no se puede detectar inconsistencias en la dispersión del proceso (variación dentro de subgrupos). Esta es una debilidad de estas cartas ya que se ha reconocido la presencia e importancia de este tipo de variación (Akao, 1970; Shindo, 1981; Quarshie y Shindo, 1994; McCullagh y Nelder, 1989; Gutiérrez y Camacho, 1996). Para atender esto Quarshie y Shindo (1994) han propuesto una nueva carta de control, basada en los V estadísticos de Potthoff y Whittinghill (1966), con la cual se puede evaluar los dos tipos de variación. El objetivo de este trabajo es ampliar los estudios de caracterización para esta nueva carta, y comparar su desempeño con la carta p.

2. MODELOS PARA LOS PROCESOS DE ATRIBUTOS

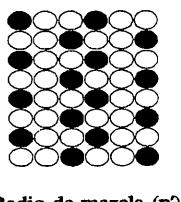
En las mezclas de sólidos hay dos características de interés: la razón de mezcla y el índice de mezcla. La primera estima la concentración o proporción de partes en la mezcla. El segundo la aleatoriedad de la distribución de las partículas. Un proceso de atributos con productos conformes y no conformes es de hecho una mezcla binaria, y en la que la razón de mezcla es conocida o fácil de conocer; en cambio para obtener información acerca del índice de mezcla se requiere un método de muestreo apropiado. Al respecto Yoshizawa y Shindo (1977) proponen el muestreo por spots; que en lugar de obtener aleatoriamente una partícula de la mezcla, se extrae aleatoriamente un grupo (spot) de partículas de la mezcla.

En la figura 1 se muestran posibles patrones de segregación de una mezcla binaria. En (c), estado completamente segregado, se da una distribución con dos puntos. En (a), estado completamente mezclado, los dos tipos de partículas se encuentran aleatoriamente en la mezcla; por lo tanto se tiene distribución binomial. En (b), mezcla incompleta, se describen mejor por la distribución beta-binomial (Quarsie y Shindo, 1994):

$$P_r[X=x] = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \frac{p^{\tau p'-1} (1-p)^{\tau (1-p')-1}}{B(\tau p', \tau (1-p'))} dp$$

Mezcla binaria

(a) Completamente mezclado



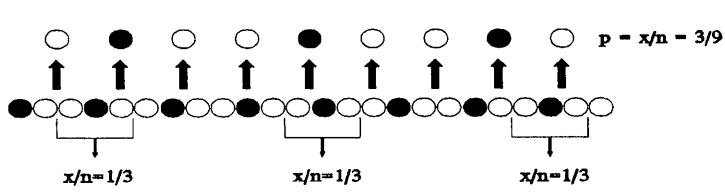
Radio de mezcla (p')

Indice de mezcla (M) = 1

Proceso de atributos

(a') Estado controlado (estable)

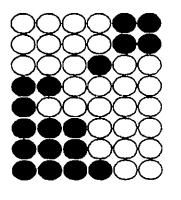
(Muestreo aleatorio simple)



$p = x/n = 3/9$

(Muestreo aleatorio por spots)

(b) Incompletamente mezclado

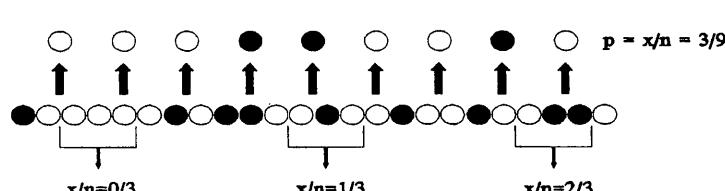


Radio de mezcla (p')

Indice de mezcla (M): $0 < M < 1$

(b') Estado fuera de control

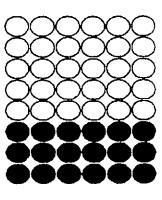
(Muestreo aleatorio simple)



$p = x/n = 3/9$

(Muestreo aleatorio por spots)

(c) Estado segregado

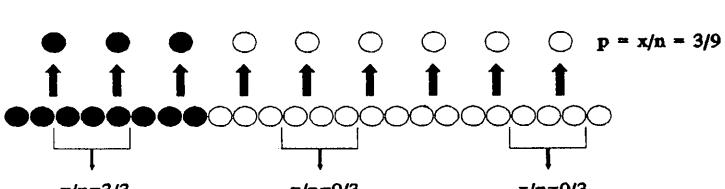


Radio de mezcla (p')

Indice de mezcla (M) = 0

(c') Extremadamente fuera de control

(Muestreo aleatorio simple)



$p = x/n = 3/9$

(Muestreo aleatorio por spots)

Figura 1. Tipos de mezclas en un proceso de atributos

Los parámetros p' y τ son positivos y determinan la razón de mezcla y el índice de mezcla, respectivamente. Además definiendo el índice, M , como $M = \frac{\tau}{\tau+1}$, se tiene que éste crece de 0 a 1 conforme la mezcla progresó del estado segregado al completamente mezclado. Para cada uno de los casos de la figura 1, con muestreo aleatorio simple sólo se puede estimar la razón de mezcla (3/9). En cambio si se toma aleatoriamente muestras por spots (3), se puede estimar tanto la razón como el índice de mezcla, y de esa forma distinguir cada caso: estado controlado (a'), estado incontrolado (b') y estado extremadamente fuera de control (c').

3. CARTA \hat{p} -V

Para estudiar el estado de mezcla del subgrupo se hace un muestreo por spots:

Subgrupo	Spots en los que se divide el subgrupo (m)						Estadísticos	
	i	1	2	...	j	...	m	\hat{p}_i
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1m}	\hat{p}_1	V_1
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2m}	\hat{p}_2	V_2
.
:	:	:	:	:	:	:	:	:
k	x_{k1}	x_{k2}	...	x_{kj}	...	x_{km}	\hat{p}_k	V_k

La proporción de unidades disconformes se estima con:

$$\hat{p}_i = \frac{\sqrt{X_{i.}}}{\sqrt{X_{i.}} + \sqrt{Y_{i.}}} \quad \text{donde } X_{i.} = \sum_{j=1}^m x_{ij} (x_{ij}-1) \quad \text{y } Y_{i.} = \sum_{j=1}^m y_{ij} (y_{ij}-1).$$

La variación entre subgrupos se detecta analizando los \hat{p}_i por medio de la carta \hat{p} . Por otra parte, bajo la hipótesis de homogeneidad de muestras binomiales vs muestras beta-binomial; la razón de verosimilitud es (Quarsie y Shindo, 1994):

$$V_i = \frac{X_{i.}}{p'} + \frac{Y_{i.}}{q'}$$

La homogeneidad del proceso cuando p' es desconocida se calcula con:

$$V_m = (\sqrt{X_{i.}} + \sqrt{Y_{i.}})^2$$

La variación dentro de subgrupos se detecta analizando los V_i por medio de la carta V.

Con base a lo anterior, las cartas \hat{p} y V, 3 sigma, se obtiene a partir de:

$$\bar{\hat{p}} = \frac{\sum_{i=1}^k \hat{p}_i}{k}, \quad \hat{\sigma}_{\hat{p}} = \sqrt{\frac{\sum_{i=1}^k (\hat{p}_i - \bar{\hat{p}})^2}{k-1}} \quad \text{y} \quad \bar{V}_m = \frac{\sum_{i=1}^k V_{m_i}}{k}, \quad \hat{\sigma}_{V_m} = \sqrt{\frac{\sum_{i=1}^k (V_{m_i} - \bar{V}_m)^2}{k-1}}$$

4. CARACTERIZACIÓN DE LA NUEVA CARTA

En esta sección se explica una parte de los estudios realizados para describir el comportamiento de la nueva carta \hat{p} -V y compararlo con el de la tradicional carta p (Gutiérrez, 1997). Los estudios de la potencia fueron hechos por simulación.

Variación entre grupos (modelo binomial). En este estudio se comparó la potencia de la carta p y la \hat{p} para detectar cambios en la variación entre subgrupos (cambios en p'). Este estudio se hizo para diferentes tamaños de subgrupo y distintas p' (un ejemplo ilustrativo de los resultados obtenidos se muestra en la figura 2). Como conclusión central del análisis de los resultados se pudo ver que la carta p y carta \hat{p} , tienen un desempeño similar bajo el supuesto de distribución binomial. Por lo que, incluso bajo estas condiciones, esta nueva carta podría reemplazar a la tradicional carta p.

Variación entre grupos (modelo beta-binomial). Este estudio fue similar al anterior, pero ahora suponiendo que el número de artículos por subgrupo sigue una distribución beta-binomial. En la figura 3 se ilustra el tipo de resultados obtenidos. Como conclusión central se pudo ver que la carta p tiene una gran cantidad de falsas alarmas (figura 3). En otras palabras, cuando la distribución de los artículos defectuosos no es homogénea en el subgrupo, la carta p es inapropiada para estudiar los cambios en la variación de p' . Por otra parte, la potencia de la carta \hat{p} es pobre, aunque si cuida el error tipo I. Este tipo de variación la registra ligeramente la carta V.

Variación dentro de grupos (modelo beta-binomial). En este estudio se comparó la potencia de la carta V, p y la \hat{p} para detectar cambios en la variación dentro de subgrupos (variación en M), bajo el supuesto de distribución beta-binomial. En la figura 4 se muestra el tipo de resultados obtenidos. Como conclusión central se pudo ver que la carta p tiene una gran cantidad de falsas alarmas. La carta V si detecta estos cambios, aunque su potencia es pobre. La carta \hat{p} prácticamente no registra estos cambios.

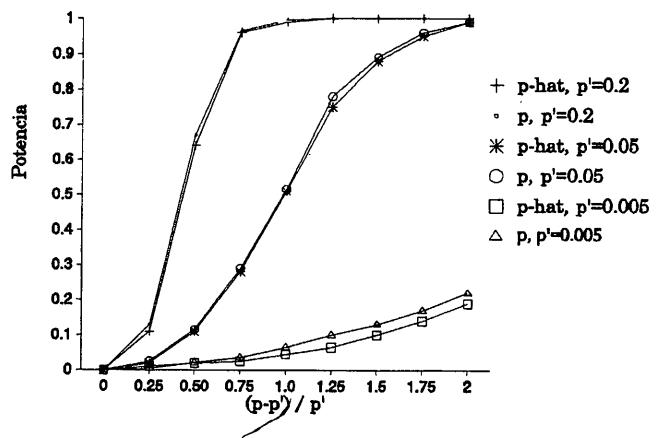


Figura 2. Comparación de la carta p y de la p-hat con $M=1$, $m=4$ y $n=50 \times 4$.

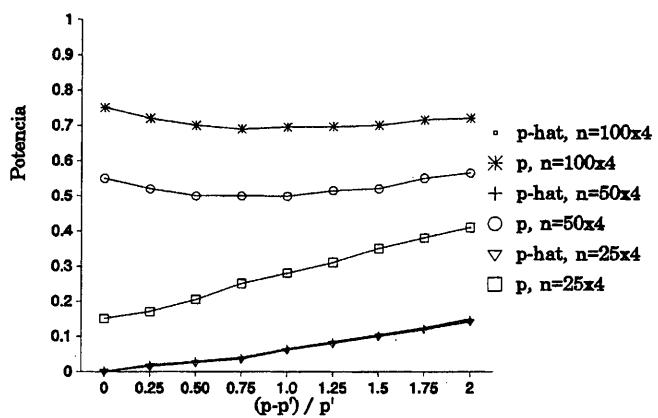


Figura 3. Comparación de la carta p y de la p-hat con $M=0.8$, $m=4$ y $p'=0.05$.

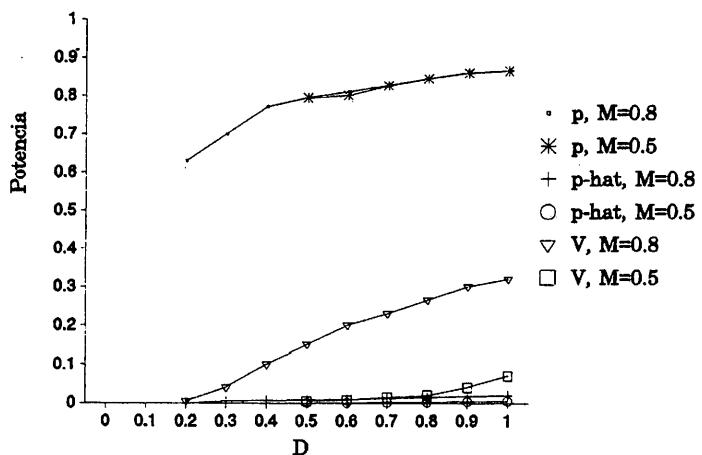


Figura 4. Comparación de la carta p, p-hat y V con $m=4$, $n=50 \times 4$ y $p'=0.2$. $D=1-M$.

REFERENCIAS

- Akao, Y. (1970). A Control Chart Based on the Mean and Range of Percent Defective by Sampling Without Replacement, *JSPC*, 21; 1646-1650.
- Gutiérrez Pulido, H. (1997). *Calidad Total y Productividad*. México: McGraw Hill.
- Gutiérrez Pulido, H. y O. Camacho Castillo (1996). Ineficiencia de la Carta p para Tamaño de Subgrupo Grande: Diagnóstico y Alternativas. *Memoria del X Foro Nacional de Estadística*, pp.152-156.
- McCullagh, P. y Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall.
- Potthoff, R.F. y Whittinghill, M. (1966). Testing for Homogeneity: the Binomial and Multinomial Distributions. *Biometrika*, 53, 167-175.
- Quarsie, L.B. y Shindo, H. (1994). A Control Chart for Atributes and its Aplications. *ASQC 49th Annual Quality Congress Proceeding*, 378-388.
- Shindo, H. (1981). A Study on the \bar{p} - R_p Chart Based on Survey Data of a Process, *JSPC*: 32, 747-753.
- Yoshizawa, T. y Shindo, H. (1977). Description of a Mixed State and the Beta-Binomial Distribution, *Hinshitsu*, *JSPC*: 7(3).

Empleo de Técnicas Computacionales Modernas para Identificar Distribuciones Posteriores

RAÚL HERNÁNDEZ-MOLINAR

Tulane University

1. INTRODUCCIÓN

La metodología que utiliza el enfoque Bayesiano está siendo utilizada con mayor frecuencia por los investigadores en el análisis de datos relacionados con el área de Confiabilidad, debido a que es posible utilizar información generada anteriormente y combinarla con datos muestrales para apoyar la toma de decisiones.

La metodología Bayesiana tiene una estrecha relación con métodos para obtener funciones de verosimilitud, y esto nos permite combinar la información que proporciona una muestra con información que se encuentra asociada a la distribución de un parámetro o parámetros. La identificación de distribuciones posteriores basada en el Teorema de Bayes, en ocasiones resulta compleja debido a la integración numérica requerida.

Es interesante encontrar que el empleo de la metodología sugerida aquí para identificar distribuciones posteriores; permite conocer con relativa facilidad el comportamiento de la(s) variable(s); además de que también nos motiva a emplear herramientas estadísticas tales como: Análisis Exploratorio de Datos o Gráficas Computacionales.

La perspectiva de *Remuestreo de Variables Aleatorias vía Bootstrap Ponderado*, utilizada en este trabajo, ofrece también un gran atractivo desde el punto de vista pedagógico debido a la facilidad para ser implementada, y a que no requiere el empleo de integración numérica sofisticada, lo que nos da la oportunidad de analizar y explorar bajo un enfoque proactivo, el comportamiento de las variables que son de interés en la investigación.

2. EL PARADIGMA BAYESIANO

Si se toma en cuenta que se ha obtenido un conjunto de datos a partir de un modelo paramétrico basado en un vector de dimensión finita θ ; es posible establecer que el Proceso de Aprendizaje Bayesiano considera:

$$p(\theta|x) = \frac{l(\theta;x)p(\theta)}{\int l(\theta;x)p(\theta)d(\theta)}$$

Es importante notar que la operación de integración tiene gran importancia en Estadística Bayesiana; tanto para definir la constante de normalización, como para definir la distribución marginal, la esperanza o la varianza de la(s) variables bajo estudio. Este trabajo sugiere el empleo de una alternativa que facilita el empleo del Teorema de Bayes en un contexto de Muestreo y Remuestreo.

3. MÉTODO DE REMUESTREO VÍA BOOTSTRAP PONDERADO

Consideremos que tenemos una función positiva $f(\theta)$ la cual puede ser normalizada, de tal forma que una función de densidad

$$h(\theta) = \frac{f(\theta)}{\int f(\theta)d(\theta)}$$

pueda ser utilizada para identificar una distribución posterior.

Como lo señalan Smith y Gelfand (1992), así como Meeker y Escobar (1996); una alternativa para producir $h(\theta)$ consiste en utilizar el método Bootstrap Ponderado, mediante el cual

$$h(\theta) = p(\theta|DATA) = \frac{L(DATA|\theta)p(\theta)}{\int L(DATA|\theta)p(\theta)d(\theta)}$$

o bien

$$f(\theta|DATA) = \frac{R(\theta)p(\theta)}{\int R(DATA|\theta)p(\theta)d(\theta)}$$

donde

$$R(\theta) = \frac{L(\theta)}{\int L(\theta)d(\theta)}$$

corresponde a la verosimilitud relativa; el cual actúa como una probabilidad de Remuestreo.

Debe notarse que esto implicaría que aquellos valores de θ en nuestra muestra previa, tendrán mayor probabilidad de ser obtenidos en una muestra posterior. Una vez que se tienen definidos:

- a) los valores muestrales ,
- b) la distribución que corresponde a estos valores muestrales, y
- c) la distribución previa del parámetro(s) bajo estudio,

es posible aplicar en forma directa este método. Notar que la función de verosimilitud puede ser obtenida de manera inmediata si se tiene la información relacionada con la muestra.

4. SIMULACIÓN MONTE CARLO DE LA DISTRIBUCIÓN POSTERIOR DE θ

Es posible aproximar la Distribución Posterior utilizando expresiones que representen correctamente la verosimilitud relativa (la probabilidad de que el valor de θ sea obtenido); así como la función inversa acumulada de la distribución de la variable que representa a la muestra obtenida. El procedimiento es el siguiente:

1. Generar los valores de θ_i , $i = 1, \dots, K$. Estos deben corresponder a una muestra aleatoria generada a partir de la distribución del parámetro, $p(\theta)$, la cual es conocida previamente.
2. Calcular $R(\theta_i)$, con base en la función de verosimilitud correspondiente. Generar U_i , a partir de una distribución uniforme con parámetros $(0,1)$.
3. Seleccionar como parte de la muestra posterior la observación i -ésima (θ_i), la cual tiene asociada una probabilidad $R(\theta_i)$ que satisface la condición:

$$U_i \leq R(\theta_i)$$

Es posible indicar que estas observaciones provienen de una distribución posterior $p(\theta | DATA)$ (véase Smith y Gelfand, 1992).

5. EJEMPLOS

5.1 El Caso de una Binomial cuyo Parámetro θ proviene de una Distribución Uniforme

Primero consideremos el caso de una población binomial, asumiendo que se tiene la siguiente información:

	i				
	1	2	3	4	5
x_i	1	2	2	3	4
n_i	3	4	5	4	5

La verosimilitud correspondiente a θ , dado que se conoce el valor de X puede representarse de la siguiente forma:

$$\theta^{\sum x_i} (1-\theta)^{\sum (n_i - x_i)} \prod \binom{n_i}{x_i}$$

La distribución asociada al parámetro θ se considera uniforme en el rango $(0,1)$. Utilizando el procedimiento descrito anteriormente, generamos la distribución posterior (aquí estamos “remuestreando” con base en la muestra previa).

5.2 El Caso de la Suma de dos Binomiales cuyos Parámetros θ_1 y θ_2 provienen de una Distribución Uniforme

Ahora veamos como generar la distribución posterior de una población, que sabemos puede ser generada a partir de la suma de dos binomiales; es decir,

$$X_{1i} \sim \text{BINOMIAL}(n_{1i}, \theta_1)$$

$$X_{2i} \sim \text{BINOMIAL}(n_{2i}, \theta_2)$$

en donde ambas son independientes y n_{1i}, n_{2i} son valores conocidos. Los valores utilizados en este ejemplo son los siguientes:

	<i>i</i>			
	1	2	3	4
n_1	3	7	3	4
n_2	4	5	5	5
Y_i	5	3	4	6

Entonces, sabemos que:

$$Y_i = X_{1i} + X_{2i}, \quad i = 1, 2, 3, 4$$

y la verosimilitud para θ_1, θ_2 dado que se tienen los valores de Y_i se puede representar de la siguiente forma:

$$\prod \sum \binom{n_{1i}}{j_i} \binom{n_{2i}}{y_i - j_i} \theta_1^{j_i} (1 - \theta_1)^{n_{1i} - j_i} \theta_2^{y_i - j_i} (1 - \theta_2)^{n_{2i} - y_i + j_i}$$

donde el dominio de j_i queda definido por:

$$\{\max\{0, y_i - n_{2i}\} \leq j_i \leq \min\{n_{1i}, y_i\}\}$$

En base a esta información, se genera la distribución posterior para tratar de identificar los valores de θ_1 y θ_2 .

Resulta interesante y útil observar que una vez generada la distribución posterior, es posible llevar a cabo transformaciones de la función encontrada en caso de ser necesario. Por ejemplo, el empleo de la transformación logit (log-odds) puede ser de interés.

5.3 El Caso de una Variable Gamma cuyos Parámetros θ_1 y θ_2 , provienen de una Distribución Uniforme con Parámetros (0,1)

Este ejemplo utiliza los siguientes valores:

$$x_i : 0.22, 0.50, 0.88, 1.00, 1.32, 1.33, 1.54, 1.76, 2.50, 3.00$$

los cuales corresponden al tiempo de vida después de una prueba realizada a ciertos componentes aéreos (Crowder, 1991).

En este caso se considera que la verosimilitud correspondiente a θ_1 y θ_2 es:

$$\frac{1}{\Gamma(\theta_1)^n (\theta_2)^{\theta_2}} \exp -\frac{1}{\theta_2} \sum X_i \prod X_i^{(\theta_1-1)}$$

REFERENCIAS

- Crowder, M.J., Kimber, A.C., Smith, R.L., and Sweeting, T.J. (1991). *Statistical Analysis of Reliability Data*. London: Chapman & Hall.
- Meeker, W.Q. and Escobar, L.A. (1997). *Statistical Methods for Reliability Data*. John Wiley & Sons Inc.
- Smith A.F.M. and Gelfand A.E. (1992). Bayesian Statistics Without Tears: A Sampling-Resampling Perspective, *The American Statistician* 46, 84-88.

Análisis de una Serie Biomédica a través de un Modelo Autorregresivo Bayesiano

GABRIEL HUERTA

ISDS, Duke University

1. INTRODUCCIÓN

Desarrollos recientes en descomposiciones de series de tiempo permiten establecer inferencias sobre las componentes subyacentes de una serie observada. De aquí que se tenga el interés de especificar distribuciones iniciales directamente sobre los parámetros que definen dichas componentes latentes. Este enfoque es presentado y desarrollado en este trabajo en el contexto de modelos autorregresivos lineales. Adicionalmente, para obtener inferencias de las distribuciones finales relevantes se propone un método Monte Carlo basado en Cadenas de Markov (MCCM). El enfoque que se adopta aquí es paralelo al de Barnett, et al. (1996a,b) donde se ilustra la versatilidad que los métodos MCCM ofrecen para estructurar el conocimiento inicial sobre parámetros interpretables desde el punto de vista práctico. Estos autores especifican distribuciones iniciales sobre la función de autocorrelación parcial, mientras que aquí las distribuciones iniciales se estructuran sobre los recíprocos de las raíces del polinomio que caracteriza al modelo. Dichas raíces determinan la descomposición de una serie autorregresiva, como se expone a continuación.

2. DESCOMPOSICIONES PARA SERIES DE TIEMPO

Suponga que la serie de tiempo $\{x_t\}$ de observaciones igualmente espaciadas sigue un modelo autorregresivo de orden p , es decir

$$x_t = \sum_{i=1}^p \phi_i x_{t-i} + \varepsilon_t$$

donde ε_t es una sucesión de errores no correlacionados de media cero y $\phi = (\phi_1, \dots, \phi_p)'$ es el vector de coeficientes AR. Los errores se suponen normales, $\varepsilon_t \sim N(\varepsilon_t | 0, \sigma^2)$ de varianza constante. Se introduce el polinomio característico

$$\phi(u) = 1 - \sum_{j=1}^p \phi_j u^j = \prod_{i=1}^p (1 - \alpha_j u)$$

donde $\{\alpha_1, \dots, \alpha_p\}$ son los recíprocos de las raíces de $\phi(u) = 0$. Si se supone que el proceso es estacionario $|\alpha_j| \leq 1$ para cada j , y estacionariedad estricta implica que ninguna raíz es de módulo unitario. Nos concentraremos en modelos estacionarios ya que de otra manera la función de pronóstico se vuelve explosiva y ninguna predicción razonable se puede obtener del modelo. Con el operador de rezago B , tenemos la representación

tradicional $\phi(B)x_t = \varepsilon_t$. Suponga que las raíces α_j son distintas y aparecen como C pares conjugados y $R = p - 2C$ raíces reales. Denote los pares conjugados como $r_j \exp(\pm i\omega_j)$ para $j = 1, \dots, C$, y las raíces reales como r_j para $j = 2C + 1, \dots, p$; aquí tanto r_j como ω_j son cantidades reales con $\omega_j > 0$.

La descomposición de la serie se deriva de que el proceso se puede invertir debido a la estacionariedad, y vía expansión en fracciones parciales se convierte en

$$x_t = \sum_{j=1}^C z_{tj} + \sum_{j=2C+1}^p a_{tj}$$

donde z_{tj} y a_{tj} son los procesos latentes que corresponden a las raíces complejas y reales respectivamente. Correspondientes a las raíces reales, se tiene que a_{tj} son procesos AR(1) y correspondientes a los pares conjugados complejos de raíces se tiene que z_{tj} son procesos ARMA(2,1) donde la componente AR representa una comportamiento periódico de frecuencia ω_j , es decir, con periodo o longitud de onda $2\pi/\omega_j$. El desarrollo de este resultado a través de modelos dinámicos lineales (modelos espacio-estado) aparece en West (1997) y es discutido con mayor amplitud en West y Harrison (1997, sección 9.5, capítulo 15).

A partir de este momento, se supone que la distribución marginal inicial en la varianza σ^2 de los errores es la usual de referencia para parámetros de escala. Adicionalmente se supone que dicha varianza es independiente de las raíces.

3. DISTRIBUCIONES INICIALES EN LAS RAÍCES

Primero se especifican cotas superiores C y R para el máximo número de pares complejos y raíces reales respectivamente. Esto implica un valor máximo en el orden del modelo $p = 2C + R$. Condicional en C y R , suponemos iniciales independientes para α_j por casos. Para cada $j = 1, \dots, R$, la raíz real r_j tiene una inicial con soporte $|r_j| \leq 1$ y densidad

$$r_j \sim \pi_{r,0} I_0(r_j) + \pi_{r,-1} I_{(-1)}(r_j) + \pi_{r,1} I_1(r_j) + (1 - \pi_{r,0} - \pi_{r,-1} - \pi_{r,1}) U(|r_j| - 1, 1)$$

donde $I(\cdot)$ es la función indicadora. De esta forma la marginal de r_j permite una raíz en la frontera $r_j = \pm 1$. La masa *a priori* en $r_j = 0$ define una probabilidad positiva de que la raíz sea igual a cero, así que el número de raíces reales puede ser menor al máximo R . Para valores fijos de $\pi_{r,\cdot}$, se infieren iniciales para cantidades como el número de raíces diferentes de cero; esta es simplemente una binomial $Bn(R, 1 - \pi_{r,0})$. Esta consideración nos permite evaluar diferentes valores para $\pi_{r,0}$. En vez de dar valores fijos a $\pi_{r,\cdot}$, estas cantidades se trabajan como hiperparámetros a estimar sobre los que se tiene que especificar una distribución inicial. Aquí se utiliza la distribución conjunta Dirichlet con marginales uniformes.

En el caso de las raíces complejas, para cada $j=1, \dots, C$, la inicial se define en el par (r_j, λ_j) con soporte $0 \leq r_j \leq 1$ y $2 < \lambda_j < \lambda_u$ para alguna cota superior λ_u en los periodos. Se toma r_j y λ_j como condicionalmente independientes con

$$r_j \sim \pi_{c0} I_0(r_j) + \pi_{c1} I_1(r_j) + (1 - \pi_{c1} - \pi_{c0}) \beta r_j^{\beta-1} I_{(0,1)}(r_j);$$

y

$$\lambda_j \sim U(\lambda_j | 2, \lambda_u)$$

Igual que con las raíces reales se tiene una masa de probabilidad en la frontera definida por la condición de estacionariedad y la masa en cero implica que el número de raíces complejas puede ser menor al máximo número C . Nuevamente se utiliza una distribución Dirichlet-Uniforme sobre los hiperparámetros π_{c*} . La combinación de las distribuciones de probabilidades sobre el número de componentes complejas y reales implica una distribución en el orden del AR.

4. ESTRUCTURA Y ANÁLISIS DE LA DISTRIBUCIÓN FINAL

Se define $\mathbf{X} = \{x_1, \dots, x_n\}$ como la serie original y dado el máximo orden del modelo p , $\mathbf{Y} = \{x_0, x_1, \dots, x_{(p-1)}\}$ denota los valores iniciales de la serie. El análisis incorpora inferencias formales sobre estos valores iniciales. Los parámetros del modelo se denotan por $\varphi = \{\alpha_j, j = 1, \dots, p; (\pi_{r,-1}, \pi_{r0}, \pi_{r1}); (\pi_{c0}, \pi_{c1}); \sigma^2\}$. Las inferencias *a posteriori* consisten en resumir la distribución $p(\varphi, \mathbf{Y} | \mathbf{X})$. Para cualquier subconjunto ξ de elementos φ , denote por $\varphi \setminus \xi$ a los elementos de φ sin considerar a ξ . El método MCCM se basa en la estructura estándar del Gibbs sampling. Concretamente,

- Para cada $j = 2C + 1, \dots, r$, se muestran las raíces reales de $p(r_j | \varphi \setminus r_j, \mathbf{X}, \mathbf{Y})$;
- Para cada $j = 1, \dots, C$, se muestran las raíces complejas de $p(r_j, \lambda_j | \varphi \setminus (r_j, \lambda_j), \mathbf{X}, \mathbf{Y})$;
- Se muestran los hiperparámetros de

$$p(\pi_{r,-1}, \pi_{r0}, \pi_{r1} | \varphi \setminus (\pi_{r,-1}, \pi_{r0}, \pi_{r1}), \mathbf{X}, \mathbf{Y})$$

y

$$p(\pi_{c0}, \pi_{c1} | \varphi \setminus (\pi_{c0}, \pi_{c1}), \mathbf{X}, \mathbf{Y});$$

- Se muestrea la varianza de $p(\sigma^2 | \varphi, \mathbf{X}, \mathbf{Y})$;
- Se muestran los valores iniciales de $p(\mathbf{Y} | \varphi, \mathbf{X})$.

De todas estas distribuciones, la única para la que no se tiene una condicional completamente especificada es la que corresponde a (r_j, λ_j) y en este caso es necesario efectuar un paso de aceptación y rechazo basado en el algoritmo Metropolis-Hastings (Tierney, 1994).

5. ANÁLISIS DE UNA SERIE BIOMÉDICA

Diggle y al Wasel (1997) presentan un análisis espectral para varias series biomédicas, una de las cuales es utilizada aquí para ilustrar el método descrito en las secciones anteriores. La serie consta de $n = 60$ observaciones y cada medición es el nivel de la hormona LH (Luteinising Hormone) registrado en intervalos de un minuto durante una hora en una mujer post-menopáusica. La serie aparece en la Figura 1a. Se seleccionó $C = 20$ y $R = 20$ de tal manera que $p = 60$ para ilustrar el ajuste de un AR(p) con $n = p$ observaciones. Además se especificaron valores de $\beta = 1$ y $\lambda_u = 30$, el máximo periodo observable en una serie de 60 observaciones. Se corrieron cuatro cadenas del MCCM con valores iniciales muestrados de la distribución *a priori*. El diagnóstico de Gelman y Rubin

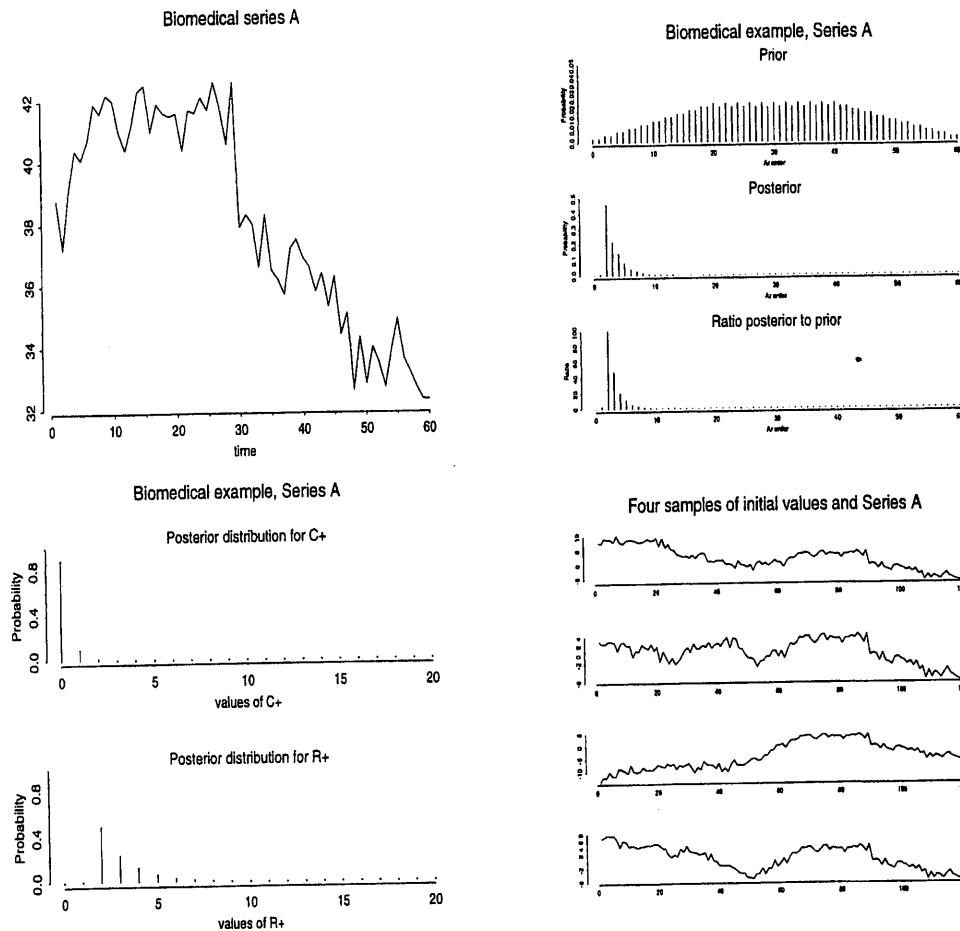


Figura 1. (a) 60 mediciones de la hormona LH tomadas en intervalos de un minuto durante una hora; (b) distribución inicial para el orden de un AR(60), $C=20$, $R=20$; distribución final para p_+ y el cociente final entre inicial con base en 10,000 muestras del MCCM; (c) distribución final para C_+ y R_+ con base en 10,000 muestras del MCCM. (d) datos originales y cuatro muestras de valores iniciales provenientes de la mezcla definida por la distribución final de p_+ .

(1992) muestra que con 10000 iteraciones el método prácticamente converge a la distribución estacionaria. Las inferencias de la distribución final se basaron en dos muestras obtenidas de dos cadenas de 10000 iteraciones donde en cada una se descartaron las primeras 5000. Las figuras 1b-1d ilustran el tipo de inferencias *a posteriori* que se obtienen de estos datos.

En la Figura 1b aparece la distribución inicial para el orden del modelo (denotado por p_+) junto con la correspondiente distribución final y el cociente final entre inicial. Aunque la inicial no es estrictamente uniforme, se comporta de esta forma, como se puede notar de comparar la distribución final con la taza final-inicial. El modelo de orden 2 se ve relativamente favorecido con un decaimiento de las probabilidades finales hasta el modelo de orden 10. La Figura 1c presenta las distribuciones finales para C_+ y R_+ , el número de raíces complejas y reales respectivamente, y muestra que la serie no presenta componentes cíclicas (cero raíces complejas) y 2 ó 3 raíces reales. Finalmente, en la Figura 1d aparecen los datos junto con 4 muestras de valores iniciales $\mathbf{Y} = \{x_0, x_{-1}, \dots, x_{-59}\}$. Este tipo de gráficos permiten valorar subjetivamente el ajuste del modelo a la serie. En este caso, las aparente tendencia en los datos originales son detectada por los valores iniciales. Este aparente comportamiento no estacionario es confirmado después de ordenar las muestras que corresponden a las raíces reales y notar que la máxima raíz real tiene un probabilidad de 0.95 de tener módulo unitario.

REFERENCIAS

- Barnett, G., Kohn, R. y Sheather, S. (1996). Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *Working Paper*, Statistics Group: Australian Graduate School of Management, University of New South Wales, Australia.
- Barnett, G., Kohn, R. y Sheather, S. (1996). Robust Bayesian estimation of autoregressive moving average models. *Working Paper*, Statistics Group: Australian Graduate School of Management, University of New South Wales, Australia.
- Diggle, P.J., y al Wasel, I. (1997). Spectral analysis of replicated biomedical time series (with discussion). *J. Roy. Statist. Soc.*, (Ser. B), (por aparecer).
- Gelman, A. y Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7, 457-511.
- Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Ann. Statist.*, 22, 1701-1762.
- West, M., y Harrison J. (1997). *Bayesian Forecasting and Dynamic Linear Models* (2da Ed.), Springer-Verlag: New York.
- West, M. (1997). Time series decomposition. *Biometrika*, (por aparecer).

Cadenas de Markov para Analizar el Comportamiento de la Flota Pesquera Mexicana en la Búsqueda de Atún

J. LARA-TEJEDA

y

R. SOLANA-SANORES

Facultad de Ciencias, U. A. B. C.

1. INTRODUCCIÓN

La dinámica del comportamiento de una flota pesquera puede descomponerse como sigue (Hilborn, 1985): inversión, movimiento, potencia de captura y descarga. En el caso de la pesca superficial del atún en el Océano Pacífico Oriental (OPO), se observa que en su movimiento existe un componente de búsqueda en el cual intervienen señales que están asociadas a la presencia de cardúmenes. Ellas son de naturaleza aleatoria, por lo que su estudio requiere de un tratamiento estocástico (Deriso *et al.*, 1991). Las señales relacionadas con un lance pesquero son: presencia de aves, mamíferos marinos, objetos flotantes y brisas.

La pesca superficial del atún en el OPO, se realiza empleando una red de cerco. Dependiendo de la señal identificada se distinguen tres tipos de maniobras: 1) lances sobre delfín; 2) lances sobre cardúmenes libres; y 3) lances sobre objetos flotantes. Los lances se realizan posterior a una serie de avistamientos (señales) y maniobras que realiza la embarcación. Estas señales y maniobras son registradas por observadores científicos del Programa Nacional de Aprovechamiento del Atún Y Protección de Delfines, PNAAPD (Compéan-Jiménez, 1993), aproximadamente cada cuatro horas, cada que exista un cambio de actividad o se dé un avistamiento. Los registros son almacenados en una base de datos denominada Informe Diario (ID), que fue la fuente principal de información. De ellos, se analizaron un total de 187,876, que corresponden al 66% de los registros tomados a bordo de viajes de la flota atunera mexicana durante los años de 1992 a 1994.

Se puede suponer que los eventos son aleatorios y forman una secuencia encadenada para conducir a la realización de un lance pesquero. Dado esto, la finalidad del presente trabajo es analizar las posibles "rutas" de señales y eventos que llevan al éxito de un lance pesquero.

2. MATRIZ DE TRANSICIÓN DEL COMPORTAMIENTO DE LA FLOTA ATUNERA MEXICANA

Se considera que la sucesión de los eventos corresponde a un proceso de tiempo discreto (Bhat, 1972). Esto, por el hecho de que el observador científico registra regularmente los datos de dichos eventos. Asimismo, se supone que cada evento es un estado el cual, para un intervalo de tiempo, es función del evento anterior y no de otros. Así, el espacio estocástico es el conjunto de todos los eventos que se registran en un viaje.

Si lo anterior es cierto, es posible determinar las probabilidades condicionales de que el sistema (un viaje pesquero), cambie de un evento a otro. Esto es de tal manera que, dado un evento, las probabilidades de transición hacia otros eventos suman 1.

Con un programa en LISP, se construyó un conjunto cuyos elementos fueron los cambios de un evento a otro o de un evento asimismo. Con esto, se identificaron todos los posibles cambios de evento y su frecuencia correspondiente. Con las frecuencias calculadas, se construyó la matriz de transición dentro de una hoja de cálculo para la posterior identificación de dependencia de los eventos. Esto último, se realizó por medio de una prueba de hipótesis con Ji-cuadrada.

Los resultados de las pruebas correspondientes sobre la no existencia de estados absorbentes y transitorios, y la recurrencia de estados, demuestran que se cumplen estos supuestos. Esto por el hecho de que partiendo de un evento en un lapso se pase a otro y no se regrese al original. Además, se observó que al registrar un evento, existe la probabilidad de que se vuelva a registrar dicho evento.

Es importante destacar que las probabilidades de transición de los eventos en un viaje pesquero correspondieron a una transición estacionaria. Esto debido a que las matrices de transición tienden a las probabilidades marginales. Sea Π_j , la probabilidad estacionaria o de estabilidad de la cadena de Markov para el estado j , entonces se satisfacen las siguientes ecuaciones de estabilidad. Sea $\Pi_j \geq 0$;

$$\Pi_j = \sum_{i=0}^M \Pi_i P_{i,j} \quad (1)$$

($j = 0, 1, \dots, M$), donde $P_{i,j}$ es la probabilidad de transición del estado i al j , y $\sum_{i=0}^M \Pi_i = 1$.

Con esto se puede calcular la probabilidad Π_j , con la que se puede estimar el valor esperado de la primera visita μ_{jj} (Bhat, 1972), ya que,

$$\Pi_j = \frac{1}{\mu_{jj}} \quad (2)$$

($j = 0, 1, \dots, M$). Con lo anterior se plantea el siguiente sistema:

$$\begin{aligned} \Pi_{BUSCAR} &= \Pi_{BUSCAR} P_{BUSCAR,BUSCAR} + \Pi_{DERIVA} P_{DERIVA,BUSCAR} + \dots + \Pi_{OTRASE} P_{OTRASE,BUSCAR} \\ \Pi_{CHAPOT} &= \Pi_{BUSCAR} P_{BUSCAR,CHAPOT} + \Pi_{DERIVA} P_{DERIVA,CHAPOT} + \dots + \Pi_{OTRASE} P_{OTRASE,CHAPOT} \\ 1 &= \Pi_{BUSCAR} + \Pi_{DERIVA} + \dots + \Pi_{OTRASE} \end{aligned} \quad (3)$$

Cuya representación matricial es $\mathbf{A} * \mathbf{B} = \mathbf{C}$. Donde \mathbf{A} es la matriz que contiene las probabilidades de transición de los eventos; \mathbf{B} es el vector de las probabilidades estacionarias de la cadena de Markov; y \mathbf{C} es el vector que contiene ceros, excepto en el último renglón cuyo valor es uno.

Encontrando la solución para \mathbf{B} y con (2), se observa que el tiempo esperado de recurrencia es $\mu_i < \infty$, $\forall i \in \{\text{Espacio de Estados}\}$. Se cumple, entonces, que todos los eventos son recurrentes positivos y esto satisface el hecho de que una cadena de Markov sólo tiene estados de este tipo (Bath, 1972).

En cuanto al tiempo de recurrencia, sea $f_{i,j}^*$ la probabilidad de que empezando en el evento i , se pase al estado j en un tiempo finito y el tiempo de recurrencia μ_i que se da en las expresiones siguientes (Prawda, 1980)

$$\mu_i = \sum_{n=1}^{\infty} n f_{i,i}^n \quad \text{y} \quad \mu_{i,j} = \sum_{n=1}^{\infty} f_{i,j}^n + \sum_{k \neq j} P_{i,k} \mu_{k,j} \quad (4)$$

esto permite estimar el tiempo de primera visita a un evento a partir de otro dado, ya que para un estado recurrente:

$$\mu_{i,j} = 1 + \sum_{k \neq j} P_{i,k} \mu_{k,j} \quad (5)$$

Aplicando (4) y (5) sobre la matriz de transición y conociendo el lapso de tiempo promedio entre registro y registro (1 hora 56 minutos), se estimó el tiempo promedio del primer lance sobre mamífero después del evento SALIDA en 47 horas 17 minutos, que corresponde a un total de 24.46 registros antes del lance. Para corroborar este resultado, se procedió a trabajar en forma exhaustiva la matriz original de los registros de observadores; esta dió como resultado 26.46 registros para llegar a un lance, después de salir de puerto y que corresponde a 51 horas 9 minutos, que cae en el intervalo de confianza.

3. ESPACIOS DE ESTADOS DE BÚSQUEDA

Con las matrices de transición es posible dibujar los espacios de estado. Cada estado, que corresponde a un evento, se une a otro por el trazo de una línea dirigida, que representa la transición correspondiente. Se escogen los estados básicos que conducen a un lance y a cada trazo de una transición se le asocia la probabilidad correspondiente que permite identificar las rutas de eventos que conducen a los lances. Por ejemplo, para un lance sobre mamífero se observa lo siguiente (fig 1): partiendo del evento BUSCAR se tienen las probabilidades de avistamientos de mamíferos (.145), caza (.3484), y lance con mamífero (.7655).

La información sobre las “rutas” de señales que sigue la flota atunera mexicana permitirá una mejor planeación en las estrategias de movimiento de los barcos. Esto, repercutirá en una mejoría de la dinámica y la captura de la flota atunera mexicana.

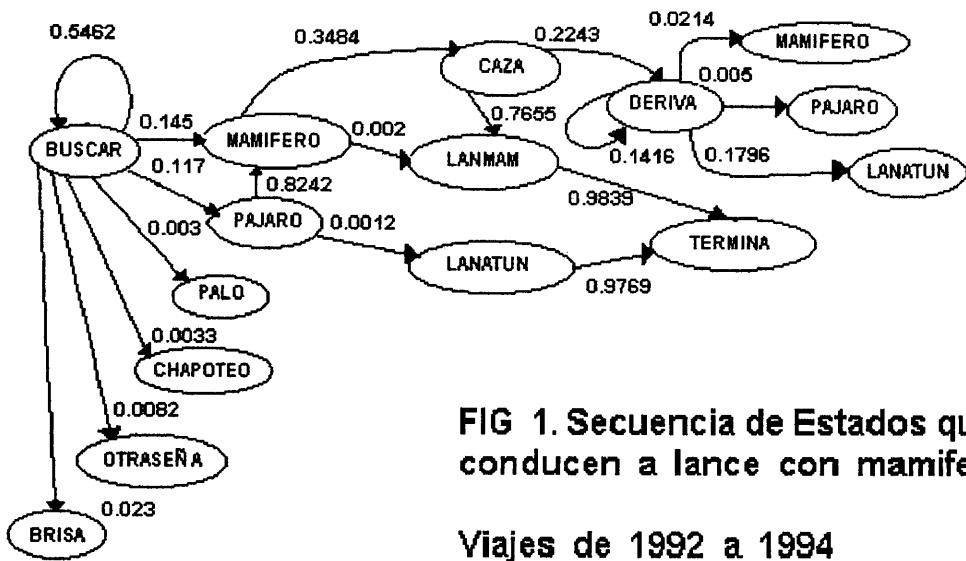


FIG 1. Secuencia de Estados que conducen a lance con mamífero

Viajes de 1992 a 1994

Agradecimientos. El presente trabajo fue parcialmente financiado por el proyecto CONACYT número 3946P-B. Los autores hacen patente su agradecimiento al Dr. Guillermo Compéan Jiménez por haber aportado la información. Asimismo, a los observadores científicos del Programa Nacional de Aprovechamiento del Atún y Protección de Delfines.

REFERENCIAS

- Bhat, U.N. (1972). *Elements of applied stochastic processes*. New York: J. Wiley & Sons.
- Compéan - Jiménez, G. (1993). Aprovechamiento del Atún y Protección del Delfín. En: *Biodiversidad marina y costera* (eds. SI Salazar-Vallejo y NE González), pp. 129-138. México: Com Nal Biodiversidad y CIQRO.
- Deriso, R.B., R.G. Punsly y W.H. Bayliff. (1991). A Markov movement model of yellowfin tuna in the Eastern Pacific Ocean and some analysis for international management, *Fisheries Research*, **11**, 375-395.
- Hilborn, R. (1985). Fleet dynamics and individual variation: Why some people catch more fish than others, *Canadian Journal of Fisheries and Aquatic Science*, **42**, 2-13.
- Prawda, J. (1980). *Métodos y Modelos de Investigación de Operaciones*. LIMUSA: México.

Introducción a los Modelos Lineales Bayesianos

ADRIANA LÓPEZ GARCÍA y GABRIEL NÚÑEZ ANTONIO

U.N.A.M.

1. INTRODUCCIÓN

En estadística existen dos grandes enfoques para el análisis de modelos estadísticos, el enfoque clásico y el enfoque Bayesiano. Los modelos de regresión lineal como una clase particular de los modelos estadísticos pueden analizarse desde ambos enfoques.

El enfoque Bayesiano, a diferencia del enfoque clásico, se caracteriza por permitir la actualización de la información inicial que se tenga sobre los parámetros desconocidos. Dicha actualización se logra al incorporar en el análisis la información que proporciona una muestra aleatoria. Esta incorporación de información muestral se logra a través del *teorema de Bayes*.

Dentro de la teoría Bayesiana existen dos formas relevantes de llevar a cabo el análisis de problemas estadísticos. Una es sin considerar y la otra es considerando los elementos de la teoría de la decisión. Esta última permite combinar la información inicial y muestral con otros aspectos importantes en el problema; por ejemplo, incorporar información sobre las posibles consecuencias que ocurrán al tomar cierta decisión, las cuales pueden ser cuantificadas mediante la especificación de las llamadas funciones de utilidad.

En este trabajo se presenta un análisis Bayesiano, considerando la teoría de la decisión, del modelo de regresión lineal múltiple.

2. TEORÍA DE LA DECISIÓN Y ESTADÍSTICA

2.1 Teoría de la Decisión

Si un decisor es capaz de aceptar unos cuantos principios de razonamiento coherente o *principios de coherencia* (Bernardo, 1981) como fundamentos axiomáticos básicos, entonces siempre dispondrá de una forma razonable y única para tomar decisiones de manera coherente. Para lograr ésto el decisor debe ser capaz de determinar el conjunto de todas las posibles decisiones, D , el conjunto Θ de los sucesos inciertos y las consecuencias, c_{ij} . Posteriormente, el decisor debe de cuantificar y evaluar la información que posea de los diferentes sucesos inciertos, así como sus preferencias entre las posibles consecuencias.

La manera de cuantificar la incertidumbre de Θ es mediante la asignación de probabilidades a sus diferentes elementos. Las preferencias entre las posibles consecuencias pueden ser evaluadas numéricamente a través de una medida de utilidad $U(\cdot)$. Así, si se tiene que C es el conjunto de todas las posibles consecuencias y \prec el orden de preferencia entre ellas, una utilidad es una función definida como $U:C \rightarrow \mathbf{R}$, donde $c_1 \prec c_2$ si y sólo si $U(c_1) < U(c_2)$, para todo par c_1, c_2 de consecuencias en C .

Por otro lado, en un problema de decisión se tiene implícita la existencia de una función

$$F : D \times \Theta \rightarrow C.$$

Ya que para cada decisión y suceso incierto se tiene asociada una consecuencia. Así, para cada decisión d_i en D y cada suceso θ en Θ se tiene una utilidad $U(d_i, \theta)$ definida a través de la composición $U \circ F$.

De esta manera, si se toma la decisión d en las condiciones H se puede obtener la utilidad esperada $U^*(d)$ para cada decisión. En caso de que Θ sea finito o numerable $U^*(d)$ se calcula como

$$U^*(d) = \sum_j U(d, \theta_j) P(\theta_j | H).$$

Cuando Θ es no numerable y $U(d, \theta)$ es P -medible, la utilidad esperada se obtiene a través de una integral de Lebesgue

$$U^*(d) = \int U(d, \theta) dP.$$

Los axiomas de coherencia garantizan que la solución óptima al problema de decisión será aquella decisión que maximice la utilidad media o utilidad esperada. A ésto se le conoce como el *criterio de decisión de Bayes*, y la bondad del mismo reside esencialmente en su fundamento axiomático.

2.2 Problemas de Decisión Estadísticos

Los problemas de inferencia estadística paramétrica se pueden ver como problemas de decisión bajo incertidumbre. Por ejemplo, para el problema estadístico de estimación puntual del parámetro θ , donde θ es el parámetro de un modelo probabilístico $f(x|\theta)$, el espacio de decisiones es igual a el espacio paramétral, es decir, $D = \Theta$, y el conjunto de sucesos inciertos estará representado por los posibles elementos de Θ . Para el problema estadístico de prueba de hipótesis el espacio de decisión estará dado por las hipótesis a contrastar, y el conjunto de sucesos inciertos igual que en el de estimación puntual está representado por los posibles valores del espacio paramétral. En ambos casos la función de utilidad puede ser una cierta medida de información.

Como lo desconocido en un problema estadístico es el valor de θ , en principio, es necesario expresar el poco o mucho conocimiento inicial que se tenga sobre θ a través de una f.d.p., $\pi(\theta)$. A esta función se le conoce como *distribución inicial*.

Una vez que se determina la forma de cuantificar o describir la información inicial que se posee sobre el parámetro θ , a través de su distribución inicial, se está en condiciones de obtener la distribución final de θ . Dicha distribución describe el conocimiento que se tiene sobre θ tras incorporar a la información inicial la información que proporcionan los resultados experimentales. El camino que permite obtener dicha distribución final, $\pi(\theta | x_1, \dots, x_n)$, es el *teorema de Bayes*.

Teorema de Bayes. Sean x_1, \dots, x_n los resultados de algún experimento ξ con f.d.p. conjunta $f(x_1, \dots, x_n | \theta)$ y sea $\pi(\theta)$ la distribución inicial de θ . La distribución final de θ está dada por

$$\pi(\theta | x_1, \dots, x_n) = f(x_1, \dots, x_n | \theta) \pi(\theta) / \pi(x_1, \dots, x_n),$$

donde

$$\pi(x_1, \dots, x_n) = \int_{\Theta} f(x_1, \dots, x_n | \theta) \pi(\theta) d\theta.$$

3. EL MODELO DE REGRESIÓN LINEAL

Definición. El modelo de regresión lineal múltiple (MRLM) que relaciona una respuesta aleatoria Y con un conjunto de variables regresoras X_1, \dots, X_{p-1} es

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N(0, \Sigma)$$

con $\mathbf{y} = (y_1, \dots, y_n)', \beta = (\beta_0, \dots, \beta_{p-1})', \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$, \mathbf{X} la matriz de $n \times p$ con renglones $x_i' = (1, X_{i1}, \dots, X_{i,p-1})$ y $\Sigma = \tau^{-1} I_n$ la matriz, de varianzas y covarianzas de ε . En este caso $\sigma^2 = \tau^{-1}$ es la varianza de cada error ε_i , por lo que τI_n es la matriz de precisión de ε .

3.1 Análisis Bayesiano Conjugado

Es común considerar como distribución inicial, en el caso de un análisis conjugado, un miembro de la familia Normal-Gamma de la forma

$$\pi(\beta, \tau) \propto \tau^{p/2} \exp\left\{-\frac{\tau}{2}(\beta - \mu)' A(\beta - \mu)\right\} \tau^{\alpha-1} \exp\{-\tau\lambda\},$$

es decir, $\pi(\beta, \tau)$ una Normal-Gamma($\mu, \tau A, \alpha, \lambda$). En este caso

$$\pi(\beta, \tau | \mathbf{y}) = \text{Normal-Gamma}\left(\mu_1^*, \tau A_1^*, \alpha_1^*, \lambda_1^*\right)$$

$$\pi(\beta | \mathbf{y}) = t_p(\mu_1^*, \tau A_1^*) \text{ con } n + 2\alpha \text{ g.de.l.}$$

$$\pi(\tau | \mathbf{y}) = \text{Gamma}\left(\alpha_1^*, \lambda_1^*\right)$$

donde

$$\alpha_1^* = (n + 2\alpha) / 2$$

$$\lambda_1^* = \lambda + [\mathbf{y}' \mathbf{y} - (\mathbf{A}\mu + \mathbf{X}' \mathbf{y})' (A_1^*)^{-1} (\mathbf{A}\mu + \mathbf{X}' \mathbf{y})] / 2$$

$$\mu_1^* = (A_1^*)^{-1} (\mathbf{A}\mu + \mathbf{X}' \mathbf{y})$$

$$A_1^* = (\mathbf{A} + \mathbf{X}' \mathbf{X})$$

A continuación se presentan las estimaciones correspondientes de los parámetros usando funciones de utilidad.

Usando una función de utilidad cuadrática se tienen los siguientes estimadores:

$$\hat{\beta} = E_{\pi(\beta|y)}(\beta) = \mu_1^*$$

$$\hat{\tau}^{-1} = E_{\pi(\tau|y)}(\tau^{-1}) = \lambda_1^* (\alpha_1^* - 1)^{-1}$$

Usando como función de utilidad la divergencia de Kullback-Leibler se tienen los siguientes estimadores:

$$\hat{\beta} = \mu_1^*$$

$$\hat{\tau}^{-1} = \text{moda de } \pi(\tau^{-1}|y) = \lambda_1^* (\alpha_1^* + 1)^{-1}$$

Las regiones de $(1-\gamma)$ de probabilidad que se obtienen son las siguientes.

Para β :

$$R_{1-\lambda}(\beta) = \left\{ \beta : p^{-1}(\beta - \mu_1^*)' A^* (\beta - \mu_1^*) \leq F_{p, 2\alpha+n, 1-\gamma} \right\}.$$

Para τ^{-1} :

$$\left(2\lambda_1^*/\chi^2_{(2\alpha_1^*, 1-\frac{\gamma}{2})}, 2\lambda_1^*/\chi^2_{(2\alpha_1^*, \frac{\gamma}{2})} \right)$$

donde $F_{p, 2\alpha+n, 1-\gamma}$ es el cuantil superior de orden $(1-\gamma)$ de una distribución F con p y $2\alpha+n$ g. de l. y $\chi^2_{(2\alpha_1^*, 1-\frac{\gamma}{2})}, \chi^2_{(2\alpha_1^*, \frac{\gamma}{2})}$ los cuantiles de orden $(1-\frac{\gamma}{2})$ y $(\frac{\gamma}{2})$, respectivamente, de una distribución chi-cuadrada con $2\alpha_1^*$ g. de l.

Si se considera una función de utilidad cuadrática, la mejor predicción (cuando $\mathbf{X}=\mathbf{Z}$) del vector $W=(W, \dots, W)'$ de k observaciones que se quieran predecir, será

$$\hat{W} = E_{\pi(W|y)}(W),$$

donde $\pi(W|y) = t_k(\Lambda^{-1}\Upsilon, (2\alpha+n)\Lambda(\Psi - \Upsilon'\Lambda^{-1}\Upsilon)^{-1})$ con $2\alpha+n$ g. de l. y

$$\begin{aligned} \Lambda &= (I - \mathbf{Z}\Phi\mathbf{Z}') \\ \Upsilon &= \mathbf{Z}\Phi(A\mu + \mathbf{X}'y) \\ \Psi &= \mathbf{y}'y + \mu'A\mu - (A\mu + \mathbf{X}'y)\Phi(A\mu + \mathbf{X}'y) + 2\lambda \\ \Phi &= (A + \mathbf{X}'\mathbf{X} + \mathbf{Z}'\mathbf{Z})^{-1} \end{aligned}$$

3.2 Análisis Bayesiano no Informativo

Es posible llevar a cabo un análisis Bayesiano aún cuando no se cuente con información inicial o se esté limitado para especificarla, en este caso lo que se puede emplear son las llamadas distribución inicial *no informativas*; ver por ejemplo Jeffreys (1961) y Broemeling (1985). La metodología de los desarrollos es similar al caso conjugado.

4. CONCLUSIONES

Se puede decir que el enfoque Bayesiano, usando teoría de la decisión, ofrece una gran flexibilidad en la solución de los problemas estadísticos. Lo anterior porque este enfoque permite incorporar al análisis estadístico la información que cada decisor tenga sobre el problema y actualizar el conocimiento inicial sobre los parámetros desconocidos.

En este trabajo se consideró como distribución inicial conjugada para el MRLM, una distribución Normal-Gamma. Sin embargo, se puede usar cualquier distribución inicial ya que esto dependerá de la información inicial con que se cuente. Aún más, cualquier distribución inicial se puede aproximar a través de combinaciones convexas (mezclas) de distribuciones conjugadas, en este caso a través de mezclas de distribuciones Normal-Gamma.

En el MRLM, es importante notar la similitud de los resultados de estadística clásica y los resultados que se pueden obtener desde el punto de vista Bayesiano usando una distribución inicial de Jeffreys.

Para una revisión rápida de los desarrollos en el caso no informativo, así como de los resultados en los tópicos de pruebas de hipótesis y diagnósticos del modelo se puede revisar el trabajo de López y Nuñez (1995).

REFERENCIAS

- Bernardo, J.M (1981). *Bioestadística. Una perspectiva Bayesiana*. Valencia: Vicens-vives.
- Bernardo, J.M. y Smith, A.F.M. (1994). *Bayesian Theory*. New York: Wiley.
- Box, G.E.P Y Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Adisson-Wesley.
- Broemeling, D.L. (1985). *Bayesian Analysis of Linear Models*. New York: Marcel Dekker.
- Jeffreys, H. (1961). *Theory and Probability*, 3rd. ed. Oxford: Clarendon Press.
- López, G.A y Nuñez, A.G (1995). Introducción a los Modelos Lineales Bayesianos. *Tesis de Licenciatura*. Facultad de Ciencias, UNAM.

Un Método para Minería de Datos con Tres Variables

ANDRZEJ MATUSZEWSKI
Institute of Computer Science, Polonia

1. BASE DE DATOS PARA PACIENTES - COMO EJEMPLO

Vamos a suponer que tenemos una base de datos sobre el conjunto representativo de pacientes. La representatividad es garantizada por el médico que quiere analizarlos. En esta base hay dos variables categóricas, X y Y, que corresponden a dos síntomas importantes para el tipo de enfermedad bajo estudio. Vamos a suponer adicionalmente que el “comportamiento” de los síntomas puede depender de la edad de los pacientes. En nuestras consideraciones la “edad” será siempre una variable (Z) con un número finito y pequeño de valores. Lo último es importante porque la herramienta analítica que vamos a usar consiste más que nada en las tablas de contingencia y los métodos estadísticos correspondientes.

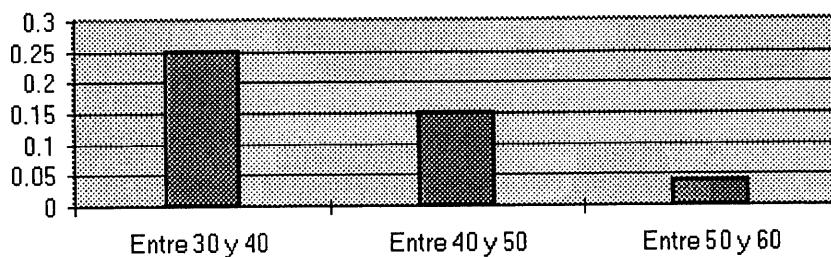


Figura 1. Correlación entre síntomas X y Y como función de la edad

La figura 1 muestra cómo se puede representar la influencia de la edad del paciente sobre la correlación entre X y Y. Es una forma muy fácil de entender para el médico. En este caso podemos concluir que al aumentar la edad, la correlación entre dos síntomas decrece. Interesan ciertas pruebas de hipótesis para responder a preguntas como:

1. ¿Cuáles de las tres correlaciones son significativas?
2. ¿La tendencia de decrecimiento de correlaciones es significativa o no?

Estas pruebas en muchas ocasiones serán tratadas como algo adicional, que no es de primera necesidad. En otras palabras tales pruebas tal vez no pertenezcan al área de minería de datos.

De hecho no siempre existe la posibilidad de que sea apropiado el coeficiente de correlación y por tanto no siempre podemos hacer una figura como la de arriba.

2. PARADOJA DE SIMPSON

Aún en análisis tan elementales como los que se pueden incluir en programas computacionales de minería de datos, hay que considerar la famosa paradoja de Simpson. Esta paradoja no comparte con las muchas otras paradojas de la estadística el valor de ser

puramente teórica. Es más, se puede decir que gran parte de los resultados de importancia de muchos análisis de minería de datos tienen que ver algo con dicha paradoja.

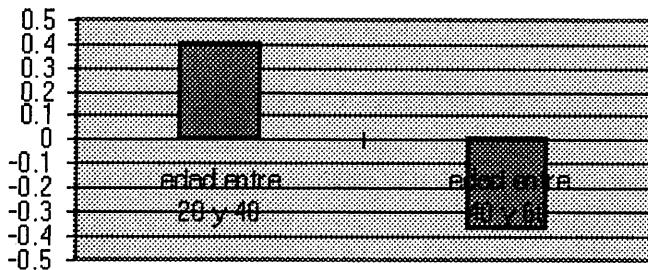


Figura 2. Correlación entre síntomas X y Y como función de la edad

El coeficiente de correlación es un concepto que nos ayuda a presentar un caso en el que aparece la paradoja de Simpson. La figura 2 presenta el hecho de que dos síntomas pueden coincidir frecuentemente en la edad más joven (20-40 años) y no coincidir para los pacientes mas avanzados de edad. Al no considerar la edad de los pacientes, uno puede llegar a la falsa conclusión de que no hay ninguna correlación entre X y Y.

Se puede decir que la paradoja de Simpson nos da una razón profunda de la necesidad de analizar triadas de variables. Cuando tenemos muchas variables (atributos) en la base de datos, casi siempre aparece la situación en la que no hay interdependencia (es decir una noción más general que la correlación) entre X y Y, en general; pero si existe interdependencia significativa entre estos síntomas para ciertos subconjuntos de pacientes. Tales situaciones pueden tener mucha importancia científica o práctica. Importancia más interesante aún si estas interdependencias (de carácter contrario en diferentes niveles de Z) no eran esperadas antes de hacer la investigación.

3. DISTANCIAS

Cuando los síntomas pueden tomar más de dos valores posibles (es decir que existen mas “valores” que los de simple existencia o no para el síntoma) aparecen problemas con el uso del coeficiente de correlación. Si por lo menos una variable de la pareja X y Y es de tipo nominal, no se puede definir un coeficiente de correlación que pueda tomar valores negativos (Goodman, Kruskal, 1954). Por esto hay dificultades para presentar evidencia para posibles triadas (X, Y, Z) con la dependencia del tipo de la paradoja de Simpson.

Uno puede considerar una solución diferente. Se puede definir la distancia que está midiendo la diferencia entre estructuras de interdependencia para X y Y. La diferencia se refiere a dos niveles de edad. Por ejemplo, si las estructuras de dependencia entre X y Y para niveles de edad 20-30 años y 30-40 años son similares, entonces la distancia toma un valor pequeño. Si los pacientes de edad 50-60 tienen una estructura de dependencia entre síntomas X y Y muy diferente a la de edades más jóvenes entonces ambas distancias: (20-30 contra 50-60 y 30-40 contra 50-60) toman valores grandes.

4. FACTORIZACIONES

En la literatura sobre el análisis con componentes estadísticos, donde hay una variable categórica, la noción de ‘factorización’ tiene importancia . Esta noción ha contribuido al desarrollo de teoría y aplicaciones de tales sistemas, e. g. Dempster (1967, 1968), Kyburg Jr (1987), Pearl (1988), Shafer y Shenoy (1990), Pearl (1990), Dubois y Prade, (1992), Spirtes et al. (1993), Spiegelhalter et al. (1994), Smets y Kennes (1994) y Klopotek (1995). Esta noción se refiere a la factorización de distribuciones probabilísticas multivariadas.

La razón para que dicha factorización tenga fuerte impacto teórico y práctico es que todo tipo de independencias estocásticas, desde la más simple (pero muy importante) de dos variables hasta otras más sofisticadas con muchas variables involucradas, sí se pueden escribir con una forma de factorización apropiada.

Existe la otra razón, puramente estadística, que también “promueve” la noción de factorización. Dentro de la metodología de análisis de tablas de contingencia son muy importantes los modelos loglineales. La metodología de estos modelos fue desarrollada entre otros por Bishop et al. (1975), Crowder (1978), Breiman et al. (1984), McCullagh y Nelder (1989), Hastie y Tibshirani (1990), Draper (1995) y Lee y Nelder (1996).

Ahora vamos a presentar en una forma específica dos modelos loglineales, ambos para tres variables discretas, subrayando el conexión que tienen estos modelos con la noción de factorización.

El primer modelo tiene la siguiente forma:

$$\log E[N_{ijk}] = m + dx_i + dy_j + dz_k + dxy_{ij} \quad (1)$$

donde: E - operator de esperanza (estadística),

N_{ijk} - elemento de la tabla de contingencia tres-dimensional como variable aleatoria,

i, j, k - son índices que toman valores correspondientes a sus respectivos síntomas: X, Y, Z,

m - una constante,

d - son los contrastes que representan influencias de respectivas variables-síntomas o interacciones entre estas variables (todas las posibles sumas por índices de los contrastes son iguales cero).

Si la matriz $[N_{ijk}]$ tiene la distribución multinomial, producto-multinomial o Poisson, entonces la estadística ji-cuadrada, basada sobre estimaciones de máxima verosimilitud de los parámetros en (1) tiene una forma cerrada (Bishop et al., 1975). Por esto es muy fácil verificar si el modelo (1) es adecuado para los datos que están en la base de datos.

Al establecer a través de los cálculos que (1) es cierto para la triada (alguna) de variables X, Y y Z, entonces se puede decir que la siguiente factorización quedó empíricamente confirmada.:

$$P(X, Y, Z) = P^Z(X, Y) * P^{XY}(Z), \quad (2)$$

donde: P es la distribución conjunta,

P con índice es una distribución marginal sumando por variables que están en el índice.

El segundo modelo loglineal tiene la forma:

$$\log E[N_{ijk}] = m + dx_i + dy_j + dz_k + dxz_{ik} + dyz_{jk}. \quad (3)$$

Al establecer la validez de este modelo con datos de alguna triada X , Y , Z , uno tiene la siguiente factorización:

$$P(X, Y, Z) = P^Y(X, Z) * P^X(Y, Z). \quad (4)$$

5. FACTORIZACIONES PARA VARIABLES DE TIPO DEMPSTER-SHAFER

El formalismo de tipo Dempster-Shafer introduce nociones que pretenden dar una alternativa pragmática para los dos tradicionales formalismos estadísticos: frecuentista y Bayesiano. La literatura sobre el “nuevo” formalismo incluye: Dempster (1967, 1968), Shafer (1976), Shafer y Shenoy (1990), Pearl (1990), Smets y Kennes (1994), Klopotek (1995).

En el aspecto de tipo de variables (atributos), dentro de formalismo de Dempster y Shafer se considera únicamente el caso discreto, finito, que por cierto para los fines de este trabajo es suficiente. Las variables de este tipo ofrecen la posibilidad de que un paciente pueda tener dos valores de esta misma variable a la vez. Por ejemplo el paciente puede tener dos síntomas de este misma naturaleza o dos diferentes diagnósticos. Esta posibilidad no puede ser realizada dentro de estadística clásica. Hay que pagar el cierto precio por la posibilidad de tener dos o más valores para el mismo paciente. Vamos a describir muy brevemente este precio en dos niveles.

En el nivel de modelo sí es posible definir lo que es lo mínimo: las factorizaciones para triadas de variables. En Matuszewski y Klopotek (1994, 1995) se han definido factorizaciones análogas a las (2) y (4) aunque en forma menos clara. No fue posible definir coeficientes de correlación pero para el modelo análogo a (2) uno puede definir un tipo de distancia. Por ejemplo, se puede definir distancia entre estructuras de interdependencia de dos síntomas X y Y para dos diagnósticos que son valores de variable Z . Es posible de que el paciente tenga dos (o más):

- posibles síntomas de tipo X conjuntamente (por ejemplo tiene el dolor de pierna derecha y izquierda a la vez),
- posibles síntomas de tipo Y conjuntamente,
- posibles diagnósticos de tipo Z conjuntamente (Z ya no es la edad de paciente porque ésta es única).

El nivel del modelo no puede por sí mismo proporcionar los métodos efectivos para establecer evidencia empírica para validez de modelo apropiado. En términos estadísticos hay que establecer en el otro nivel algún tipo de prueba para poder calcular la significancia estadística (p-valor) del modelo. Luego, calculando el p-valor de la

estadística de prueba uno puede tener evidencia en favor o en contra de modelo para los datos que se tiene en la base de datos.

En Klopotek, Matuszewski y Wierzchon (1996) hay una descripción del procedimiento estadístico que tiene el propósito de calcular p-valores que miden la validez de modelo análogo a (2). Hemos utilizado una generalización del método descrito en Conaway (1994).

Es más difícil todavía establecer validez estadística para el modelo con distancia entre diagnósticos, pues en este caso deben utilizarse resultados del tipo de los de Thomson (1995).

REFERENCIAS

- Bishop, Y.M.M., Fienberg, S. E., Holland P. W. (1975). *Discrete Multivariate Analysis: Theory and practice*. MIT Press, Cambridge.
- Breiman L., Friedman J.H., Olshen R.A., Stone C. J. (1984). *Classification and regression trees*. Wadsworth, Belmont.
- Conaway M.R. (1994). Causal nonresponse models for repeated categorical measurements. *Biometrics*, **50**, 1102-1116.
- Crowder M. J. (1978). Beta-binomial for proportions, *Applied Statistics*, **27**, 34-37.
- Dempster A. P. (1967). Upper and lower probabilities induced by a multi-valued mapping. *Ann. Math. Stat.*, **38**, 325-339.
- Dempster A. P. (1968). A generalization of Bayesian inference, *J. R. Stat. Soc., Ser. B*, **30**, 205-247.
- Draper D. (1995). Inference and hierarchical modelling in the social sciences (with discussion). *J. Educ. Behav. Statist.*, **20**, 115-147 & 228-233.
- Dubois D., Prade H. (1992). Evidence, knowledge and belief functions. *Int. J. of Approximate reasoning*, **6**, 295-319.
- Gelman A., Carlin J. B., Stern H. S., Rubin D. B. (1995). Bayesian data analysis. *Chapman and Hall*, London.
- Goodman J. A., Kruskal W. H. (1954). Measures of association for crossclassifications. *J. Amer. Stat. Ass.*, **49**, 732-64.
- Hastie T. J., Tibshirani R. J. (1990). Generalized Additive models. *Chapman and Hall*.
- Klopotek M. A. (1995). Interpretation of belief function in Dempster Shafer theory. *Found. Comp & Dec. Sc.*, **20**, 289-306.
- Klopotek M. A., Matuszewski A., Wierzchon S. T. (1996). Overcoming negative-valued conditional belief functions when adapting traditional knowledge acquisition tools to Dempster-Shafer theory. *Proc. CESA '96 IMACSMulticonference*, **2**, 948-953.
- Kyburg jr H. E. (1987). Bayesian and non-Bayesian evidential updating. *Art. Intelligence*, **31**, 271-293.
- Lee Y., Nelder J. A. (1996). Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc. Ser. B*, **58**, 619-678.
- McCullagh P., Nelder J. A. (1989). *Generalized linear models*. 2nd ed., Chapman and Hall, London.

- Matuszewski A., Kłopotek M. (1994). What does a belief function believe in? *ICS PAS Reports* No 758, Warsaw.
- Matuszewski A., Kłopotek M. (1995). Factorization of Dempster-Shafer belief functions based on data. *ICS PAS Reports* No 798, Warsaw.
- Pearl J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan & Kaufmann.
- Pearl J. (1990). Bayesian and belief-function formalisms for evidential reasoning: a conceptual analysis. In: G. Shafer, J. Pearl (eds), "Readings in uncertain reasoning", Morgan Kaufman Pub. Inc., San Mateo CA, 540-569.
- Shafer G. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton.
- Shafer G., Shenoy P. (1990). Axioms for probability and belief function propagation. In: R. D. Shachter (ed.), *Uncertainty in artificial intelligence 4*. North-Holland, New York, pp. 169-198.
- Smets P., Kennes R. (1994). The transferable belief model. *Artificial Intel.*, **66**, 191-234.
- Spiegelhalter D., Thomas A., Best N., Gilks W. (1994). BUGS, reference manual, version 0.3. *Medical Research Council Biostatistics Unit, Cambridge*.
- Spirites P., Glymour C., Scheines (1993). Causation, prediction and search. *Lecture Notes in Statistics*, **81**, Springer Verlag.
- Shenoy P. (1994). Conditional independence in valuation-based systems. *Int. Journal of Approximate Reasoning*, **10**, 203-234.
- Thomson P. C. (1995). A hybrid paired and unpaired analysis for the comparison of proportions. *Statistics in Medicine*, **14**, 1463-1470.

La Influencia de Quetelet en México: una Polémica con un Océano de por Medio

LETICIA MAYER

IIMAS-UNAM

1. PRINCIPALES CONTRIBUCIONES

El trabajo de Quetelet representa el primer escalón racionalmente organizado del desarrollo de la estadística. Quetelet promovió la fundación y regulación de varias asociaciones de estadística. Entre ellas podemos incluir: *Royal Statistical Society of London*, *The Statistical Section of the British Association*, entre otras. Aunque su formación fue de astrónomo, su reputación internacional se desarrolló principalmente ligada a sus trabajos estadísticos. Desde 1826 comenzó una correspondencia con los diversos buroes estadísticos de Europa. En sus primeros trabajos estadísticos exploró los problemas relacionados con la natalidad y la mortalidad particularmente en Francia, sin embargo nunca pudo obtener datos completos, ni realizar el censo de ese país.

Quetelet hizo dos aportaciones importantes en el avance estadístico: la formulación del concepto de hombre promedio u hombre tipo y la prueba de distribuciones. Desde 1827 hasta 1835 examinó las frecuencias que se encontraban en las tablas secuenciales y su distribución gráfica (Stigler, 1986). Con algunas excepciones sólo comparó dos características al mismo tiempo: los nacimientos y las muertes, cruzandolos por mes, temperatura, altitud, ciudad e incluso hora del día. También investigó la mortalidad por edad, profesión, localidad y estación del año. Además consideró la mortalidad en cárceles y hospitales. Tomó en cuenta otros atributos humanos: la altura, el peso, el crecimiento y la fuerza. Parte de su interés estaba en extender las características físicas a cualidades morales. De aquí surgió la necesidad de crear estadísticas de alcoholismo, insalubridad, suicidio y crimen. En 1835 publicó su escrito más conocido: *Sur l'homme et le développement de ses facultés, ou essai de physique social*. El cual fue traducido al inglés en 1842 en Escocia. En esta publicación desarrolló ampliamente su concepto de "hombre tipo". La idea de un hombre tipo o promedio cautivó la imaginación de los científicos desde 1835. El concepto encerraba, aparentemente, la idea del hombre común de una manera precisa y científica. Este concepto sirvió como elemento válido de la estadística y congenió con el pensamiento político del siglo XIX. Para Quetelet el "hombre promedio" era un individuo que no pertenece a ninguna época, ni posición social y las cualidades del hombre tipo representan lo que es grande, bueno y hermoso. El "hombre tipo" se convirtió en un símbolo concreto de la sociedad. Pero más que eso fue la forma de comenzar a crear una física social, la puerta de entrada a una ciencia social matemática.

2. LA CRIMINALIDAD UNA DESVIACIÓN MORAL

Con el recuento de los seres humanos y sus hábitos la sociedad llegó a ser objeto de la estadística. Pero las tablas secuenciales llevaban consigo las connotaciones de lo normal y las desviaciones a la norma. La mayor parte de las regularidades, semejantes a las leyes, fueron recolectadas en relación con lo que se consideraba desviaciones morales: suicidio, crimen, vagancia, locura, prostitución y enfermedad. La criminalidad tenía una regularidad tal que resultaba imposible atribuirla al azar. Incluso en una carta que Guerry le escribió a Quetelet en 1832 hizo la siguiente observación: “nos vemos obligados a reconocer que los hechos del orden moral están sujetos, lo mismo que los del orden físico, a leyes invariables” (Hacking, 1991). De 1821 a 1829, aunque con cierta irregularidad, apareció una publicación con datos estadísticos sobre la criminalidad, *Recherches statistique sur la ville de Paris et le département de la Seine*. Para 1830 las regularidades que se observaron sobre crímenes, suicidios, prostitución, vagancia y alcoholismo no dejaron de llamar la atención de los científicos.

Las regularidades que presentaban las tablas secuenciales de la estadística dieron lugar a una reflexión que podría considerarse más bien una metaciencia. La constancia en las desviaciones llevó a Quetelet a presentar una curva humana similar a la de la “ley de los errores” que se desarrolló en astronomía. Quetelet, al introducir parámetros de la astronomía a la sociedad, les dio un valor de medición y cuantificación que antes no tenían. Además a la ficción analítica del “hombre tipo” le confirió un valor real al medir y contar propiedades físicas, pero lo que es más importante, al cuantificar características morales. El “hombre tipo” se definió de acuerdo con su origen nacional o bien racial. Dejó de concebirse un pueblo únicamente de acuerdo con su geografía, lengua, historia o religión. Ahora también lo caracterizaban las cualidades antropomórficas de sus habitantes.

Los científicos no sólo descubrían leyes estadísticas sobre la moral, sino que pensaban que había una explicación a ellas en la naturaleza que tenía que ver con el determinismo. En este sentido el “hombre tipo”, que se definió no universalmente, sino nacionalmente, implicó una especie de raza que era buena o mala determinada en forma natural. A partir de este momento se empezaron a exaltar las características de los hombres dependiendo de su nacionalidad.

3. EL MEXICANO COMO “HOMBRE TIPO”: UNA POLÉMICA A DISTANCIA

En 1839 se publicó el primer número del *Boletín*, órgano informativo del Instituto Nacional de Geografía y Estadística. Esta revista especializada editó un trabajo de José Gómez de la Cortina intitulado “Población” (Gómez de la Cortina, 1839). Este fue el primer artículo de estadística moderna que se publicó en México. El trabajo abordó cuatro temas, todos relacionados con la problemática de la estadística: los censos, el balance de los sexos, las estadísticas de la moral y el problema del analfabetismo. Con las estadísticas de la moral

el conde de la Cortina intentó demostrar que la población desviada de México era una minoría comparada con la de países como Francia. Esta demostración apuntaba a que la población, considerada como la verdadera riqueza de las naciones, en México casi no registraba desviaciones. En otras palabras era prácticamente perfecta y esto estaba determinado en forma natural.

El recuento de causas criminales fue conocido en México desde la memoria del estado de Guanajuato de 1826. Sin embargo la preocupación por la criminalidad y la forma de controlarla venía en aumento; lo innovador en el artículo de Gómez de la Cortina fue el análisis y las conclusiones a las que llegó. El autor fue gobernador del Distrito Federal entre 1835 y 1836, con lo que tuvo la posibilidad de hacer una serie de observaciones y cuantificaciones personales con base en las cuales elaboró sus tablas de delitos en la ciudad de México. Estos estados, que como se ha dicho, fueron ejecutados con toda la exactitud y escrupulosidad posible, dan lugar a las observaciones siguientes.

1a. Siendo 202 los criminales de este período, en una población de 205.430 habitantes resulta 1 99/101, o cerca de dos de los primeros, por cada 1016 de los segundos, o lo que es lo mismo, menos de un criminal por cada 508 habitantes, debiendo notarse que en las ciudades populosas, y con especialidad en las capitales, abundan más los alicientes al crimen, la gente ociosa y las ocasiones de corrupción.

2a. Siendo 29 el término medio que corresponde a cada mes, en los mismos estados, resulta menos de un criminal por día. En París, por ejemplo, el número de personas encarceladas cada veinticuatro horas por robo, riña y otras infracciones de policía, es de 25 a 30; si se añaden las personas apresadas por delitos de mayor importancia, puede calcularse aquel número en 35 a 40, de lo que resulta que la población de la ciudad de México, apenas más de tres veces menor que la de París, produce un número de delincuentes *más de treinta veces menor* que el que produce la de la capital de Francia (Gómez de la Cortina 1839).

Los datos de Gómez de la Cortina seguramente resultaron elocuentes en su momento. La criminalidad en la ciudad de México era ¡treinta veces menor que la de París! En la primera mitad del siglo XIX, la vagancia, la miseria, la criminalidad, la prostitución fueron motivo de preocupación para la mayoría de los grandes novelistas europeos, basta recordar a Eugenio Sue con Los misterios de París, o Los miserables de Víctor Hugo, o bien las novelas de Dickens. París y Londres representaron el ejemplo de las grandes ciudades llenas de problemas, principalmente la población desviada, las clases peligrosas (Chevalier, 1984). La pequeña comunidad científica mexicana, junto con los burócratas e intelectuales interesados en la criminalidad, conocían las estadísticas de París y el Sena y, al comparar éstas con las de la ciudad de México, es probable que se sintieran reconfortados.

Las demás conclusiones de Gómez de la Cortina siguieron apuntando a un “hombre tipo” excepcional, no sólo por la baja desviación de la norma, sino por las razones mismas de la criminalidad:

3a. De los 202 crímenes que contienen los estados, 138 son contra la propiedad, y 64 contra las personas: por consiguiente resulta 1 de los primeros por cada 1.488 habitantes, y 1 de los segundos por cada 3.209 habitantes; viéndose en el exceso que el

número de los primeros lleva al de los segundos, los efectos de la miseria y del abandono que producen los hábitos adquiridos en las guerras civiles, más bien que la perversidad de una intención dirigida al mal (Gómez de la Cortina, 1839).

Para el autor la mayoría de los delitos no implicaban maldad, sino necesidad. Sólo una tercera parte se cometieron en contra de la persona y dos terceras partes en contra de la propiedad. Gómez de la Cortina confió en la bondad natural de los mexicanos, al grado de dejar su seguridad personal y la de su familia en manos de exdelincuentes aparentemente reformados: Puso de portero a un capitán de ladrones, y le ordenó que permaneciera en la puerta con la obligación de aprehender a cualquiera de sus antiguas amistades que acertara a pasar enfrente de la casa; y de su conducta dependía el que le perdonaran sus fechorías. Otra vez en compañía del mismo individuo, entonces mozo de espuela, se dirigía a su casa de campo con la Condesa, cuando les alcanzó un mensajero que requirió al Conde el inmediato regreso a la ciudad para el arreglo de un urgente e importante negocio. Anochecía, y sin embargo, el Conde, fiado en el pundonor del ladrón, le ordenó conducir a la señora hasta la hacienda, y ella sola, a caballo, y acompañada de este alarmante guía, hizo la jornada sin novedad (Calderón de la Barca, 1959).

La cuarta y quinta conclusiones a las que llegó el autor nuevamente apuntan a una población en la cual la desviación resulta fácil de corregir y encauzar dado que implica, en su mayoría, un solo tipo de delincuente: varón, soltero y de 25 a 40 años. Los datos del autor parecen apuntar a que efectivamente la población de la ciudad de México, comparada con la de París, era mucho más sana moralmente en términos estadísticos del siglo XIX.

Otro de los temas preferidos de las estadísticas de la moral fue la prostitución:

En los padrones que con la mayor escrupulosidad mandó formar el gobierno del Distrito desde Octubre de 1835 hasta Agosto de 1836, aparecen 322 mujeres públicas en la ciudad de México, incluyéndose en este número 53, que sin ser enteramente públicas, o como vulgarmente se dice *callejeras*, sino mantenidas por varios particulares, debió el gobierno considerarlas como pertenecientes a la clase de que se trata. Resulta, pues, una prostituta por cada 637 158/161 habitantes. En París, el año de 1832 se registraron en los asientos de la prefectura de policía 42.699 prostitutas, [...] Resulta, pues, que en la población de París, algo más de *tres veces* mayor que la de la ciudad de México, hay constantemente un número de prostitutas casi sesenta y siete *veces mayor* que en la de esta última ciudad (Gómez de la Cortina, 1839). Para el autor el nuevo elemento apuntaba a lo mismo, la prostitución era muy baja en comparación a la de ciudades como París. En México la desviación a la norma moral por parte de los varones era menor y factible de controlar, lo mismo sucedía con las mujeres -cuyos datos de criminalidad en todos los países eran más bajos que los de los hombres- pero además la prostitución ni siquiera tenía punto de comparación con la de París. Hombres y mujeres poseían costumbres más sanas en México que en Francia.

El suicidio que fue tema de debate en Europa durante todo el siglo XIX, hasta culminar con el estudio de Durkheim, en México no se tocó. Las estadísticas de criminalidad no lo registraron. Bien puede ser, como lo apuntó Gómez de la Cortina, que

esta desviación fue prácticamente desconocida en México, o bien, porque al ser algo sancionado por la religión católica, los familiares de suicidas procuraron ocultarlo.

Otros crímenes que el autor consideró poco comunes en México fueron: envenenamiento, asesinatos pagados, asesinato con premeditación y sacrilegio:

Son desconocidos entre nosotros los *asesinatos pagados*, y muy raros también aquellos en que se echa de ver el grado a que puede llegar la perversidad humana, por el refinamiento de las circunstancias con que se premeditan, o con que aumenta la crueldad de la ejecución (Gómez de la Cortina, 1839). Los científicos de la primera mitad del siglo XIX, al acumular datos estadísticos sobre la criminalidad, la prostitución y el suicidio, llegaron a imaginar leyes universales, casi biológicas, que determinaban la conducta moral de los individuos divididos por su origen nacional. Los franceses contaban con un alto porcentaje de población desviada y tendían al suicidio, por el contrario, los mexicanos eran buenos por naturaleza. Atrás de todas estas reflexiones estaba el pensamiento determinista.

Gómez de la Cortina, al patentizar las bondades del pueblo mexicano, lo que quería era salvarlo; demostrarle al mundo, en forma absolutamente científica, que México no sólo contaba con los mejores recursos materiales, como lo había demostrado Humboldt, sino que además su población se acercaba a la perfección moral. Todo esto no con base en las constantes, sino de acuerdo con las desviaciones de la norma. Lo que a primera vista parecía una ingenuidad del conde de la Cortina, plasmada en su documento, analizado éste en el contexto científico del siglo XIX vemos que en verdad respondía a una idea determinista de su época: el azar no podía existir, pues la naturaleza imponía leyes a la sociedad al igual que las leyes físicas de la naturaleza, por lo tanto tenía que haber alguna constante que hacía que el pueblo de México fuera bueno en esencia. La estadística, en estos términos, respondió a la creación del imaginario nacional.

4. QUETELET Y GÓMEZ DE LA CORTINA

Quetelet era sólo tres años mayor que el conde de la Cortina. El primero nació en 1796 y el segundo en 1799. Es posible que ambos personajes se hubieran conocido dado que Gómez de la Cortina vivió en Europa desde 1814 hasta 1832, además de que, durante esa época, fue agregado de la embajada de España en Holanda, Austria, Inglaterra y Francia. Gómez de la Cortina fue un hombre muy culto, un verdadero sabio, como se les llamaba en aquella época. Un literato, matemático y conocedor profundo de la estadística decimonónica. Por su misma formación fue un hombre muy polémico. Son ampliamente conocidos sus desacuerdos literarios con personajes como Lacunza, Rodríguez Galván e incluso Guillermo Prieto y Manuel Payno. Por lo mismo resulta notable la falta de polémica con respecto a los planteamientos estadísticos. Partiendo de esta reflexión propongo que el verdadero debate se daba en forma transoceánica: directamente con Quetelet, Guerry y Villermé. Por desgracia el archivo personal de Gómez de la Cortina, hasta la fecha, se encuentra perdido. Sin embargo, partiendo de su propio artículo de “Población”, es de

suponer que la defensa del mexicano es en realidad una discusión con los europeos, particularmente con Quetelet quien consideraba a los mexicanos, junto con los chinos, grupos de salvajes.

REFERENCIAS

- Calderón de la Barca, F. (1959). *La vida en México*. México: Editorial Porrúa.
- Chevalier, L. (1984). *Masses, basses labourieuses et classes dangereuses*. París: Hachette.
- Gómez de la Cortina, J. (1839). Población. *Boletín del Instituto Nacional de Geografía y Estadística*, No. 1, pp. 13-37.
- Hacking, I. (1991). *La domesticación del azar*. España: Gedisa.
- Stigler, S. (1986). *The History of Statistics*. The Belknap Press of Harvard University Press.

Estimación Bayesiana de Proporciones Condicionales

MANUEL MENDOZA y MANUEL MEZA

I.T.A.M.

1. INTRODUCCIÓN

El problema de estimar proporciones es uno de los más conocidos en la literatura estadística. Si se cuenta con una muestra aleatoria de la población de interés, la solución puede obtenerse sin mayores complicaciones tanto con un enfoque frecuentista como Bayesiano. Las soluciones que estos dos enfoques producen no son las mismas en general pero asintóticamente coinciden. Si el análisis Bayesiano se lleva a cabo con poca información inicial y se utiliza alguna de las funciones de pérdida más comunes, los resultados son muy similares aun para tamaños muy reducidos, excepto en casos extremos. Este es un caso donde la solución frecuentista satisface prácticamente todos los criterios con que habitualmente se juzga la bondad de sus estimaciones. Por otra parte, bajo el enfoque Bayesiano la solución cumple, como siempre, con el único criterio de optimalidad aplicable. El propósito de este reporte es ilustrar la situación que se presenta cuando el problema se aborda a partir de una muestra aleatoria que procede no de la población de interés sino de una población mayor que la contiene. En estas circunstancias, la solución Bayesiana sigue siendo óptima pero no puede asegurarse lo mismo de la solución frecuentista, aun cuando puntualmente los resultados siguen siendo muy similares, como en el caso más simple.

La comparación de los métodos frecuentistas y Bayesianos ha sido objeto de investigación durante un buen número de años y así como existen trabajos donde esa comparación se lleva a cabo desde una perspectiva formal, insistiendo en la naturaleza axiomática de la estadística Bayesiana que garantiza su coherencia metodológica, también se han publicado contribuciones en las que se examinan en detalle algunos problemas o aplicaciones concretas donde los resultados particulares de ambos enfoques se comparan para establecer sus ventajas relativas en la práctica. Aquí, para ilustrar las diferencias que existen entre los dos enfoques, se considera una situación en la que la aplicación de ambos enfoques produce resultados similares pero en donde, dependiendo de la naturaleza del muestreo, la solución frecuentista puede perder su condición de solución óptima mientras que tal efecto no se presenta bajo el enfoque Bayesiano.

2. EL PROBLEMA ORIGINAL

Considere la situación en la que cada uno de los elementos de una población puede ser clasificado en una y sólo una de dos categorías (A y B) y θ es la proporción desconocida de elementos que pertenecen a la primera categoría. El problema, en su versión más simple, consiste en estimar el parámetro θ a partir de una muestra de individuos de la población de

interés. Es decir, a partir de una muestra aleatoria $\mathbf{Z} = (X_1, X_2, \dots, X_n)$ de una variable *Bernoulli* con función de probabilidad

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}; \quad x = 0,1. \quad (1)$$

Desde un punto de vista frecuentista, los métodos habituales de estimación (*momentos y máxima verosimilitud*, por ejemplo) conducen a la estimación

$$\hat{\theta} = n_1 / (n_1 + n_2) = n_1 / n \quad (2)$$

donde $n_1 = X_1 + \dots + X_n$ y $n_2 = n - n_1$. Así, el estimador resulta la proporción muestral. Esta solución es muy simple y tiene un indudable atractivo intuitivo. Más aun, reúne propiedades (*insesgamiento, varianza mínima, etc.*) que, desde la propia perspectiva frecuentista, permiten considerarla óptima. La solución Bayesiana, por su parte plantea el problema como uno de decisión y requiere la especificación de una distribución inicial o *a priori* para θ lo mismo que una función de pérdida. En cualquier caso a partir de estos elementos, y con la información muestral, se obtiene la solución de *pérdida esperada mínima a posteriori* que, por construcción, es óptima. Si, se utiliza una distribución inicial (conjugada) Beta($\theta|\alpha, \beta$), la correspondiente final resulta

$$p(\theta | \mathbf{Z}) = \text{Beta}(\theta | \alpha + n_1, \beta + n_2). \quad (3)$$

Si además, se adopta una función de pérdida cuadrática, el estimador de θ está dado por

$$\tilde{\theta} = E(\theta | \mathbf{Z}) = \frac{\alpha + n_1}{\alpha + \beta + (n_1 + n_2)}. \quad (4)$$

Como se puede observar, los estimadores frecuentista y Bayesiano no coinciden aunque son asintóticamente equivalentes y en general producen valores estimados muy similares si α y β son sensiblemente menores que n_1 y n respectivamente. Al respecto es interesante notar que si se utiliza la idea de análisis de referencia (Bernardo 1979) para describir una estado de ignorancia relativa inicial la correspondiente distribución *a priori* es precisamente una Beta con parámetros $\alpha = 1/2, \beta = 1/2$ de manera que en ese caso, la distribución final es una Beta($\theta | n_1 + 1/2, n_2 + 1/2$) y salvo resultados muestrales extremos, los estimadores frecuentista y Bayesiano son prácticamente iguales. Así, cuando en el enfoque Bayesiano se utiliza un distribución de referencia y una pérdida cuadrática, los estimadores $\hat{\theta}$ y $\tilde{\theta}$ son, en la práctica iguales y además, y muy importante, óptimos respecto a sus correspondientes criterios. Por supuesto, si se trata de establecer la incertidumbre asociada a estas estimaciones es importante mencionar que los dos enfoque proceden de manera distinta. Para el enfoque Bayesiano toda la información relevante sobre θ está contenida en la

distribución final y a partir de esta distribución es posible obtener intervalos de *probabilidad* final para θ . Por otra parte, con el enfoque frecuentista, es necesario establecer la distribución muestral de $\hat{\theta}$. Para tamaños de muestra finitos, $\hat{\theta}$ se distribuye como un múltiplo de una variable Binomial y asintóticamente, es Normal. En cualquier caso, los dos primeros momentos son de cálculo inmediato y se puede proceder a producir los conocidos intervalos de confianza para θ .

3. PROPORCIONES CONDICIONALES

Consideré ahora el caso en que la muestra aleatoria de individuos no se extrae de la población de interés sino de una población mayor que contiene a ésta. Ahora, un individuo seleccionado en la muestra puede no pertenecer a la población de interés (categoría C) y sólo en caso contrario tiene sentido clasificarlo en una de las dos categorías previas (A y B). Esta situación aparece de manera muy natural en una variedad de aplicaciones. Por ejemplo, si se realiza una encuesta de preferencias electorales para determinar la proporción de lectores que votarán en favor del candidato A en una elección futura, es perfectamente posible, y razonable, que se utilice como marco de muestreo el padrón electoral correspondiente. Sin embargo, si se obtiene una muestra de ciudadanos a partir del padrón, puede ocurrir que algunos de los encuestados hayan decidido no votar o se encuentren en el proceso de decisión correspondiente. En cualquier caso la muestra no procede de la población de los electores que efectivamente ejercerán su derecho al voto sino, de una población que contiene a la de los votantes efectivos.

Formalmente, se cuenta con una muestra aleatoria $\mathbf{Z} = (X_1, X_2, \dots, X_n)$ con función de probabilidad conjunta

$$p(\mathbf{Z}|\underline{\theta}) = \theta_1^{n_1} \theta_2^{n_2} \theta_3^{n_3} \quad (5)$$

donde $\underline{\theta}^t = (\theta_1, \theta_2, \theta_3)$; $\theta_1 + \theta_2 + \theta_3 = 1$. De esta manera, θ_1, θ_2 y θ_3 , representan las proporciones, en la población muestreada, de individuos que pertenecen a las categorías A, B y C respectivamente. En estas circunstancias, es claro que únicamente se cuenta con $n_1 + n_2$ elementos en la muestra que provienen de la población de interés y que de ellos, sólo n_1 pertenecen a la categoría A y un estimador intuitivamente razonable podría ser $\hat{\theta}_c = n_1 / (n_1 + n_2)$. Esta expresión es similar a la del estimador $\hat{\theta}$ de la sección 1. Sin embargo, es necesario notar que en este caso $n \neq n_1 + n_2$ y de hecho, $n = n_1 + n_2 + n_3$. Ahora bien, es interesante observar que si se aplican los mismos métodos frecuentistas que se han citado, el resultado es precisamente, como consecuencia directa de la propiedad de invarianza, $\hat{\theta}_c$. Más aun se puede observar que

$$\hat{\theta}_c = n_1 / (n_1 + n_2) = \{n_1 / n\} / \{(n_1 + n_2) / n\} = \hat{p}(A) / \hat{p}(A \cup B) \quad (6)$$

de donde es claro que se está estimando una probabilidad o proporción condicional. De hecho, el parámetro de interés está dado por

$$\theta_c = \theta_1 / (\theta_1 + \theta_2) = p(A) / p(A \cup B) = p(A | A \cup B). \quad (7)$$

Las dificultades aparecen cuando se analizan las propiedades de $\hat{\theta}_c$. Desde una perspectiva frecuentista, este estimador no sólo no satisface los criterios de insesamiento y varianza mínima sino que no existen expresiones explícitas para sus dos primeros momentos. Como una primera consecuencia, la evaluación del nivel de precisión de este estimador es sólo aproximada o, en el mejor de los casos, conservadora. Mucho más importante, este estimador no puede considerarse, desde ningún punto de vista razonable, óptimo. El contraste con el estimador (2) es evidente. Por lo que toca al enfoque Bayesiano, la función de verosimilitud (5) depende no sólo de θ_c sino de alguna otra función ω del vector $\underline{\theta}$. Es oportuno recordar que si bien $\underline{\theta}$ es un parámetro con tres componentes, éstas satisfacen una restricción lineal de manera que, en realidad, el espacio paramétral tiene sólo dos dimensiones. En consecuencia, es conveniente reparametrizar la verosimilitud en términos de θ_c y alguna elección de ω . Una posibilidad que resulta muy conveniente es la $\theta_c = \theta_1 / (\theta_1 + \theta_2)$ y $\omega = \theta_1 + \theta_2$ que conduce al modelo reparametrizado

$$p(\mathbf{Z}|\theta_c, \omega) = \{\theta_c^{n_1} (1 - \theta_c)^{n_2}\} \{\omega^{n_1+n_2} (1 - \omega)^{n_3}\} \quad (8)$$

donde el resultado más relevante es la estructura de factorización de la verosimilitud que se descompone como el producto de una función de θ_c y otra de ω . Precisamente si para este modelo se aplica, de nuevo, la idea del análisis de referencia considerando que θ_c es el parámetro de interés y ω representa un parámetro de ruido, se puede verificar que, como consecuencia particular de la factorización, se tiene que

$$p(\theta_c, \omega) = \text{Beta}(\theta_c | \alpha_1, \beta_1) \text{Beta}(\omega | \alpha_2, \beta_2) \quad (9)$$

donde $\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1/2$. En otras palabras, de acuerdo con esta distribución de referencia, θ_c y ω son independientes *a priori* y la distribución marginal para θ_c es la misma que se obtuvo en (3). Naturalmente, la distribución final conjunta satisface la condición

$$p(\theta_c, \omega | \mathbf{Z}) \propto \{\theta_c^{n_1+1/2} (1 - \theta_c)^{n_2+1/2}\} \{\omega^{n_1+n_2+1/2} (1 - \omega)^{n_3+1/2}\} \quad (10)$$

de donde es inmediato verificar que θ_c y ω son independientes *a posteriori* y, por tanto, la distribución *a posteriori* marginal para θ_c es simplemente una $\text{Beta}(\theta_c | n_1 + 1/2, n_2 + 1/2)$ que coincide con la distribución *a posteriori* para el parámetro de interés que se obtuvo en la

sección 1. En esas condiciones, si se utiliza una función de pérdida cuadrática, el estimador Bayesiano para este caso resulta

$$\tilde{\theta} = E(\theta | \mathbf{Z}) = \frac{n_1 + 1/2}{(n_1 + n_2) + 1} \quad (11)$$

En otras palabras, el procedimiento Bayesiano opera exactamente igual que en el caso más simple, ignorando por completo las observaciones en la muestra que no pertenecen a la población de interés. Un análisis superficial de (6) podría sugerir que lo mismo ocurre con el estimador frecuentista. Esto no es así, si bien es cierto que en esa expresión aparece la cantidad $n_1 + n_2$ jugando el papel del tamaño de la muestra. En la evaluación de la incertidumbre en el proceso de estimación, particularmente para calcular los momentos muestrales de $\hat{\theta}$, se utiliza el tamaño muestral $n = n_1 + n_2 + n_3$, involucrando así a las observaciones que no proceden de la población de interés. De hecho, las dificultades aparecen precisamente debido a que tanto el numerador (n_1) como el denominador ($n_1 + n_2$) se consideran aleatorios.

4. COMENTARIOS FINALES

En este trabajo se ha presentado el problema de estimación de proporciones condicionales que, además de ser común en la práctica, constituye un elemento de contraste entre los enfoques frecuentista y Bayesiano de la Estadística. Con frecuencia y posiblemente con razón, se presenta el argumento de acuerdo al cual es absolutamente injusto juzgar los resultados de un procedimiento de inferencia de uno de los enfoques con los criterios del otro. Aquí, se han juzgado los resultados de la estimación frecuentista de una proporción con sus propios criterios mientras que el procedimiento Bayesiano ha sido juzgado sólo con sus propios criterios. El resultado establece que, dependiendo del marco de muestreo disponible, es decir, según se trate de un problema con proporciones simples o condicionales, los resultados frecuentistas son o dejan de ser óptimos juzgados *con sus propios criterios*. Desde la perspectiva Bayesiana, utilizando también sus propios términos de evaluación, la optimalidad permanece.

REFERENCIA.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B*, **41**, 113-147, (with discussion).

Modelación de Perfiles de Crecimiento Usando Modelos Lineales con Coeficientes Aleatorios

MIGUEL OJEDA SOSA-LANDA H.

LINAE, Universidad Veracruzana

y

NÚÑEZ-ANTÓN V.

*Facultad de Ciencias Económicas y Empresariales,
Universidad del País Vasco*

1. INTRODUCCIÓN

La modelación de curvas de crecimiento es un área de la modelación estadística que ha atraído la atención de estadísticos y biometristas desde hace varias décadas. (Ver para referencias, por ejemplo, Seber, 1984). Se han planteado enfoques diversos, que van desde la consideración de las respuestas para un individuo como un vector de observaciones con una distribución multivariada; así se ha justificado el uso del modelo lineal general multivariado de efectos fijos (Pottoff y Roy, 1964). También se ha enfocado el problema considerando que cada curva se describe a partir de un polinomio de orden k en el tiempo, y entonces la modelación se da en dos fases: (1) modelar cada curva; y (2) modelar los coeficientes asociados a la muestra de curvas. Esta idea fue primeramente propuesta por Rao (1965), y posteriormente una diversidad de contribuciones parten de ella y presentan procedimientos para mejorar la estimación y prueba de hipótesis, así como para permitir la inclusión de covariables tanto en cada ocasión como a nivel del individuo o unidad sobre la que se mide el crecimiento. (Ver Tian et al., 1994, para una revisión). Recientemente enfoques unificadores, como el del modelo lineal jerárquico (Bryk y Raudenbush, 1992), el del modelo de coeficientes aleatorios (Longford, 1993) y el de modelos multinivel (Goldstein, 1995), plantean un marco teórico y recomendaciones metodológicas para realizar la modelación de este tipo de datos de una manera más realista; es decir, considerando todas las complejidades de este tipo de datos.

El enfoque de modelos lineales jerárquicos con coeficientes aleatorios considera que en primera instancia cada individuo en el estudio tiene su propia trayectoria de crecimiento, la cual es modelada por una ecuación individual, donde los coeficientes son aleatorios. En esta ecuación individual se puede incluir como covariables explicatorias al tiempo de observación y transformaciones de él, pero también es posible incluir los datos de otras covariables. Para completar la postulación del modelo se plantea que la variabilidad entre los coeficientes es modelada usando un modelo también lineal, en el cual se incorporan los datos de covariables al nivel del individuo.

En este trabajo se formula un modelo de regresión con coeficientes aleatorios, el cual es una generalización directa de un modelo anterior presentado en Ojeda y Juárez-Cerrillo (1996). Este modelo es mucho más flexible y permite mayor realismo en el proceso de modelación. Para ilustrar la propuesta se presenta la modelación de la tendencia y

variabilidad de una muestra de 74 perfiles de crecimiento temprano para igual número de familias de coníferas, que fueron estudiadas en condiciones de invernadero.

2. EL MODELO

Sea Y la variable de crecimiento y T la variable del tiempo de observación. Sea y_{ij} la j -ésima medición sobre el i -ésimo sujeto, realizada en el tiempo t_{ij} . Sea n el número de sujetos en la muestra y m_i el número de mediciones hechas sobre el individuo i -ésimo. Entonces cada curva de crecimiento individual se puede modelar por la ecuación de regresión:

$$y_{ij} = \beta_{0i} + \sum_{s=1}^S \beta_{si} x_{sij} + e_{ij}; \quad i = 1, 2, \dots, n; j = 1, 2, \dots, m_i \quad (1)$$

donde x_{sij} puede ser t_{ij} o una transformación, o incluso alguna medición sobre alguna otra variable explicatoria X . Los coeficientes aleatorios son β_{is} , los cuales están asociados con cada curva de crecimiento, y los e_{ij} son errores aleatorios no observables correspondientes a cada sujeto en cada observación.

El modelo en (1) se puede escribir en notación compacta como:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \mathbf{e}_i; \quad i = 1, 2, \dots, n \quad (2)$$

donde \mathbf{X}_i es la correspondiente matriz de diseño asociada al i -ésimo individuo, $\boldsymbol{\beta}_i^t = (\beta_{0i}, \beta_{1i}, \dots, \beta_{Si})$ y $\mathbf{e}_i^t = (e_{i1}, e_{i2}, \dots, e_{im_i})$.

Para estudiar la variabilidad en los coeficientes β_i se propone el siguiente modelo

$$\boldsymbol{\beta}_i = \mathbf{W}_i \Gamma + \mathbf{u}_i; \quad i = 1, 2, \dots, n \quad (3)$$

donde $\mathbf{W}_i = \mathbf{I}_{s+1} \otimes \mathbf{w}_i^t$, con $\mathbf{w}_i^t = (1, w_{i1}, w_{i2}, \dots, w_{iq})$, que es el vector de datos para el individuo i -ésimo en las covariables W_1, W_2, \dots, W_q ; el símbolo \otimes denota el producto Kronecker, y $\Gamma = (\gamma_0^t, \gamma_1^t, \dots, \gamma_q^t)$ es el vector de coeficientes fijos, con $\gamma_s^t = (\gamma_{s1}, \dots, \gamma_{sq})$. Al vector $\mathbf{u}_i^t = (u_{0i}, u_{1i}, \dots, u_{Si})$ se le denomina el vector de errores aleatorios de segundo nivel.

Sustituyendo la ecuación (3) en la (2), haciendo $\mathbf{Z}_i = \mathbf{X}_i \mathbf{W}_i$, se obtiene el, así llamado, modelo lineal general mixto:

$$\mathbf{y}_i = \mathbf{Z}_i \Gamma + \mathbf{X}_i \mathbf{u}_i + \mathbf{e}_i; \quad i = 1, 2, \dots, n \quad (4)$$

Las suposiciones asociadas a esta formulación plantean que $E\{e_i\} = 0$, $\text{var}\{e_i\} = \Sigma_i$, $E\{u_i\} = 0$, $\text{var}\{u_i\} = \Omega$. Asimismo se supone que los errores a nivel de individuos, y de ocasiones a individuos están incorrelacionados; esto es, $\text{Cov}\{e_i, e_{i*}\} = 0$, con $i, i^*=1, 2, \dots, n$

3. EL EJEMPLO

Se tiene una muestra de 74 perfiles de crecimiento de igual número de familias de coníferas, que fueron determinados en base a 10 plantas por familia, medidas en su altura (Y) durante dos meses, tomando mediciones semanales, pero algunas familias no fueron medidas en la última semana, lo que produjo perfiles con desigual número de ocasiones. Dado que la variabilidad intraplanta no resultó significativa en análisis preliminares y en razón de que a los investigadores en genética les interesa estudiar la tendencia y la variabilidad del crecimiento entre familias, se decidió modelar los perfiles. Para mayores detalles sobre los datos ver Mendizabal-Hernández (1995).

Siguiendo la estrategia general se realizaron los análisis descriptivos y exploratorios. En la Figura 1 presentamos las distribuciones de crecimiento. De los análisis descriptivos se decidió realizar ajustes por separado de polinomios de orden 3, evaluándose buenos ajustes. Sin embargo la familia de modelos de Horel (Daniel y Wood, 1980; páginas 22-23) dio mejores resultados al realizar ajustes por separado. Así el modelo propuesto para el perfil de crecimiento fue finalmente:

$$\ln(y_{ij}) = \beta_{0i} + \beta_{1i} \ln(t_{ij}) + \beta_{2i} (t_{ij}) + e_{ij}$$

donde y_{ij} es la observación en el tiempo t_{ij} , en la ocasión j -ésima ($j=1, 2, \dots, m_i$) de la familia i -ésima ($i=1, 2, \dots, 74$).

Inicialmente, usando el ML3E (Prosser et al., 1990), se realizó un ajuste sin considerar covariable alguna al nivel de perfil; es decir, se consideraron como modelos de nivel 2 a:

$$\begin{aligned}\beta_{0i} &= \gamma_{00} + u_{0i} \\ \beta_{1i} &= \gamma_{10} + u_{1i} \\ \beta_{2i} &= \gamma_{20} + u_{2i}\end{aligned}$$

Se obtuvieron los residuos de segundo nivel; es decir, $\hat{u}_{0i}, \hat{u}_{1i}$, y \hat{u}_{2i} y estos se estudiaron gráficamente contra diferentes propuestas de variables explicatorias, encontrando que $W = (Y_2 - Y_1) / Y_1$, donde Y_1 y Y_2 son el crecimiento en la primera y segunda semana respectivamente, que denominamos “crecimiento temprano”, resultó ser una buena variable explicatoria a nivel del perfil. Así entonces, las ecuaciones al segundo nivel fueron finalmente:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}w_i + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}w_i + u_{1i}$$

$$\beta_{2i} = \gamma_{20} + \gamma_{21}w_i + u_{2i}$$

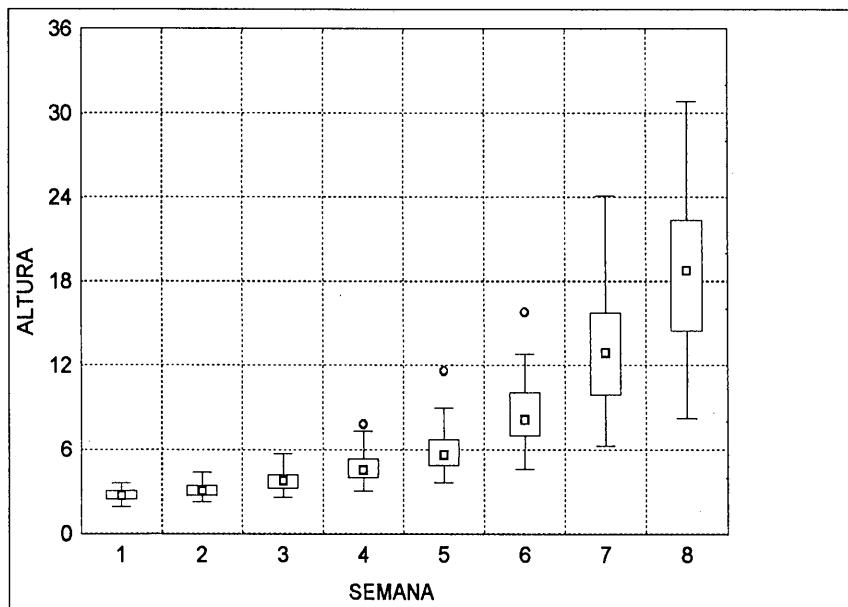


Figura 1. Distribución de los perfiles de crecimiento de las 74 familias.

De los resultados obtenidos del ajuste se obtuvo el siguiente modelo estimado:

$$\log(y_{ij}) = 0.530 - 0.694 \log(t_{ij}) + 0.420t_{ij} + 0.006w_i + 0.011w_i \log(t_{ij})$$

Por otro lado la matriz de componentes de varianza y covarianza, con sus respectivos errores estándar en paréntesis, resultó:

$$\hat{\Omega} = \begin{bmatrix} [0.0232] & & \\ (0.0044) & & \\ & [0.0121] & 0.0020 \\ & (0.0031) & (0.0039) \\ & [-0.0053] & [-0.0031] & [0.0028] \\ & (0.0014) & (0.0016) & (0.0007) \end{bmatrix}$$

de lo que se sigue que la variabilidad entre perfiles de crecimiento es significativa (los coeficientes están marcados en la matriz) en todos los sentidos, excepto en la varianza del coeficiente a la variable $\log(T)$.

REFERENCIAS

- Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publicatios: Thousand Oaks, California, USA.
- Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*. Wiley, New York.
- Goldstein, H. (1995). *Multilevel Statistical Models* (2nd. Edition). Halsted Press: New York.
- Longford, N. T. (1993). *Random Coefficient Models*. Oxford University Press: New York.
- Mendizabal-Hernández, L. C. (1995). Evaluación de la progenie de huertos semilleros de Pinus Patula. *Tesis de Licenciatura*, Facultad de Biología, Universidad Veracruzana, Xalapa Ver., México.
- Ojeda, M. M. and Juárez-Cerrillo, S. F. (1996). Biplot display for diagnostics in a two-level regression model for growth curves analisys. *Computational Statistics and Data Analysis* (in press).
- Potthoff, R. F. and Roy, S. N. (1964) A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313-326.
- Prosser, R., Rassbash, J. and Goldstein, H. (1990). ML3: Software for Tree Level Analysis. Institute of Educational, University of London, London.
- Rao, C. R. (1965) The theory of least squares when the parameters are stochastic and its application to growth curves. *Biometrics*, **52**, 447-458.
- Seber, G. A. F. (1984) *Multivariate Observations*. Wiley, New York.
- Tian, J. J. Shukla, R. and Buncher, R. (1994). On prediction of future observation in growth curve model. *Statistics in Medicine*, **13**, 2205-2217.

Modificación de un Análisis Bayesiano para Factoriales no Replicados

JORGE OLGUÍN y PATRICIA ROMERO

IIMAS-UNAM

1. INTRODUCCIÓN

En investigaciones industriales es común listar un número grande de variables o factores que se piensa podrían tener algún efecto sobre determinada característica de calidad (variable respuesta) de un producto o proceso. La experiencia ha mostrado que con mucha frecuencia se cumple el “Principio de Pareto”, que, en este contexto, establece que la mayor parte de la variabilidad de la respuesta se debe a un número reducido de factores. En estos casos, los experimentos factoriales no replicados resultan de gran utilidad, pues permiten estudiar simultáneamente un número grande de factores e identificar aquellos cuyos efectos son relevantes. En diseño de experimentos, cuando el principio de Pareto se cumple, se dice que hay “esparcidad de efectos”.

Una vez efectuado el experimento (sin réplicas), comúnmente se tienen estimaciones de m contrastes obtenidas de un diseño ortogonal y el problema consiste en decidir, en ausencia de un estimador del error experimental, que en otros experimentos se obtiene con base en réplicas, cuáles de los contrastes estimados tienen un *tamaño estadísticamente significativo*. Los contrastes que resultan “significativos” se denominan *contrastos activos* y los restantes son considerados *contrastos nulos*.

Box y Meyer (1986) presentaron un procedimiento Bayesiano para el análisis de factoriales no replicados. En este resumen, presentamos una modificación de dicho método así como los resultados de un estudio de Monte Carlo en el que comparamos las características operacionales de la modificación con el método original, al que también nos referiremos como método estándar o B-M.

2. MÉTODO DE BOX Y MEYER

Supóngase que un contraste τ_i ($i = 1, \dots, m$) tiene una probabilidad α de ser activo y $1 - \alpha$ de ser nulo. Los contrastes activos son independientes $N(0, \sigma_\tau^2)$ mientras que los nulos son 0. Sean T_1, \dots, T_m los contrastes estimados y, si fuese necesario, estandarizados, de modo que, dado τ , todos tengan la misma varianza σ^2 . Por consiguiente los T_i pueden representarse como sigue:

$$T_i = \begin{cases} e_i & \text{cuando el contraste } i \text{ es nulo} \\ \tau_i + e_i & \text{cuando el contraste } i \text{ es activo} \end{cases}$$

donde las e_i son variables aleatorias independientes $N(0, \sigma_e^2)$. Haciendo $(\sigma^2 + \sigma_\tau^2)/\sigma^2 = k^2$, T_1, \dots, T_m son observaciones independientes de la mezcla de normales denotada por

$$(1 - \alpha)N(0, \sigma^2) + \alpha N(0, \sigma_\tau^2)$$

El procedimiento consiste en obtener, para cada contraste, la probabilidad posterior de que sea activo. Para facilitar la exposición, definimos las variables aleatorias Bernoullis ϕ_i , $i = 1, \dots, m$ de tal forma que $\phi_i = 1$ si el i -ésimo contraste es activo y $\phi_i = 0$ si no lo es. Aplicando el teorema de Bayes, e integrando para remover la condicionalidad en σ (ver Box y Meyer 1986, Olguín 1994), se tiene que la probabilidad posterior de que el i -ésimo contraste sea activo, dado el vector $T = (T_1, \dots, T_m)$ es

$$\begin{aligned} \Pr[\phi_i = 1 | T] &= \int_0^\infty \Pr[\phi_i = 1 | T_i, \sigma] f(\sigma; T) d\sigma \\ &= \frac{\int_0^\infty \Pr[\phi_i = 1 | T_i, \sigma] f(\sigma; T) d\sigma}{f(T)} \end{aligned} \quad (1)$$

El método requiere de los parámetros α y k . Con base en resultados de varios experimentos reportados en la literatura, Box y Meyer (1986) utilizan $\alpha = 0.20$ y $k = 10$.

3. MODIFICACIÓN PROPUESTA

El método de Box y Meyer tiene características que lo hacen más atractivo que otros métodos; por ejemplo, su flexibilidad en la suposición de esparcidad de efectos así como la posibilidad de utilizar información adicional a la de la muestra de contrastes. Sin embargo, la aplicación de este método en ejemplos reales ha mostrado que, en ciertas situaciones, puede pasar por alto contrastes activos que son fácilmente detectados por otros métodos. Esto sucede principalmente cuando alguno(s) de los contrastes es (son) muy grandes con respecto a otro(s) contraste(s) también activo(s) (ver Olguín, 1994).

Con el propósito de darle mayor robustez y flexibilidad al método, nos hemos propuesto estudiar modificaciones en dos direcciones. Por una parte, consideramos interesante que el experimentador pueda expresar su grado de creencia o ignorancia acerca de los parámetros α y k mediante la asignación de distribuciones *a priori* para estos parámetros; y por otra parte consideramos que el uso de una distribución de colas pesadas como la Student para la modelación de los contrastes activos le puede dar mayor robustez. En este resumen presentamos algunos resultados del análisis de una modificación con base en esta última idea. Utilizando una Student con 3 grados de libertad para la modelación de los contrastes activos, los contrastes estimados T_1, \dots, T_m se suponen observaciones independientes, provenientes de la mezcla de distribuciones denotada por

$$(1-\alpha)N(0, \sigma^2) + \alpha St(0, 3, k^2 \sigma^2 / 3)$$

Las probabilidades posteriores de que cada contraste sea activo se obtienen como se describió en la sección anterior, es decir, utilizando (1) bajo los nuevos supuestos.

4. COMPARACIONES ENTRE EL MÉTODO ESTÁNDAR Y LA MODIFICACIÓN PROPUESTA

4.1 Tasas de Error

Con el propósito de comparar la modificación propuesta con el método estándar, se calibraron ambos métodos de manera que la proporción de falsos positivos por experimento bajo la hipótesis de nulidad de los m contrastes ($m = 7, 15, 31$) fuera de $\gamma = 0.05$ para ambos métodos. Esto se hizo por simulación, obteniendo probabilidades posteriores de ser activo “críticas” que produjeran, para cada método y valor de m , dicha proporción o tasa de error.

4.2 Características Operacionales

Una vez hechos comparables los métodos, se procedió a comparar sus características operacionales, es decir, su funcionamiento bajo la presencia de contrastes activos en un experimento. Con este propósito se definieron los siguientes índices.

Tasa de detección (td).- Es la proporción de contrastes activos detectados por experimento.

Tasa de falsos positivos (fp).- Proporción de contrastes nulos declarados como activos con respecto al número de contrastes activos por experimento.

Se realizó entonces un estudio de Monte Carlo para el caso $m = 7$ para situaciones con $r = 1, 2, 3$ contrastes activos. Para cada valor de r se utilizaron diversos tamaños de los contrastes activos y con técnicas de superficie de respuesta se caracterizaron las superficies de td_j y fp_j , así como de las diferencias $td_2 - td_1$ y $fp_2 - fp_1$, donde $j = 1$ para el método estándar (B-M) y $j = 2$ para el método modificado.

4.3 Resultados

- Caso $r = 1$.

En la figura 1 se presenta una gráfica comparativa de las tasas de detección td de ambos métodos en la presencia de un solo contraste activo cuando éste es mayor que 2σ . Como se puede observar, la modificación presentó una td ligeramente superior a la versión original del método (B-M); la diferencia es mayor para un tamaño del contraste activo en el intervalo $(2\sigma, 5\sigma)$. Para un tamaño del contraste activo menor que 2σ las td presentadas por los dos métodos son prácticamente iguales.

En la figura 2 se muestra una gráfica comparativa de las fp para ambos métodos cuando el tamaño del contraste activo es mayor de 2σ . El funcionamiento del método modificado es ligeramente superior para cualquier tamaño del contraste activo. Las diferencias mayores se presentan para un tamaño del contraste activo de alrededor de 3σ .

- Casos $r = 2$ y $r = 3$.

Para las situaciones de dos y tres contrastes activos presentes, en amplias regiones determinadas por los tamaños de los contrastes activos, ambos métodos presentaron valores

de td y fp muy similares. Sin embargo, en regiones caracterizadas por la presencia de un contraste activo muy grande con respecto al otro (los otros), el método modificado parece superior.

En la figura 3 se presenta la superficie ajustada para la diferencia $Dtd = td_2 - td_1$ entre las tasas de detección de los dos métodos para $r = 2$, para situaciones en las que uno de los contrastes activos está en el intervalo $(2\sigma, 7\sigma)$ y el otro en $(10\sigma, 40\sigma)$. Como se puede apreciar en la gráfica, la superficie es positiva sobre una amplia región.

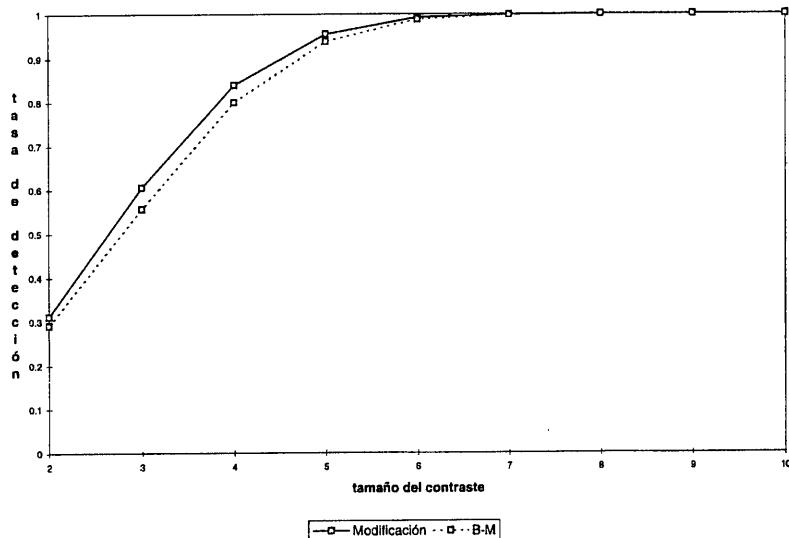


Figura 1. Comparación de la tasa de detección.

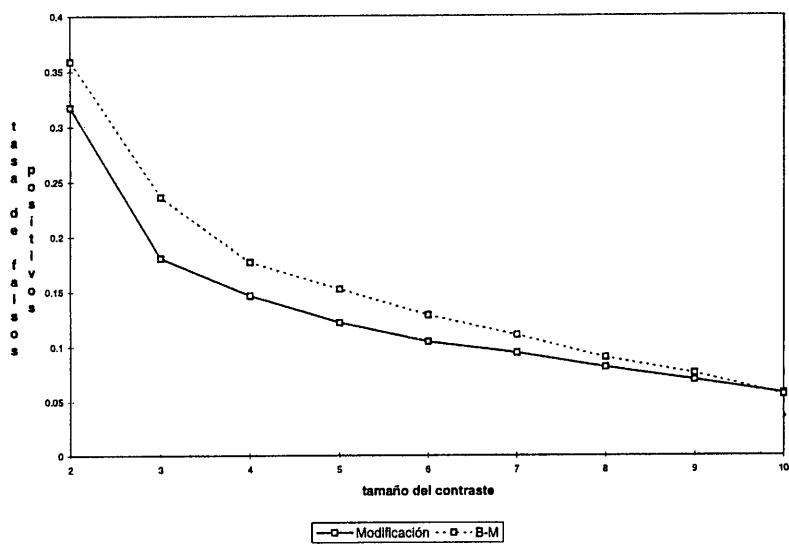


Figura 2. Comparación de la tasa de falsos positivos.

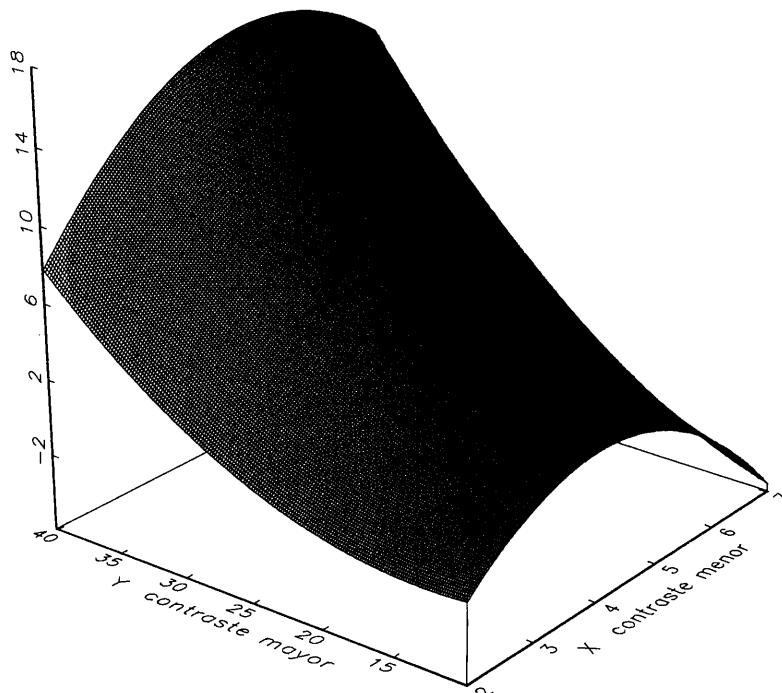


Figura 3. Superficie ajustada de la diferencia en detección entre ambos métodos para $r=2$.

REFERENCIAS

- Box, G. E. P. and Meyer, R. D. (1986). An Analysis for Unreplicated Fractional Factorials, *Technometrics*, **28**, 11-18.
- Olgún, J. (1994). *The Analysis of Unreplicated Factorial Experiments*. Ph.D. Thesis, University of London, University College London, August 1994.

Singularidades en la Distribución Fiducial

FEDERICO O'REILLY y RAÚL RUEDA

IIMAS, UNAM

1. INTRODUCCIÓN

Se considera un problema de inferencia en el que el parámetro $\theta \in \Theta$ (Θ un intervalo) y T es suficiente minimal para θ . Es posible que el contexto anterior sea una sobresimplificación; sin embargo, si θ fuese el parámetro de interés en un problema más complicado, se supone en este trabajo que ya se hizo la correspondiente reducción vía condicionalidad o con suficiencia marginal, etc. (ver por ejemplo, Barndorff-Nielsen, 1978, en lo relativo a “nonformation”). La verosimilitud para θ es proporcional a $g(t; \theta)$, la densidad de T , y siendo $G(t; \theta)$ su función de distribución, toda la información necesaria para inferir sobre θ está contenida en $G(t; \theta)$.

Se supone que existe una relación entre T y θ en el sentido de que los valores grandes de θ implican valores “grandes” de T y valores chicos de θ implican valores “chicos” para T .

Lo anterior significa que los valores grandes para T representan evidencia en contra de valores pequeños de θ . Por ejemplo, para el problema $H_0: \theta \leq \theta_0$ vs. $H_1: \theta > \theta_0$, se dice que “se rechaza H_0 para valores grandes de T ”. Si se fija una hipótesis $H_0: \theta \leq \theta_0$, entonces $1 - G(t; \theta_0) = P[T > t; \theta_0]$, será una cantidad decreciente en t ya que para la hipótesis fija, la evidencia (representada por $T=t$) al hacerse más extrema en contra de H_0 produce un nivel de significancia descriptivo (*p-value*) más chico. Esta aseveración nadie la cuestiona por ser $G(t; \theta_0)$ una función de distribución. De manera análoga, supóngase ahora que T es fijo y que θ_0 se hace variar. Sigue siendo cierto que el nivel de significancia descriptivo para $H_0: \theta \leq \theta_0$ es $1-G(t; \theta_0)$ y debiera ocurrir que al considerar a θ_0' , el correspondiente valor de $1-G(t; \theta_0')$ sea más chico ya que representa una situación en la que la hipótesis y la evidencia distan más. Esto significa que $G(t; \theta)$ debiera ser monótona decreciente como función de θ , para cada t fijo.

Esta propiedad la denominaremos *contrastabilidad monótona*. La noción puede extenderse aún más, si se pide que para casos extremos de “distanciamiento” entre evidencia e hipótesis, exista certidumbre (casi) en la contrastabilidad. Por ejemplo, si θ_0 es fijo y se considera a t convergiendo a su límite superior, el valor de significancia descriptivo converge a cero; sin embargo, no es evidente que cuando t es fijo y θ va convergiendo a su límite inferior, la correspondiente significancia decrezca a cero.

Se dirá entonces que en el problema hay contrastabilidad monótona si

- (i) $G(t; \theta)$ es decreciente como función de θ ;
- y habrá certidumbre en la contrastabilidad para casos extremos si
 - (ii) $G(t; \theta) \rightarrow 0$ si $\theta \rightarrow$ límite inferior y $G(t; \theta) \rightarrow 1$ si $\theta \rightarrow$ límite superior.

2. EJEMPLOS

Ejemplo 1. Si θ es un parámetro de localización para T , entonces $\theta = \mathbf{R}$ y $G(t; \theta) = G_0(t - \theta)$, con G_0 función de distribución.

Evidentemente se satisface que

- (i) $G(t; \theta)$ es decreciente como función de θ ,
- (ii) $G(t; \theta) \rightarrow 0$ si $\theta \rightarrow -\infty$ y $G(t; \theta) \rightarrow 1$ si $\theta \rightarrow \infty$.

Ejemplo 2. Si θ es de escala para T , entonces $\theta = \mathbf{R}^+$ y $G(t; \theta) = G_0(t/\theta)$, con G_0 función de distribución. Por construcción se tiene que (i) y (ii) se satisfacen.

Nótese que (ii) es equivalente a afirmar que en la familia de distribuciones $\{G(t; \theta) : \theta \in \Theta\}$, cuando $\theta \rightarrow$ valor extremo, se obtiene una distribución degenerada. ¿Qué pasa si en los caso límite se obtiene una distribución no degenerada?

Ejemplo 3. En Pedersen (1978), con una observación $X \sim N(\mu, 1)$, se desea inferir sobre μ^2 . Sin entrar en detalles y por ser totalmente equivalente pero más sencillo de ilustrar, se toma $\theta = |\mu|$ y $T = |X|$, en cuyo caso $G(t; \theta) = \Phi(t/\theta) - \Phi(-t/\theta)$ con $t, \theta \geq 0$. Puede verificarse que

- (i) G es monótona decreciente en θ y que,
- (ii) $\lim_{\theta \rightarrow \infty} G(t; \theta) = 0$, pero $\lim_{\theta \rightarrow 0} G(t; \theta) = 1 - 2\Phi(-t) < 1 \quad \forall t \geq 0$.

La interpretación para esta cantidad menor que uno, de acuerdo a lo visto anteriormente, es que para t fijo el nivel de significancia descriptivo para la hipótesis $H_0: \theta = 0$ no se pudo hacer arbitrariamente pequeño, por lo que no puede haber certeza en la contrastación de esta situación extrema.

En un contexto equivalente al problema anterior y que es simplificación al ejemplo dado en Stein (1959), considere a T con distribución ji-cuadrada no central de 1 grado de libertad y parámetro de no centralidad $\theta \in [0, \infty)$. Puede verificarse que $G(t; \theta)$ es monótona decreciente y que $\lim_{\theta \rightarrow \infty} G(t; \theta) = 0$, pero en $\theta = 0$ T tiene como distribución una ji-cuadrada central de 1 grado de libertad.

Ejemplo 4. Para inferir sobre la deriva (*drift*) en un Browniano con base en observaciones de “primera visita” (*first passage time*) Nádas (1973), Seshardi y Schuster (1974) y Patil y Kovner (1976) discuten la prueba óptima y la construcción de intervalos de confianza para el cociente de los parámetros de la correspondiente distribución Gaussiana inversa. Dicho cociente aparece también como un parámetro de interés en O'Reilly y Rueda (1992) ya que indexa las tablas allí construidas para el procedimiento de bondad de ajuste desarrollado. Por otro lado Hsieh (1990) enfrenta también el problema de inferir sobre este parámetro con soluciones tanto clásicas como Bayesianas.

Sin entrar en detalles sea $\{X_1, X_2, \dots, X_n\}$ una muestra aleatoria de una distribución Gaussiana inversa, y sean $\theta = \lambda / \mu$ y $T = \hat{\lambda} / \hat{\mu}$, el cociente de estimadores máximo verosímiles.

La distribución para T está dada por

$$G(t; \theta) = 1 - \frac{e^r r^m}{2^{m-1} (m-1)!} \int_t^\infty e^{-ru} (u^2 - 1)^{m-1} \sum_{i=0}^m \frac{(m+i)!}{(m-i)! i! 2^i r^i} u^{-(i+m)} du$$

donde $t^* = \sqrt{1+1/t}$, $n = 2(m+1)$, y $r = (2m+1)\theta$.

Aunque no se ha exhibido una demostración analítica, se puede constatar numéricamente que

- (i) G es monótona decreciente en θ , y
- (ii) si bien puede demostrarse que $\lim_{\theta \rightarrow \infty} G(t; \theta) = 0$ se tiene que $\lim_{\theta \rightarrow \infty} G(t; \theta) < 1$, $\forall t \geq 0$ al igual que en el ejemplo 3.

De hecho $\lim_{\theta \rightarrow \infty} G(t; \theta)$ corresponde a una función de distribución relacionada con una F (los detalles aparecen en Seshardi y Schuster, 1974).

Al igual que en el ejemplo 3, en este caso, con T observada, a la hipótesis $H_0: \theta = 0$ corresponde un valor de significancia que no pudo hacerse arbitrariamente pequeño aún tomando el valor de θ más chico posible.

3. DISTRIBUCIÓN FIDUCIAL

En un problema de inferencia con G monótona, se define la distribución fiducial para θ (habiéndose observado $T = t$) simplemente como $H(\theta; t) = 1 - G(t; \theta)$, distribución también conocida como *confidencial* o *de significancia*.

Si el problema satisface la condición de certeza (ii), esta distribución es propia; sin embargo, si no satisface esta condición se ha denominado en la literatura como una distribución “defectuosa”, nombre que consideramos incorrecto, pues lo que ocurre es que tiene una masa en el extremo del intervalo.

La construcción de intervalos de confianza puede hacerse con una H aún si posee una masa en un extremo, como en el caso de los ejemplos 3 y 4. Para el ejemplo 4, en Rueda (1988) aparece una comparación de intervalos para θ de la forma (θ^*, ∞) obtenidos de la fiducial, con intervalos de probabilidad final obtenidos con una distribución inicial particular. Ahí se puede observar que para los tamaños de muestra utilizados y para los valores explorados de la estadística T , no existen diferencias entre los dos enfoques, sin haberse detectado dificultades por la presencia de la masa fiducial en el cero.

También relacionado con el ejemplo 4, Hsieh (1990) al comentar la solución clásica en la obtención de sus intervalos (llamada **exacta** por él), no hace mención específica sobre el hecho de que sus intervalos provienen de la distribución fiducial, ni tampoco comenta sobre la masa.

4. COBERTURA FRECUENTISTA

Si, en el ejemplo 3, se desea obtener un intervalo de la forma $[0, \theta^*]$, este puede construirse a partir de la fiducial $H(\theta; t) = 1 - [\Phi(t-\theta) - \Phi(-t-\theta)]$ simplemente resolviendo $H(\theta^*; t) = 1 - \alpha$ si es que la masa fiducial en el cero no excede $1 - \alpha$; si este valor se excediera, el intervalo resultante sería el $\{0\}$. Se ha criticado al método fiducial por el siguiente razonamiento que consideramos **parcial**. Dado que $\theta = |\mu|$ y $T = |X|$, y originalmente $X \sim N(\mu, 1)$, se pudo haber obtenido la distribución fiducial para μ , que resulta ser una $N(X, 1)$, y que coincide con la final que se obtiene al suponer como inicial a $\pi(\mu) \propto 1$. Teniendo la fiducial para μ , se seguiría que la distribución (inducida) para θ sería $H_i(\theta; t) = \Phi(\theta - t) - \Phi(-\theta - t)$, que es continua en $[0, \infty)$.

Como un simple ejercicio, se presentan a continuación los resultados de cobertura empírica basados en 1,000 simulaciones, para los intervalos $[0, \theta^*]$ obtenidos usando H y H_i , simplemente construidos de manera que en θ^* la correspondiente distribución valga $1 - \alpha$.

θ	$1 - \alpha$	Cobertura	
		H	H_i
0.50	95 %	96 %	100 %
0.50	50 %	49 %	100 %
0.50	25 %	26 %	40 %
0.75	95 %	96 %	100 %
2.00	95 %	95 %	98 %
3.00	95 %	95 %	95 %

El uso de H_i no produce intervalos con la cobertura correcta debido a que se intenta con H_i hacer inferencias sobre θ (parámetro de interés), usando un procedimiento que se utilizó para hacer inferencias sobre μ ; y esto creemos que es incorrecto y no descalifica al procedimiento inferencial utilizado (en este caso, el fiducial).

Como se dijo, la fiducial para μ coincide con la distribución final si se utiliza la distribución no informativa usual. En el enfoque Bayesiano, cuando se utilizan distribuciones no informativas, debe identificarse primero el parámetro de interés y encontrar la distribución inicial correspondiente. No creemos que sea adecuado utilizar la final para μ y simplemente transformar, ya que se estaría proponiendo implícitamente para el problema el uso de la inicial $\pi(\mu) \propto 1$.

Consideramos que la crítica al procedimiento fiducial con base en el ejemplo 3 se centra en un argumento equivocado, ya que compara problemas **distintos**.

REFERENCIAS

- Barndorff-Nielsen, O. (1978). *Information and exponential families*. New York: Wiley.
- Hsieh, H.K. (1990). Inferences on the coefficient of variation of an inverse Gaussian distribution. *Comm. Statist. (Theory and Methods)* **19**, 1589-1605.
- Nádas, A. (1973). Best tests for zero drift based on first passage times in Brownian motion. *Technometrics* **15**, 125-132.
- O'Reilly, F.J. y Rueda, R. (1992). Goodness of fit for the inverse Gaussian distribution. *Can. J. Statist.*, **20**, 387-397.
- Patil, S.A. y Kovner, J.L. (1976). On the test and power of the zero drift on first passage times in Brownian motion. *Technometrics* **18**, 341-342.
- Pedersen, J.G. (1978). Fiducial inference. *Int. Stat. Rev.*, **46**, 147-170.
- Rueda, R. (1988). Intervalos de confianza y de probabilidad en la distribución Gaussiana inversa. *Aportaciones Matemáticas, Comunicaciones*, **5**, 369-375.
- Seshardi, V. y Schuster, J.J. (1974). Exact tests for zero drift based on first passage times in Brownian motion. *Technometrics*, **16**, 133-134.
- Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Annals of Math. Statist.*, **30**, 877-880.

Estimación de Componentes de Varianza de un Modelo Estadístico Particionado

EMILIO PADRÓN
U.A. de Coahuila

y LUIS LATOURNERIE
U.A.A.A.N. Coahuila

1. INTRODUCCIÓN

Un programa de mejoramiento genético debe estar apoyado en los métodos estadísticos para avanzar con mayor firmeza en la selección de plantas (genotipos) que coadyuven a incrementar la producción por unidad de superficie, al seleccionar la técnica adecuada que conlleva a una mejora en la toma de decisiones. Por lo tanto se estudia el comportamiento de genotipos de maíz en diferentes ambientes, y se logra obtener el verdadero efecto de sus varianzas en base a las esperanzas de cuadrados medios, de acuerdo a la técnica tradicional del análisis de varianza (ANOVA) como se aprecia en Searle (1987). Searle et al. (1992), comentan que las sumas de cuadrados del análisis de varianza para datos desbalanceados siguen siendo los mismos que para datos balanceados a excepción de tener n_i en lugar de n y $N = \sum n_i$, y el que los datos sean desbalanceados no elimina la posibilidad de obtención de estimadas negativas de σ^2_τ (componente de varianza para tratamiento) en el análisis de varianza. La información agronómica proviene de ensayos de 35 líneas S_2 de porte enano, las cuales fueron sometidas a evaluaciones tempranas para estimar su ACG, comportamiento agronómico y rendimiento; evaluándose en tres localidades (Torreón, Coah. Río Bravo, Tamps. y Celaya, Gto.) en cruzas con tres probadores de estrecha base genética (cruzas simples): dos de porte normal y una de porte enano. Utilizando dos repeticiones por localidad en comparación con 37 testigos, de los cuales dos son comerciales y los demás experimentales. La teoría aquí desarrollada se aplicó a los datos obtenidos por Latournerie (1990) en su trabajo de tesis. Esta investigación forma parte del programa de mejoramiento genético del Instituto Mexicano del Maíz "Mario E. Castro Gil" de la Universidad Autónoma Agraria Antonio Narro.

2. DESCRIPCIÓN DEL MODELO

En el modelo definido a continuación se utilizaron 142 genotipos de los cuales 105 fueron cruzas y 37 fueron testigos.

Dicho modelo es:

$$Y_{ijk} = \mu + L_k + R_{j(k)} + G_i + (LG)_{ki} + E_{ijk}$$

donde

i	=	1,2,3, ...,t	genotipos
j	=	1,2,3, ...,r	repeticiones
k	=	1,2,3, ...,l	localidades

Y_{ijk} :	Variable aleatoria observable de la k -ésima localidad en la j -ésima repetición del i -ésimo genotipo
μ :	Media general
L_k :	Efecto de la k -ésima localidad
$R_{j(k)}$:	Efecto de la j -ésima repetición dentro de la k -ésima localidad
G_i :	Efecto del i -ésimo genotipo
$(LG)_{ki}$:	Efecto conjunto de la k -ésima localidad y del i -ésimo genotipo
E_{ijk} :	Componente aleatoria asociada con la k -ésima localidad en la j -ésima repetición del i -ésimo genotipo

Asumiendo que los efectos son variables aleatorias independientes, esto genera el modelo infinito o modelo II por lo tanto de acuerdo a dicho modelo, el cuadrado medio de la interacción Localidad-Genotipo es el apropiado cuadrado medio del error para probar genotipos. Además se asume que las esperanzas de efectos son cero, es decir

$$E[L_k] = E[R_{j(k)}] = E[G_i] = E[(LG)_{ki}] = E[E_{ijk}] = 0$$

También se supone que las esperanzas de productos cruzados de los diferentes efectos son cero, se tiene además que

$$E[L_k^2] = \sigma_L^2 \quad E[R/L]^2 = \sigma_{R/L}^2 \quad E[G_i^2] = \sigma_G^2 \quad E[(LG)_{ki}^2] = \sigma_{LG}^2 \quad E[E_{ijk}^2] = \sigma_e^2$$

Del modelo dado anteriormente se encontraron las esperanzas de cuadrados medios de cada uno de sus componentes de acuerdo a Rodríguez (1992). Antes de continuar se definirán algunos conceptos agronómicos referentes a la partición de los tratamientos.

Genotipos	=	individuos, plantas, animales etc.(En este caso se usaron plantas)
Cruzas	=	Serie de nuevas plantas sometidas a ensayo con una característica deseable conocida.
Probadores	=	Plantas con características deseables ya conocidas por pruebas estadísticas y agronómicas preliminares
Lin / p_1	=	Número de plantas que se están probando en cruza con la característica germoplásrica uno
$(Lin / p_2 \text{ vs } Lin / p_3)$	=	Contraste o grado de potencialidad entre las plantas de línea dentro de probador dos con las plantas de línea dentro de probador tres.
Testigos	=	Plantas comerciales (Que ya están en el mercado).
Cruza vs Testigo	=	Este es un contraste que mide el grado de potencialidad entre cruzas nuevas y las cruzas comerciales.

Todos estos efectos ya mencionados se interactúan con localidad con el objeto de analizar su contribución correspondiente.

3. ESPERANZA DE CUADRADOS MEDIOS

Se obtuvieron las esperanzas de la suma de cuadrados y en ellas se generan los grados de libertad de cada efecto y después se divide dicha esperanza por sus respectivos grados de libertad, formándose la esperanza del cuadrado medio; en este caso sólo se presenta el resultado final aplicado a los datos de campo.

4. APLICACIÓN A LOS DATOS DE CAMPO

A continuación se resume, presentando solo los valores de las componentes de varianza estimadas (C.V.E) de efectos principales e interacciones, para la variable rendimiento de mazorca en ton/ha del experimento.

F.V	C.V.E.
Localidad	6.02056
Rep/Loc	0.11647
Genotipos	1.0915
Gen x Loc	3.254
Error	2.649
Total	13.13153

En este trabajo se obtuvieron las esperanzas de cuadrados medios correspondientes a cada efecto y al aplicar en ellas los datos de campo (Rendimiento) se obtuvo no solo la magnitud relativa de las varianzas, sino también los porcentajes de la suma de las varianzas estimadas, como se ver a continuación.

σ_L^2	representa	$\frac{6.02056 \times 100}{13.13153}$	45.8481 %
$\sigma_{R/L}^2$	representa	$\frac{0.11647 \times 100}{13.13153}$	0.88694 %
σ_G^2	representa	$\frac{1.0915 \times 100}{13.13153}$	8.31205 %
σ_{GL}^2	representa	$\frac{3.254 \times 100}{13.13153}$	24.78005 %
σ_e^2	representa	$\frac{2.649 \times 100}{13.13153}$	20.17282 %

5. CONCLUSIONES

De acuerdo a los datos se observa que fueron los efectos del ambiente los que más contribuyeron a la respuesta, esto se explica debido a que en la localidad de Celaya, Guanajuato los genotipos rindieron más que en Río Bravo, Tamaulipas y Torreón, Coah. En base a las estimaciones de componentes de varianza y a futura explotación comercial en diferentes localidades se espera encontrar la mejor combinación que facilite al investigador genetista seleccionar nuevo germoplasma con mayor grado de confiabilidad.

REFERENCIAS

- Latournerie, M.L. (1990). Comportamiento de 35 líneas S_2 de maíz (*Zea mays L.*) derivadas del sintético ideotipo trópico seco en un estudio de aptitud combinatoria con tres probadores. *Tesis licenciatura U.A.A.A.N.*
- Rodríguez, B.L. (1992). Esperanza de cuadrados medios de un diseño de bloques al azar con arreglo factorial combinatorio y partición de efectos. *Tesis Posgrado en Estadística U.A.A.A.N.*
- Searle, S.R. (1987). *Linear Models for Unbalanced Data*. New York: Wiley.
- Searle, S.R.; Casella,G. and McCulloch, C.E. (1992). *Variance Components*. New York: Wiley.

Clasificación Lineal Bayesiana

BEATRIZ PEÑALOZA NYSSEN

I.T.A.M.

1. INTRODUCCIÓN

El problema de clasificación es un problema real y cotidiano que consiste en decidir, a partir de observar ciertas cualidades de un individuo, a cual de una serie de clases éste pertenece. El problema de clasificación estadística, con un enfoque Bayesiano, siempre se puede resolver. Sin embargo, hay ocasiones (muy comunes) en que la solución es complicada algebraicamente y por lo mismo poco manejable, lo cual hace la toma de decisiones correcta pero ineficaz e impráctica. Este trabajo presenta una solución sencilla y manejable al problema de clasificación estadística que ha sido propuesta en la literatura, considerando todos los factores relevantes del problema. Su propósito es contribuir a que esta técnica sea más conocida y, esencialmente más utilizada.

2. EL PROBLEMA DE CLASIFICACIÓN ESTADÍSTICA

El problema de clasificación estadística consiste en decidir, a partir de observar ciertas cualidades de un individuo, a cual de una serie de clases pertenece. Por supuesto de entrada se desconoce la clase a la que pertenece el individuo, y por ende, cuál es la consecuencia de una asignación específica (puede ser un acierto o un error). Esta incertidumbre se puede deber a varias causas, entre las que se distinguen: desconocimiento del futuro, información costosa o imposible de adquirir, que se requiera la destrucción del individuo para la correcta clasificación, etc.

El problema de clasificación en general se aborda bajo algunos supuestos:

- Las clases forman una partición de la población, i.e. son conocidas, exhaustivas y excluyentes; por lo que un individuo que se sabe pertenece a la población pertenece a una y sólo una de las clases.
- Cada clase o grupo se identifica con el comportamiento de ciertas variables \underline{X} de los individuos que pertenecen a ella.

Este comportamiento es expresado a través de una función de densidad $P_i(\underline{X})$; $i=1,2,\dots,k$ que es, al menos en parte, conocida. Las densidades $P_1(\underline{X}) \dots P_k(\underline{X})$ no necesariamente tienen una estructura "similar" (e.g. pertenecer a la misma familia paramétrica). Cuando se considera más de una variable para la toma de decisiones, su comportamiento conjunto se describe a través de una función de densidad multivariada. Si se consideran variables discretas y continuas los modelos necesariamente son más complejos.

El problema de Clasificación Estadística ha sido estudiado desde hace muchos años, y quizás la técnica más conocida y utilizada sea el llamado Análisis Discriminante. En este

trabajo el problema se aborda en un contexto de toma de decisiones, con un enfoque Bayesiano y se comenta una aproximación que simplifica, en muchos casos, la solución del problema.

3. EL ANÁLISIS ESTADÍSTICO BAYESIANO

La estadística Bayesiana formula y resuelve todos los problemas de inferencia estadística como problemas de decisión. La operación de cualquier técnica Bayesiana se rige por los mismos principios básicos: los axiomas de coherencia. Si se aceptan los axiomas de coherencia entonces se puede demostrar que:

- Se puede y debe definir una función de pérdida (utilidad) que describa el orden de preferencia de las consecuencias.
- Se puede y debe definir una función de probabilidad que describa la incertidumbre sobre los sucesos inciertos que aparecen en el problema.
- La decisión óptima d^* es aquella que minimice la pérdida esperada (maximice la utilidad esperada).

De esta manera la estadística Bayesiana:

Incorpora información subjetiva relevante para la solución del problema.
Proporciona la mejor solución, sin inconsistencias.

4. SOLUCIÓN BAYESIANA AL PROBLEMA ORIGINAL

En base a lo mencionado anteriormente el problema de clasificación estadística se resuelve como un problema de decisión con incertidumbre. El problema es que el clasificador debe decidir a cuál clase asigna al individuo. Para estructurar el problema considere los siguientes elementos:

- Espacio de decisiones: $D=\{d_1 \dots d_k\}$
 d_i : el individuo se asigna a la clase i , donde las clases forman una partición conocida de la población en estudio.
- Espacio de eventos inciertos: $E=\{e_1 \dots e_k\}$
 e_i : el individuo pertenece a la clase i .
- Espacio de consecuencias: $C=\{c_{11}, c_{12}, \dots, c_{kk}\}$
 c_{ij} : la consecuencia por asignar un individuo a la clase i dado que pertenece a la clase j .

Como consecuencia de los axiomas de coherencia se deben asignar las funciones de pérdida y de probabilidad, en relación al contexto del problema, y se encuentra la mejor solución minimizando la pérdida esperada (en términos de la distribución diagnóstica y función de pérdida).

Cabe mencionar que cuando se asigna correctamente a un individuo la pérdida debe ser cero. Se pueden observar algunas características del individuo (a clasificar). Por ejemplo, para clasificar a un individuo como sano o enfermo medir su temperatura, coloración de

ojos, etc. También debe asignarse una función de probabilidad $P(E_i|\underline{x})$ $i=1,\dots,k$, que se determina por el Teorema de Bayes a través de $P(\underline{x}|E_i)$ $i=1,\dots,k$, la función predictiva de \underline{X} en la clase i , y la distribución inicial, i.e.

$$P(E_i|\underline{x}) = \frac{P(\underline{x}|E_i)P(E_i)}{P(\underline{x})}.$$

A esta distribución se le conoce como distribución posterior o **Distribución Diagnóstica** (probabilidad de pertenecer a la clase i dado que se han observado las características \underline{x} en el individuo). Si \underline{X} es un vector de dimensión r “grande” y sobretodo si combina variables discretas con continuas la especificación de la predictiva se puede complicar, y volver el proceso de clasificación lento y difícil.

5. CLASIFICACIÓN LINEAL BAYESIANA

Hasta ahora se ha expuesto la solución al problema de clasificación estadística, que se obtiene en términos de la distribución diagnóstica. En esta sección se presenta una propuesta (Bernardo, 1988) dentro del enfoque Bayesiano, para la obtención de una distribución diagnóstica más sencilla y manejable, que simplifica y agiliza la solución al problema de clasificación, a través de una función predictiva manejable.

Sea t una función relevante del vector de atributos observados que concentra la mayor parte de la información diagnóstica proveniente del vector de atributos \underline{x} , i.e. tal que

$$P(E_i|\underline{x}, D) \cong P(E_i|t, D).$$

Esto con el objetivo de condensar la información proveniente de las diferentes variables, tanto discretas como continuas, en un vector de dimensión menor, sin tener una pérdida considerable de información. Se supone que su comportamiento se puede modelar con “facilidad”. Se requiere un banco de datos D con el vector de atributos y clase a la que pertenece, para conocer a la población. La distribución diagnóstica se obtiene a partir de la Regla de Bayes,

$$P(E_i|t, D) \propto P(t|E_i, D)P(E_i|D),$$

y depende de la asignación de t en cada una de las clases. Por esto es esencial su cuidadosa selección, así como la asignación de su distribución muestral. Al asignar diferentes funciones los resultados también serán diferentes. La distribución inicial depende del tipo de muestreo usado.

Lo relevante de esta propuesta es la simplificación para obtener una distribución predictiva razonable y en consecuencia, una diagnóstica manejable.

5.1 Selección de $t(\underline{x})$

Se busca una función predictiva fácil (de dimensión menor) y “parecida” (del mismo tipo, familia) en las diferentes clases, i.e., estructuralmente homogéneas y de fácil manejo.

El uso de una ecuación lineal o suma ponderada de las variables independientes (atributos) para discriminar (separar) los grupos previamente definidos es motivado por la simplicidad de las funciones lineales de \mathbf{X} .

Este método se puede considerar como uno de separación. Su principal objetivo es encontrar una transformación lineal de los atributos de los individuos, tal que sus valores para individuos de una misma clase sean relativamente iguales (cercanos), mientras que con respecto a la de otras clases sean diferentes, es decir, que contengan toda la información discriminante proveniente de las muestras.

Este criterio de clasificación no es explicativo en el sentido que no describe el comportamiento de las variables originales en cada clase, es decir a partir de él no se puede deducir $P(x|E_i)$ $i=1,\dots,k$. Sin embargo, partiendo del supuesto que la Función Discriminante contiene toda (o gran parte de) la información relevante proveniente de las muestras, y si el número de atributos observados p es grande, que es lo más común, entonces se puede hacer uso del Teorema de Límite Central para garantizar que dentro de cada clase su distribución muestral es aproximadamente normal, sin importar el tipo de variables usadas (continuas y/o discretas).

$$t | E_i \xrightarrow{\text{Prob}} N_s(t | \mu_i, \Sigma_i) \quad i = 1, \dots, k$$

Por lo general es muy poca la información que se tiene sobre los parámetros, por lo que éstos se estiman Bayesianamente usando la información proveniente del banco de datos D . A partir de la distribución de t en cada clase y las correspondientes distribuciones iniciales de los parámetros se obtiene la distribución predictiva de t (la posterior de los parámetros se obtiene vía Bayes), quedando una distribución t-student multivariada. La distribución diagnóstica que soluciona el problema de clasificación estadística esta dada entonces por:

$$P(E_i | t, D) \propto St(t | m_i, V_i, n_i - s) P(E_i | D) \quad (1)$$

y da lugar al llamado Análisis Discriminante Bayesiano Estándar. La distribución posterior de la población depende del tipo de muestreo usado (prospectivo, retrospectivo).

La asignación de la Función Discriminante como una función relevante del vector de atributos $t=t(x)$ simplifica la solución al problema de clasificación, siendo su distribución conocida y hasta cierto punto manejable, ya que para clasificar a un nuevo individuo se debe calcular el vector t y después obtener su distribución diagnóstica de acuerdo con (1), lo que requiere un cómputo que puede ser considerado sofisticado para usuarios no entrenados.

A continuación se presenta una opción para aproximar la distribución diagnóstica (1), de tal forma que sea fácilmente manejable, sin alejarse demasiado de la verdadera distribución diagnóstica.

5.2 Una Aproximación Lineal

Se pretende ilustrar una estrategia para aproximar la relación entre la distribución diagnóstica y t mediante un modelo de Regresión Logística que involucra a las variables t sólo a través de una combinación lineal

$$\log \left(\frac{P(E_i | t, \omega)}{P(E_k | t, \omega)} \right) = \omega_{i0} + \omega_{i1}t_1 + \dots + \omega_{is}t_s, \text{ para } i=1, \dots, k-1$$

Se plantea como un problema de estimación paramétrica Bayesiana, donde ω^* (la mejor decisión) proporciona la mejor aproximación a la distribución diagnóstica, en el sentido que minimiza la pérdida esperada. Se utiliza una Medida de Divergencia Logarítmica como función de pérdida.

La solución se obtiene al encontrar las ω^* que maximicen alguna de las dos siguientes ecuaciones, dependiendo del tipo de muestreo usado (prospectivo (2), retrospectivo (3)).

$$\frac{1}{n} \sum_{j=1}^n \sum_{i=1}^k P(E_i | t_j, D) \log [P(E_i | t_j, \omega)] \quad (2)$$

$$\frac{1}{k} \sum_{h=1}^k \frac{1}{n_h} \sum_{j=1}^{n_h} \sum_{i=1}^k P(E_i | t_j, D) \log [P(E_i | t_j, \omega)] \quad (3)$$

Para la solución de cualquiera de estas ecuaciones, es necesario utilizar algún método numérico, en particular el método de Newton-Raphson es una buena opción.

Una vez obtenida ω^* , se obtiene una regla **lineal** de clasificación óptima, en el sentido que está libre de inconsistencias y se basa en la mejor aproximación a la distribución diagnóstica. Lo anterior resulta en un proceso de clasificación al alcance de cualquier persona, sin importar su profesión ni gusto por las matemáticas.

6. CONCLUSIONES

La estadística Bayesiana, al plantear los problemas de inferencia como problemas de decisión y abordarlos con un enfoque axiomático, produce soluciones óptimas y libres de inconsistencias. Este es el caso para el problema de clasificación estadística. Por otra parte, en la medida en que la implementación de las técnicas Bayesianas sea más simple sus ventajas metodológicas serán aprovechadas por un mayor número de usuarios, no necesariamente expertos estadísticos.

En este trabajo se presentó la solución Bayesiana general al problema estadístico de clasificación y en consonancia con el párrafo anterior, se presentó también una técnica que permite obtener una solución aproximada pero mucho más sencilla y manejable, que además reúne todas las cualidades del enfoque Bayesiano.

REFERENCIA

Bernardo, J.M. (1988). Bayesian Linear Probabilistic Classification. *Statistical Decision Theory and Related Topics IV*, 1. New York: Springer-Verlag. 151-161.

Solución al Problema de Programación Cuadrática usando un Método Heurístico

BLANCA ROSA PÉREZ SALVADOR

SERGIO DE LOS COBOS SILVA

UAM - I.

y

MIGUEL ANGEL GUTIÉRREZ ANDRADE

UAM-A.

1. INTRODUCCIÓN

El problema de programación cuadrática, cuyo enunciado es:

$$\text{Maximizar } (x - x_{op})^T A(x - x_{op}) \text{ sujeta a } Rx \geq r \text{ donde } x \in R^n \quad (1.1)$$

A una matriz $n \times n$ negativa definida y R una matriz $p \times n, p \leq n$, de rango completo, surge de manera natural en algunos problemas de la teoría de la estadística, lo que justifica su estudio. Su solución puede obtenerse por el método de Kunh-Tucker, por el algoritmo de puntos interiores o al resolver un problema de complementaridad lineal (Stablein, 1983), cuyo enunciado es:

Encontrar dos vectores ν y ω tales que $\nu^T \omega = 0; \nu, \omega \leq 0$; y $P\nu = \omega + q$.

En el caso que nos ocupa, $q = r - Rx_{op}$ y $P = -RA^{-1}R^T$. La solución de (1.1) es $x_{op} = x_{op} + A^{-1}R^T \nu$.

Por otro lado, un gran número de situaciones problemáticas pueden resolverse utilizando el problema combinatorio, cuyo enunciado es:

Optimizar la función $C(x)$ donde $x \in U \subset Z^n$.

El problema combinatorio no puede ser resuelto mediante los métodos clásicos de cálculo diferencial, ya que el dominio de C es discreto. Encontrar la solución requiere calcular los valores de $C(x)$ para todos los elementos de su dominio, y hasta entonces se puede elegir el punto x_0 que optimiza a $C(x)$. Estos cálculos pueden ser extremadamente numerosos, por lo que resulta práctico tener algún método que con sólo unos cuantos elementos del dominio proporcione "buenas" soluciones. Esto es lo que se tiene con los métodos heurísticos. Este trabajo pretende presentar al problema de programación cuadrática como un problema combinatorio, lo que permitiría aplicar las heurísticas conocidas.

La estructura del trabajo es la siguiente: en la sección 1, se describe dos ejemplos de la teoría de la estadística que dan lugar a un problema de programación cuadrática; en la

sección 2, se describen dos ejemplos del problema combinatorio y se describe el método de búsqueda tabú; en la sección 3, se muestra cómo el problema de programación cuadrática se transforma en un problema combinatorio.

2. EL PROBLEMA COMBINATORIO Y LOS MÉTODOS HEURÍSTICOS

Entre los métodos heurísticos más utilizados están: los algoritmos genéticos, la búsqueda tabú o el de recocido simulado. En la práctica estos métodos han dado muy buenos resultados para resolver el problema combinatorio.

Ejemplos clásicos del problema combinatorio son:

- El problema del agente viajero: este es un problema muy conocido, consiste en determinar el recorrido óptimo por n ciudades, sin pasar dos veces por una misma ciudad. Se considera que $c_1, c_2, c_3, \dots, c_n$ son las diferentes ciudades que se van a visitar, y que d_{ij} representa la distancia entre las ciudades c_i y c_j . Cada recorrido corresponde a una permutación de las n ciudades. Entonces cada recorrido puede ser representada por el vector $x = (i_0, i_1, \dots, i_n)$ $1 \leq i_j \leq n$, $i_j \neq i_k$, $j = 1, 2, \dots, n$ e $i_0 = i_n$, porque se debe visitar cada ciudad una sola vez y se regresa al punto de partida. La función objetivo es $C(x) = \sum_{j=1}^n d_{i_{(j-1)}i_j}$. Se pide encontrar el vector x_o tal que $C(x_o) \leq C(x)$ para todo x en el dominio de C .
- El problema de la implementación de m plantas en n posibles localidades ($m < n$): este problema consiste en determinar las localidades en las que debe instalarse m plantas de tal manera que minimice los costos de producción (o que maximice la utilidad esperada). En este caso el dominio de la función objetivo es igual a

$$U = \{(x_1, x_2, \dots, x_n) \mid x_i = 0 \text{ ó } x_i = 1 \text{ y } \sum_{i=1}^m x_i = m\},$$

donde cada coordenada de los elementos de U representa una localidad; si $x_i = 0$, en la localidad i no se implementa la planta, si $x_i = 1$ en la localidad i sí se implementa la planta. Se busca encontrar el vector $x_o \in U$ que optimice la función objetivo.

2.1 Búsqueda Tabú

Este es uno de los métodos que mejores resultados proporciona y fue propuesto en los años de 1989 y 1990 por Fred Glover. El método consiste en elegir un punto de inicio, y buscar entre sus “vecinos” uno que mejore la función objetivo. Los vecinos de un punto pueden ser las permutaciones en que todos los elementos permanecen en su lugar y sólo se permutan dos de ellos (las coordenadas que estén con una separación predeterminada). Una vez que se tiene el punto que mejora la función objetivo, se continua con el proceso. Se

busca entre los “vecinos” del nuevo punto aquel que mejore la función objetivo. Los puntos considerados vecinos en las iteraciones anteriores no se consideran en los procesos subsecuentes porque se convierten en tabú para evitar el ciclado.

Se ha observado en la práctica que este método mejora con muy pocas iteraciones las soluciones obtenidas utilizando otros métodos (Glover 1989, 1990, y de los Cobos 1994).

3. EJEMPLOS DE LA TEORÍA DE LA ESTADÍSTICA EN LOS QUE SURGE EL PROBLEMA DE PROGRAMACIÓN CUADRÁTICA

3.1 Un Ejemplo de Regresión Lineal Múltiple

Considere que en el modelo

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon \quad (\text{cuya forma matricial es } Y = X\beta + \varepsilon),$$

donde X_{1i} representa el número de semestres que un estudiante (el estudiante i) ha estado inscrito en una universidad, X_{2i} representa el número de horas que el mismo estudiante acostumbra estudiar a la semana y Y_i representa el número de materias que el estudiante lleva acreditadas. Es natural suponer que β_0, β_1 y β_2 son números positivos, ya que mientras más tiempo lleva inscrito en la universidad y mientras más horas dedique al estudio el número de materias acreditadas es mayor. De esta forma la ecuación de mínimos cuadrados se escribe como:

$$\text{Minimizar } (Y - X\beta)^T (Y - X\beta) \quad \text{sujeta a } \beta \geq 0.$$

Si la matriz X es de rango completo, se tiene el enunciado equivalente:

$$\text{Minimizar } (\beta - (X^T X)^{-1} X^T Y)^T (\beta - (X^T X)^{-1} X^T Y) \quad \text{sujeta a } -\beta \leq 0,$$

el cual corresponde al enunciado del problema de programación cuadrática.

3.2 Un Ejemplo en Superficies de Respuesta

Considere el modelo

$$Y = \beta_0 + b^T x + x^T B x + \varepsilon$$

con $x^T = (x_1, x_2, x_3)$, b un vector y B una matriz negativa definida. En este modelo Y representa la resistencia de un material para recubrir neumáticos de automóvil; x_1 representa el tiempo de horneado en el procesamiento del material; x_2 representa la temperatura

del horno y x_3 representa la cantidad de una sustancia A que se agrega al material. Las restricciones del sistema son: la temperatura máxima que puede alcanzar el horno T_f , y el costo máximo permitido para la sustancia A , el cual es igual a c .

Así, el problema se enuncia como:

$$\text{maximizar } Y = \beta_0 + b^T x + x^T B x + \varepsilon \quad \text{sujeta a } x_2 \leq T_f \text{ y } x_3 \leq c, \text{ o bien,}$$

$$\text{maximizar } \left(x - \frac{1}{2} B^{-1} b \right)^T B \left(x - \frac{1}{2} B^{-1} b \right) \text{ sujetas a las mismas restricciones.}$$

4. EL PROBLEMA DE COMPLEMENTARIEDAD LINEAL VISTO COMO UN PROBLEMA COMBINATORIO

Dado el vector $x = (x_1, x_2, \dots, x_n)$, se define la matriz diagonal

$$D(x) = \begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & x_n \end{pmatrix}.$$

Si $x_i = 0$ ó $x_i = 1$, $i = 1, 2, \dots, n$, entonces $\nu, \omega \leq 0$ y $\nu^T \omega = 0$ si y sólo si $\nu = D(x)\nu^*$ y $\omega = (I - D(x))\omega^*$ para un vector x y los vectores ν^* y $\omega^* \leq 0$.

La solución al problema de programación cuadrática, en términos de los vectores ν y ω del problema de complementariedad lineal es:

$$x_{op_r} = x_{op} + RA^{-1}R^T \nu.$$

Así, el problema de programación cuadrática se puede escribir en función de la matriz P y el vector q del problema de complementariedad como:

Maximizar la función

$$C(x) = q \{ PD(x) + (D(x) - I) \}^{-1} D(x) P D(x) \{ PD(x) + (D(x) - I) \}^{-1} q$$

donde $x = (x_1, x_2, \dots, x_n)^T$ y $x_i = 0$ ó $x_i = 1$, $i = 1, 2, \dots, n$.

Este último enunciado corresponde a un problema combinatorio, ya que se debe encontrar el vector x (vector discreto) tal que maximice la función objetivo. La solución se encuentra utilizando el método de búsqueda tabú, el cual como ya se mencionó, generalmente da buenas soluciones.

5. CONCLUSIONES

En particular, consideramos que los ejemplos analizados no han sido suficientes como para poder afirmar que por este procedimiento se obtienen mejores resultados; sin embargo, a pesar de lo limitado de su número si se ha podido observar que mejoran los resultados obtenidos con el método de Kuhn-Tucker. Este hecho, aunado a los resultados observados del método tabú en una gran variedad de situaciones, permiten tener esperanza que al revisar mayor número de casos se tendrá evidencias que el procedimiento propuesto es aceptable.

REFERENCIAS

- De los Cobos, S. (1994). *La técnica de la búsqueda tabú y sus aplicaciones*. Tesis Doctoral, DEP-FI, UNAM.
- Glover, F. (1989). Tabú Search, Part I, ORSA, *Journal on Computing* 1, 190-206.
- Glover, F. (1990). Tabú Search, Part II, ORSA, *Journal on Computing* 2, 4-31.
- Stablein,D. M., Carter, W. H. Jr. and Wampler G. L. (1983). "Confidence Regions for Constrained Optimal in Response Surface Experiments", *Biometrics*, 39, 759-763.

Comparación de la Capacidad Inferencial y Predictiva entre el Estimador de Máxima Verosimilitud y Estimadores Alternativos en Presencia de MV-Colinealidad

GUSTAVO RAMÍREZ VALVERDE
ISEI, Colegio de Postgraduados

JANET RICE
Universidad de Tulane, E.U.A.

e

IGNACIO MÉNDEZ RAMÍREZ
IIMAS-UNAM

1. INTRODUCCIÓN

El Estimador de Máxima Verosimilitud (EMV) presenta serios problemas en el modelo de regresión logística cuando la matriz de información esta mal condicionada. Existen dos diferentes fuentes de mal condicionamiento de la matriz de información (Lesaffre y Marx, 1993). 1) Colinealidad entre las variables explicatorias y 2) MV-Colinealidad. Varios estimadores han sido propuestos para disminuir los efectos negativos de colinealidad en regresión logística, sin embargo, no todos ellos han sido comparados contemplando la MV-colinealidad. En este articulo, mediante un estudio de simulación, se evalúan efectos de cada una de las diferentes fuentes de mal condicionamiento bajo distintas condiciones que han mostrado tener influencia en los efectos de la colinealidad. Se compararon la capacidad inferencial y predictiva del estimador de Máxima Verosimilitud con estimadores alternativos incluyendo estimadores tipo Ridge, tipo Stein y Componentes Principales.

2.- ESTIMADORES EMPLEADOS

Los estimadores alternativos estudiados en este trabajo se pueden agrupar en 4 grupos a)Estimadores tipo ridge, b)Estimadores tipo ridge iterativos, c) Estimadores tipo ridge generalizados y d) Estimadores tipo Stein. a) REGRESION TIPO RIDGE . Schaefer (1979), propuso estimadores tipo ridge en regresión logística dado por : $\hat{\beta}_R = (X^T \hat{V} X + kI)^{-1} X^T \hat{V} X \hat{\beta}$ donde $\hat{V} = \text{diag}(\hat{v}_1, \hat{v}_2, \dots, v_n)$ y $\hat{v}_i = \hat{\pi}_i(1 - \hat{\pi}_i)$, $k \geq 0$. La selección de un valor optimo de k continua irresuelto, Schaefer (1979 y 1986) y Schaefer et al. (1984) proponen los siguientes valores para k produciendo distintos estimadores ridge a) $\hat{k} = [\max(\delta_j \hat{\beta})]^{-1}$, (RID1) b) $\hat{k} = (\hat{\beta}^T \hat{\beta})^{-1}$ (RID2) y c) $\hat{k} = (p+1) / (\hat{\beta}^T \hat{\beta})$, (RID3). Donde $\hat{\beta}$ es el EMV de β , δ_j es el eigenvector asociado a el $j^{\text{ésimo}}$ eigenvalor de la matriz de información estimada y p es el numero de variables explicatorias. Lee y Silvapulle (1988) proponen otro estimador ridge (RID4) dado por $\hat{k} = [\text{trace}(X^T \hat{V} X)] / [\hat{\beta}^T X^T \hat{V} X \hat{\beta}]$.

b) ESTIMACION TIPO RIDGE ITERATIVA. Los estimadores ridge propuestos adolecen del problema de que fueron construidos pensando en la matriz de información y se utilizo la matriz de información estimada sustituyendo β por su EMV, sin embargo la norma esperada del EMV es mas grande que la norma del parámetro $\hat{\beta}$, por lo que se esperaría que se tuviera una subestimación del parámetro ridge k . Esto sugiere la utilización de un enfoque iterativo para estimar k , se podría tomar los valores del EMV como valores iniciales del proceso y en la segunda etapa utilizar uno de los estimadores ridge para estimar la matriz de información, recalcular iterativamente el estimador ridge hasta que se tenga un cambio relativo pequeño, esto es parar cuando $(\hat{k}_{i+1} - \hat{k}_i)/\hat{k}_i < \delta$, la determinación de k no es analíticamente clara, empíricamente, se diseña la siguiente regla de decisión $\delta = 30T^R$, donde T es $\text{tr}(X^TX)^{-1}/p+1$, R es el cociente del número condición de X^TVX entre el número condición de X^TX . De esta forma la regla de decisión contempla el condicionamiento de la matriz X^TX independientemente de la dimensión de X y la presencia de MV-colinealidad. Se dejó un límite de seguridad para evitar la posibilidad de valores de \hat{k} demasiado grandes, el límite usado fue $\hat{k} = 0.25$. En este estudio se usaron las versiones iterativas de los estimadores RID1, RID2, RID3 y RID4 definidos en 2.2, los cuales serán denotados por IRID1, IRID2, IRID3 e IRID4 respectivamente.

c) ESTIMACION TIPO RIDGE GENERALIZADO. Marx (1988) sugirió la posibilidad de usar regresión ridge en el modelo lineal generalizado, dado por: $\hat{\beta}(R) = (X^T\hat{V}X + QRQ^T)^{-1}X^T\hat{X}\hat{\beta}$, donde $X^T\hat{V}X$ es la matriz de información estimada, $R = \text{dig}(r_1, r_2, \dots, r_p)$. En este trabajo se usaron dos tipos de regresión generalizada el primero a) Regresión ridge generalizado (GRID). Se usaron los valores de r_i dados por $r_i = 1/\delta_i^2$, donde δ_i es el i -ésimo elemento de $\underline{\delta} = Q^T\hat{\beta}$ y Q es la matriz de eigenvectores de $X^T\hat{V}X$. b) Basándose en el trabajo de Ali (1991) se construye un valor alternativo para los r_i , dando lugar a un estimador mejorado de ridge generalizado (IRID). El estimador mejorado de ridge generalizado está dado por: $\hat{\beta}_I(R) = A^*\hat{\beta}$, donde $A^* = (I - S)^{1/2}$ con $S = I - (X^T\hat{V}X + QRQ^T)^{-1}X^TX$.

d) ESTIMADORES TIPO STEIN. El estimador tipo Stein en regresión logística está dado por: $\hat{\beta}_s = c\hat{\beta}$, donde $\hat{\beta}$ es el EVM y $0 < c < 1$. En este trabajo se usaron dos tipos de estimadores de Stein definidos por: a) El valor de c que minimiza el Error Cuadrático Medio del estimador (STEIN1). Este estimador fue propuesto por Schaefer (1986) y el valor de c está dado por: $c = \hat{\beta}^T\hat{\beta}/[\hat{\beta}^T\hat{\beta} + \text{trace}(X^T\hat{V}X)]$ y b) El valor c que minimiza $L = E[(\hat{\beta}_s - \beta)^T X^T V X (\hat{\beta}_s - \beta)]$ $= E[(c\hat{\beta} - \beta)^T X^T V X (c\hat{\beta} - \beta)]$ (STEIN2). Este estimador fue propuesto por Marx (1988) y el valor de c está dado por $C = \left[\sum_{i=1}^{p+1} \alpha_i^{-2} \lambda_i \right] / \left[p + 1 + \sum_{i=1}^{p+1} \alpha_i^{-2} \lambda_i \right]$, donde α_i es el i -ésimo elemento

de $\underline{\alpha} = M^T \underline{\beta}$, con M la matriz de eigenvectores, y λ_i es el i -ésimo eigenvalor de la matriz $X^T V X$.

e) ESTIMADOR DE COMPONENTES PRINCIPALES. El estimador de componentes principales (EPC) usado en este trabajo fue propuesto por Schaefer en 1986 y esta dado por: $\hat{\underline{\beta}}_{ep} = (X^T \hat{V} X)^+ (X^T \hat{V} X) \underline{\beta}$, donde $(X^T \hat{V} X)^+ = \sum_{i=1}^{p+1-r} \underline{m}_i \underline{m}_i^T / \lambda_i$ con \underline{m}_i el eigenvector asociado a λ_i el i -ésimo eigenvalor de la matriz $X^T V X$.

3. EL ESTUDIO DE SIMULACIÓN

Con la finalidad de comparar los distintos estimadores se realizó un estudio de simulación, los casos simulados tuvieron 2 variables y un tamaño de muestra de $n = 25$. Los factores bajo estudio fueron: a) Correlación entre variables explicatorias. Dos diferentes correlaciones fueron usadas ($r = 0.95662$ y $r = 0.99572$), dando un número de condición para la matriz $X^T X$ de 45.106 y 466 respectivamente; b) Dirección de la colinealidad. Dos ángulos (0° y 90°) entre el parámetro y el eigenvector asociado con el menor eigenvalor de la matriz $X^T X$ (llamado dirección de la colinealidad) y c) Tamaño de la norma del vector de parámetros. Dos diferentes tamaños fueron usados $|\underline{\beta}| = 1$ ó 2. Las situaciones estudiadas se resumen en todas las combinaciones entre los factores estudiados. Una vez que la matriz diseño y el parámetro $\underline{\beta}$ es determinado, los valores de π_i fueron calculados de acuerdo a la situación simulada. La condición de MV-colinealidad depende de la variable respuesta, por lo que, para la determinación de su presencia se calculó el número de condición de la matriz de información estimada (k_I). La simulación fue llevada a cabo hasta que se obtuvieron 1000 ensayos con k_I mayor que 100 (situaciones con MV-colinealidad) si el número de condición de la matriz $X^T X$ (k_X) era 45.106 ($r=0.95662$) y para las situaciones con k_X de 466 ($r=0.99572$) hasta que se obtuvieron 1000 ensayos con $k_I > 1000$ (situaciones con MV-colinealidad). El número de ensayos donde el EMV no existe no fueron contabilizados, entonces los resultados están condicionados a la existencia de el EMV. Los criterios para comparar los estimadores fueron: a) El error Cuadrático medio estimado (ECM), calculado por $ECM = \frac{1}{L} \sum_{j=1}^L (\hat{\beta}_{ij} - \beta_i)^2$, donde L es el número de ensayos

obtenidos de acuerdo a la forma de parar definida arriba, $\hat{\beta}_{ij}$ es el estimador de β_i el i -ésimo elemento de $\underline{\beta}$ en el j -ésimo ensayo. b) El sesgo medio (SM) dada por: $SM = \frac{1}{L} \sum_{j=1}^L (\hat{\beta}_{ij} - \beta_i)$, $i=0,1,2$; c) La suma de cuadrados totales de predicción (SCTP) calculada por: $SCTP = \frac{1}{L} \sum_{j=1}^L \sum_{i=1}^n (\hat{\pi}_{ij} - \pi_i)^2$, donde $\hat{\pi}_{ij}$ es el valor estimado de la probabilidad de éxito en la i -ésima observación del j -ésimo ensayo y π_i es la probabilidad de éxito en la i -ésima observación.

d) El error medio de clasificación dado por

$$EMC = \frac{1}{nL} \sum_{j=1}^L \sum_{i=1}^n [I_i(\hat{\pi}_{ij} < .5 \wedge \pi_i \geq .5) + I_i(\hat{\pi}_{ij} > .5 \wedge \pi_i \leq .5)].$$

4. RESULTADOS Y CONCLUSIONES

Por restricciones de espacio sólo se presentarán algunos de los principales resultados, la información completa está disponible con el primer autor. Los resultados presentaron bastante regularidad, primero se muestran los efectos de los factores estudiados en el EMV, y posteriormente la comparación de los distintos estimadores. Bajo todas las situaciones estudiadas los dos tipos de colinealidad afectaron ECM, el SCTP, el SM y el EMC (figuras 1 y 2 ejemplifican estos resultados), sin embargo, los efectos fueron más marcados en casos con MV-colinealidad, mostrando además un gran efecto negativo en el sesgo. La dirección de β mostró ser uno de los factores mas importantes en el ECM del EMV, mostrando valores mayores cuando se tenían 90° de colinealidad, también mostró efecto en el sesgo medio en presencia de colinealidad entre las variables explicatorias, sin embargo, no mostró grandes efectos en las demás propiedades. En cuanto a la comparación de los estimadores, se encontró que en casos de ML-colinealidad todos los estimadores alternativos tuvieron mejor comportamiento que el EMV en cuanto a SM, ECM y SCTP, pero no tuvieron diferencias claras en cuanto a EMC. El ordenamiento fue bastante consistente en los demás factores, los estimadores tipo Ridge (RID1, RID3, RID4 y RIDG) tuvieron un comportamiento similar con excepción de RID2 que tuvo mejor comportamiento, el estimador STEIN1 tuvo un comportamiento similar a los del tipo Ridge y el estimador STEIN2 tendió a parecerse a RID2 en su comportamiento. Los estimadores tipo Ridge iterativo mostraron en general mejor comportamiento que sus versiones no iterativas con excepción de RID2 que algunas en algunas condiciones fue superior a su versión iterativa. Los estimadores con mejor comportamiento fueron IRID y EPC siendo los más prometedores (la figura 3 ejemplifica las diferencias entre estimadores).

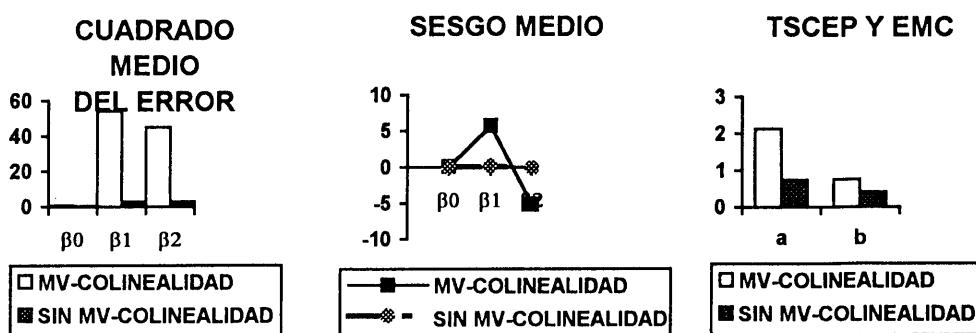


Figura 1. Efecto de MV-colinealidad en el error cuadrático medio, sesgo medio, TSCEP (a) y EMC (b) con $n=25$, 0° de colinealidad, $r=0.99$ y parámetro igual a 1.

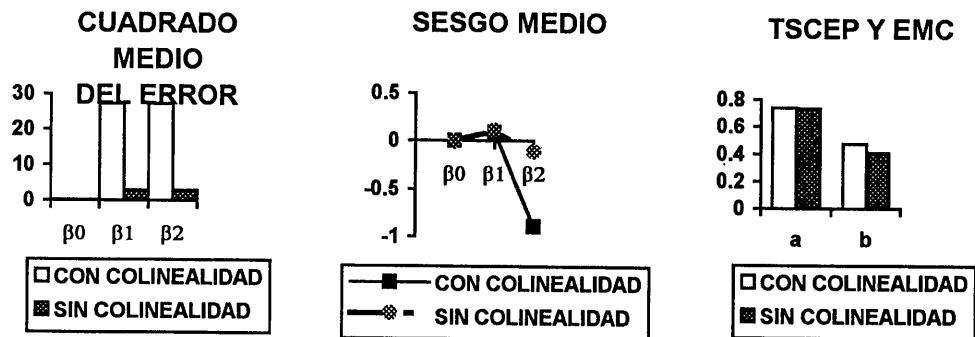


Figura 2. Efecto de Colinealidad entre las variables explicatorias en el error cuadrático medio, sesgo medio, TSCEP (a) y EMC (b) con $n=25$, 0° de colinealidad, $r=0.99$ y parámetro igual a 1.

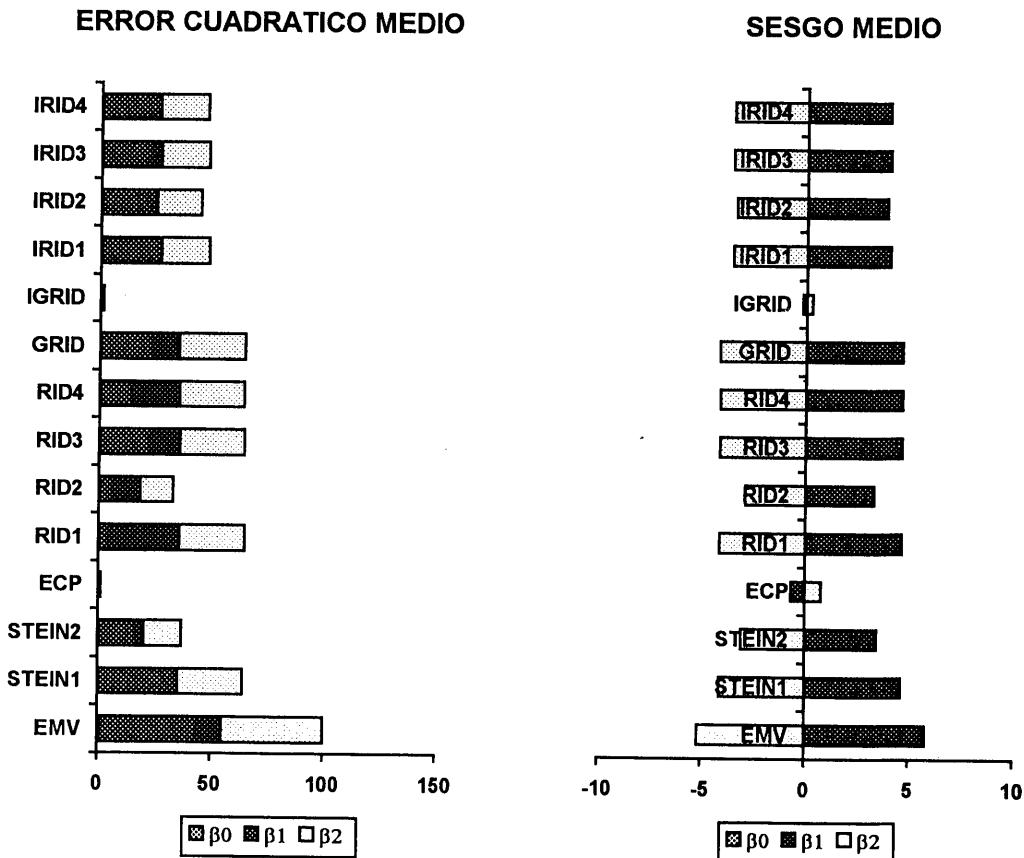


Figura 3. Ejemplo del comportamiento de los distintos estimadores para $n=25$, 0° de colinealidad, $r=0.99$ y parámetro igual a 1.

REFERENCIAS

- Ali, M.A. (1991). Restricted Minimum Bias Linear Estimator in Regression. *Communications in Statistics-Theory and Methods.* **20**, 3751-3760.
- Lee, A.H. y Silvapulle, M.J. (1988). Ridge Estimation in Logistic Regression. *Communications in Statistics-Simulation and Computation.* **4**, 1231-1257.
- Lesaffre, E. y Marx B.D. (1993). Collinearity in Generalized Linear Regression. *Communications in Statistics-Theory and Methods.* **22**, 1933-52.
- Marx, B.D. (1988). Ill-conditioned Information Matrices and the Generalized Lineal Model: An Asymptotically Biased Estimation Approach. *Doctoral Dissertation.* Virginia Politechnic Institute and State University.
- Schaefer, R.L. (1979). Multicollinearity and Logistic Regression. *Doctoral Dissertation.* University of Michigan.
- Schaefer, R.L. (1986). Alternative Estimators in Logistic Regression when the Data are Collinear. *Journal of Statistical Computations and Simulations* **25**, 75-91.
- Schaeffer, R. L., Roi, L.D. y Wolfe R.A. (1984). A Ridge Logistic Estimator. *Communications in Statistics-Theory and Methods.* **13**, 99-113.

Verosimilitud para Mezclas vía Aproximación Estocástica

MARÍA GUADALUPE RUSSELL NORIEGA

CIMAT

1. INTRODUCCIÓN

Los modelos de mezclas finitas de distribuciones surgen de manera natural en diversas áreas del conocimiento, tales como biología, química, física, medicina, y estadística. De ello se desprende el interés en estudiar teoría de inferencia para estos modelos, particularmente en aquellos casos en los que existan parámetros desconocidos.

El presente trabajo considera una clase paramétrica de mezclas de distribuciones, y aborda el problema de estimación por máxima verosimilitud. Motivados por el hecho de que la verosimilitud en cuestión es el valor esperado de una función aleatoria observable que puede simularse de manera sencilla en la computadora, se propone aquí una metodología basada en técnicas de aproximación estocástica con el objeto de obtener un estimador que aproxima al estimador de máxima verosimilitud.

2. MEZCLAS

Consideremos una familia de densidades $\{g(\mathbf{z}|\theta)\}$, descritas por un parámetro θ . Por una *mezcla de densidades* entendemos a la densidad obtenida de interpretar a θ como una variable aleatoria que ocurre con distribución H . Más específicamente, conceptualizamos la mezcla de densidades como

$$f_H(\mathbf{z}) = \int g(\mathbf{z}|\theta) dH(\theta), \quad (1)$$

donde la función f_H es la *densidad de la mezcla* correspondiente a la distribución H , la *distribución de la mezcla*.

A partir del modelo general de mezclas (1) podemos especificar familias adicionales de densidades en al menos dos formas distintas. Por un lado es posible que la distribución de la mezcla, H , sea considerada a su vez como proveniente de otra familia de distribuciones, que puede ser paramétrica o no. Por otra parte, también la función de densidad condicional $g(\mathbf{z}|\theta)$ para θ fijo puede considerarse miembro de una familia, que puede depender de otros parámetros. Una familia paramétrica de este tipo se describiría por

$$f(\mathbf{z}; \delta, \lambda) = \int g_\delta(\mathbf{z}|\theta) dH_\lambda(\theta), \quad (2)$$

y uno de los objetivos inmediatos para este modelo sería la estimación de sus parámetros λ y δ . El problema abordado en el presente trabajo considera un escenario paramétrico en el cual las funciones g_δ y H_λ ya se encuentran especificadas.

Un caso particular de (2) que da lugar a una situación más simple, se obtiene cuando la distribución H es fija y conocida, es decir, los parámetros desconocidos se involucran únicamente a través de la densidad g . En este caso se obtiene la siguiente familia paramétrica de densidades

$$f(\mathbf{z}; \boldsymbol{\beta}) = \int g_{\boldsymbol{\beta}}(\mathbf{z}|\boldsymbol{\theta}) dH(\boldsymbol{\theta}), \quad (3)$$

Tratar de resolver directamente los problemas de estimación en mezclas de distribuciones resulta difícil y costoso, aún cuando el avance actual en posibilidades de cómputo es considerable. Por ello es interesante el planteamiento y desarrollo de nuevas técnicas de estimación que faciliten y optimicen el tiempo y esfuerzo computacional para efectuar la estimación.

3. APROXIMACIÓN ESTOCÁSTICA

Robbins y Monro (1951) dan inicio a los métodos de aproximación estocástica. El problema originalmente abordado por ellos consiste en determinar el parámetro $\boldsymbol{\theta}$ que sea solución de la ecuación $M(x) = \alpha$, con M definida sobre \mathbb{R} . La función $M(x)$ no es observable, pero para cada $x \in \mathbb{R}$ es posible observar realizaciones de una variable aleatoria $Y(x)$ con algún error aleatorio asociado. La distribución de probabilidad de los errores generalmente se conoce. De este modo la variable aleatoria $Y(x)$ cumple que $E[Y(x)] = M(x) \forall x$. La versión multivariada del método de Robbins-Monro se debe a Blum (1954), y consiste en lo siguiente: sea $\mathbf{Y}(x)$ un vector aleatorio m -dimensional observable $\forall \mathbf{x} \in \mathbb{R}^m$, con alguna distribución asociada, de modo que $M(\mathbf{x}) = E[\mathbf{Y}(\mathbf{x})]$. El interés es estimar $\boldsymbol{\theta}$, la única raíz del sistema de m ecuaciones simultáneas $M(\boldsymbol{\theta}) = \boldsymbol{\alpha}$. La fórmula recursiva para aproximar $\boldsymbol{\theta}$ se define como $\mathbf{X}_{n+1} = \mathbf{X}_n - \{\mathbf{Y}(\mathbf{X}_n) - \boldsymbol{\alpha}\}$, donde la sucesión $\{\boldsymbol{\alpha}_n\}$ es una sucesión de números reales positivos tal que $\boldsymbol{\alpha}_n \rightarrow 0$ y $\sum_n \boldsymbol{\alpha}_n = 0$. Asimismo, se demuestra que $\mathbf{X}_n \xrightarrow[n \rightarrow \infty]{c.s.} \boldsymbol{\theta}$ (Blum, 1954).

4. OPTIMIZACIÓN VÍA APROXIMACIÓN ESTOCÁSTICA

El artículo de Ruppert *et al.* (1984) muestra una opción de optimización, mediante el uso de aproximación estocástica. En su formulación más general, el problema abordado consiste en maximizar la función $F(\boldsymbol{\beta}) = E\{f(\boldsymbol{\beta}, \mathbf{e})\}$, donde $\boldsymbol{\beta}$ es un parámetro p -variado, \mathbf{e} es un vector aleatorio m -dimensional, y $f(\boldsymbol{\beta}, \mathbf{e})$ es una función conocida de valor real.

Denotemos por $\mathbf{D}(\boldsymbol{\beta}, \mathbf{e})$, $\mathbf{D}(\boldsymbol{\beta}) = \mathbf{H}(\boldsymbol{\beta}, \mathbf{e})$ y $\mathbf{H}(\boldsymbol{\beta})$ los gradientes y Hessianos de $f(\boldsymbol{\beta}, \mathbf{e})$ y $F(\boldsymbol{\beta})$, respectivamente. Para un valor fijo dado $\boldsymbol{\beta} \in \mathbb{R}^p$, se supone que existen condiciones para obtener una realización, por simulación, para una respuesta $y = f(\boldsymbol{\beta}, \mathbf{e})$, utilizando la distribución conocida m -variada del vector de errores \mathbf{e} . Con ello es posible también proporcionar realizaciones simuladas del gradiente $\mathbf{D}(\boldsymbol{\beta}, \mathbf{e})$ y del Hessiano $\mathbf{H}(\boldsymbol{\beta}, \mathbf{e})$ en términos de las realizaciones de la función $f(\boldsymbol{\beta}, \mathbf{e})$. Si hay condiciones tales que

$E\{\mathbf{D}(\beta, \mathbf{e})\} = \mathbf{D}(\beta)$, es posible aplicar el procedimiento de Robbins-Monro descrito en la sección 3 directamente a $\mathbf{D}(\beta)$. El objetivo sería encontrar β^* que sea solución de la ecuación $\mathbf{D}(\beta) = 0$, con el propósito de que β^* maximice a $F(\beta)$.

5. DESCRIPCIÓN DEL MÉTODO

Una vez obtenidas las aproximaciones numéricas para el gradiente y el Hessiano, Ruppert *et al.* (1984) definen la sucesión de estimaciones del tipo Robbins-Monro, mediante las siguientes ecuaciones recurrentes:

$$\hat{\beta}_{N+1} = \hat{\beta}_N - (N + k_1)^{-1} \hat{\mathbb{H}}_N^{-1} \mathbf{D}(\hat{\beta}_N, \mathbf{e}_N), \quad \hat{\mathbb{H}}_{N+1} = \hat{\mathbb{H}}_N + (N + k_2)^{-1} \{ \hat{\mathbb{H}} (\hat{\beta}_N, \mathbf{e}_N) - \hat{\mathbb{H}} \}$$

donde $\hat{\mathbb{H}}_N$ es una estimación de $\mathbb{H}(\beta^*)$ con β^* el verdadero valor para el cual la función $F(\beta)$ es máxima y N es el número de iteraciones. Por otra parte tenemos que el gradiente $\mathbf{D}(\beta^*) = E\{\mathbf{D}(\beta^*, \mathbf{e})\} = 0$, de modo que la matriz de covarianza para $\mathbf{D}(\beta^*, \mathbf{e})$, denotada por \mathbf{s} , está dada por $\mathbf{s} = E\{\mathbf{D}(\beta^*, \mathbf{e})\mathbf{D}(\beta^*, \mathbf{e})^t\}$,

Ruppert *et al.* (1984) argumentan que, bajo condiciones regulares,

$$\hat{\beta} \xrightarrow[N \rightarrow \infty]{c.s.} \beta^*, \quad \hat{\mathbb{H}}_N \xrightarrow[N \rightarrow \infty]{c.s.} \mathbb{H}(\beta^*) \quad (4)$$

y que

$$N^{1/2}(\hat{\beta}_N - \beta^*) \xrightarrow[N \rightarrow \infty]{d} N(\mathbf{0}, \mathbb{H}(\beta^*)^{-1} \mathbf{s} \mathbb{H}(\beta^*)^{-1}). \quad (5)$$

El criterio de paro es proceder hasta la N -ésima iteración tal que

$$\frac{-tr(\hat{\mathbf{s}}_N \hat{\mathbb{H}}_N^{-1})}{N \hat{F}_N} < 2\Delta,$$

donde \hat{F} y $\hat{\mathbf{s}}_N$ son estimaciones del tipo Robbins-Monro.

6. ADAPTACIÓN A MÁXIMA VERO SIMILITUD

Concentrémonos en el objetivo de estimar el parámetro β considerando para esto el modelo paramétrico de mezclas dado por la ecuación (3) y una muestra observada z_1, \dots, z_n . Cabe notar que la densidad de cada observación puede verse como el valor esperado de la variable aleatoria $g_\beta(z | \theta)$ con respecto a la distribución H para θ , de manera que la ecuación (3) puede escribirse como $f(z; \beta) = E_H[g_\beta(z | \theta)]$. Con esto, la función de verosimilitud basada en la muestra z_1, \dots, z_n es

$$L(\beta; z_1, \dots, z_n) = \prod_{i=1}^n f(z_i; \beta) = \prod_{i=1}^n E_H[g_\beta(z_i | \theta)].$$

Denotemos con H^* la medida producto H^n , es decir, $dH^*(\theta_1, \dots, \theta_n) = \prod_{i=1}^n dH(\theta_i)$. Luego, la verosimilitud puede reescribirse como

$$L(\beta; z_1, \dots, z_n) = E_{H^*} \prod_{i=1}^n g_\beta(z_i | \theta_i). \quad (6)$$

Ya que la distribución H está especificada, entonces es factible generar realizaciones independientes del error aleatorio $\mathbf{e} = (\theta_1, \theta_2, \dots, \theta_n)$. Definamos la función

$$\varphi(\beta, \mathbf{e}) = \prod_{i=1}^n g_\beta(z_i | \theta_i). \quad (7)$$

Entonces de la ecuación (6), resolver el problema de optimización en este caso de mezclas de distribuciones equivale a encontrar el estimador de máxima verosimilitud para β .

Por analogía directa con la metodología propuesta, la función de verosimilitud $L(\beta; z_1, \dots, z_n)$ dada en la ecuación (6), es la función desconocida a maximizar, $F(\beta)$. La función $\varphi(\beta, \mathbf{e})$ en (7) juega el papel de la función aleatoria observable $f(\beta, \mathbf{e})$ para distintos valores de β . En lo sucesivo denotemos por $\hat{\beta}$ el estimador obtenido por el método descrito en Ruppert *et al.* (1984), y por $\hat{\beta}_{MV}$ el verdadero valor del estimador de máxima verosimilitud. La intención del algoritmo es producir un valor de $\hat{\beta}$ que se aproxime a $\hat{\beta}_{MV}$.

Para dotar a la técnica de interpretación en el contexto de estimación estadística, nos interesaría establecer la consistencia de $\hat{\beta}$, así como estimar la variabilidad de $\hat{\beta}$ como estimador del verdadero valor β . La teoría asintótica de Ruppert *et al.* (1984) no resuelve del todo este aspecto, ya que estima la variabilidad de $\hat{\beta}$ con respecto al estimador máximo verosímil, por lo que es necesario obtener propiedades complementarias vía aproximaciones asintóticas.

Suponiendo condiciones de regularidad el estimador $\hat{\beta}_{MV}$ cumple que

$$\hat{\beta}_{MV} \xrightarrow[n \rightarrow \infty]{c.s.} \beta, \text{ y } \sqrt{n}(\hat{\beta}_{MV} - \beta) \xrightarrow[n \rightarrow \infty]{d} N(0, I(\beta)^{-1}), \quad (8)$$

donde $I(\beta)$ es la matriz de información de Fisher y n es el tamaño de la muestra considerada. Por otra parte, de las propiedades asintóticas de Ruppert *et al.* (1984) tenemos que para $\hat{\beta}_{MV}$ fijo,

$$\hat{\beta} \xrightarrow[n \rightarrow \infty]{c.s.} \hat{\beta}_{MV}, \quad (9)$$

y

$$\sqrt{N}(\hat{\beta} - \hat{\beta}_{MV}) \xrightarrow[N \rightarrow \infty]{d} N(0, H^{-1}(\hat{\beta}_{MV}) S(\hat{\beta}_{MV}) H^{-1}(\hat{\beta}_{MV})), \quad (10)$$

donde N es el número de iteraciones.

De las ecuaciones (8) a (10), obtenemos el siguiente resultado importante: para n, N grandes,

$$\text{Var}(\hat{\beta}) \approx \frac{1}{n} [I(\beta)]^{-1} + \frac{1}{N} H^{-1}(\beta) S(\beta) H^{-1}(\beta),$$

cuya consecuencia práctica es

$$\hat{\text{Var}}(\hat{\beta}) = \frac{1}{n} \left[\hat{\mathbf{I}}(\hat{\beta}) \right]^{-1} + \frac{1}{N} \hat{\mathbf{H}}^{-1}(\hat{\beta}) \hat{\mathbf{s}}(\hat{\beta}) \hat{\mathbf{H}}^{-1}(\hat{\beta}) ,$$

con $\hat{\mathbf{I}}(\hat{\beta})$, $\hat{\mathbf{H}}^{-1}(\hat{\beta})$ y $\hat{\mathbf{s}}(\hat{\beta})$ son estimaciones del tipo Robbins-Monro.

Es posible también mostrar que $\hat{\beta}$ es asintóticamente normal. Este hecho, en conjunto con la estimación $\hat{\text{Var}}(\hat{\beta})$, nos permite la construcción de intervalos de confianza y pruebas de hipótesis sobre $\hat{\beta}$ mediante estadísticas del tipo Wald (ver Serfling, 1980, Sección 4.4).

7. COMENTARIOS Y CONCLUSIONES

Aproximación estocástica tal como se desarrolló en este trabajo es capaz de dar información sobre la cercanía entre la N -ésima estimación y el valor real. Una de las ventajas de la metodología aquí propuesta, es que en teoría funciona para cualquier dimensión, lo cual no siempre es practicable, ya que podríamos caer en un problema de sobreparametrización, que no hemos considerado en la teoría implementada. Sin embargo, para problemas de dimensión moderada, es una alternativa que sustituye a integración numérica con resultados favorables. Sin lugar a dudas el algoritmo tiene sus complicaciones y para algunos problemas de estimación resulta complicado y costoso computacionalmente. Sin embargo, aún en estos casos difíciles cabe señalar que debe valorarse el uso del método ya que obtenemos mayores beneficios como lo es el hecho de obtener mayor eficiencia en las estimaciones. El poder del método se aprecia cuando se trabaja constantemente en un ámbito de modelos de mezclas de distribuciones, como por ejemplo en contaminación atmósferica vía los modelos de receptores. Si el lector desea profundizar en el tema expuesto vea Russell-Noriega (1996).

REFERENCIAS

- Blum, J. R. (1954). Multidimensional stochastic approximation methods, *Ann. Math. Statist.*, **25**, 737-744.
- Robbins, H. and Monro, S. (1951). A stochastic approximation method, *Ann. Math. Statist.* **22**, 400-407.
- Ruppert, D., Reish, R.L., Deriso, R. B., and Carroll, R. J. (1984). Optimization Using Stochastic Approximation and Monte Carlo Simulation (with Application to Harvesting of Atlantic Menhaden), *Biometrics*, **40**, 535-545.
- Russell-Noriega, M.G. (1996). Verosimilitud para mezclas vía aproximación estocástica con aplicación a un modelo de receptores. *Tesis de maestría*. Universidad de Guanajuato.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

Evaluación de la Robustez del Nivel de Significancia de la Prueba de Mann-Whitney usando Simulación Monte Carlo

DAVID SOTRES RAMOS

y

LUIS E. CASTILLO MÁRQUEZ

ISEI - Colegio de Postgraduados

1. INTRODUCCIÓN

Uno de los objetivos que se formulan con mayor frecuencia en las investigaciones experimentales es el de la comparación de los efectos de dos tratamientos. El modelo estadístico que usualmente se emplea para representar este tipo de experimentos es el siguiente: Sean $\{X_1, X_2, \dots, X_m\}$ y $\{Y_1, Y_2, \dots, Y_n\}$ dos muestras aleatorias independientes con distribución $F_X(x)$ y $G_Y(x)$ respectivamente, y se desea contrastar las hipótesis

$$[H_0: F_X(x) = G_Y(x)] \text{ vs. } [H_a: G_Y(x) = F_X(x-\Delta)] \quad (\Delta \neq 0) \quad (1)$$

Bajo normalidad (i.e. $F_X(x) = \Phi(x), \forall x$) es bien conocido que la prueba uniformemente más potente e insesgada para contrastar las hipótesis en (1) es la prueba t de Student, ver por ejemplo Lehmann (1959). Sin embargo, cuando se desconoce la forma de $F_X(x)$ se recomienda emplear la prueba de Mann-Whitney (1947) (MW), en lugar de la prueba de t, por sus diferentes e importantes propiedades, ver por ejemplo Randles y Wolfe (1979). Todas éstas propiedades de la prueba de MW se probaron suponiendo como válido al modelo en (1), en particular este modelo supone que $\sigma_x^2 = \sigma_y^2$ tanto en la hipótesis nula como en la alternativa. Sin embargo, resulta claro que las distribuciones de X y Y pueden llegar a tener varianzas diferentes ($\sigma_x^2 \neq \sigma_y^2$).

En este trabajo se consideró una amplia variedad de distribuciones para las cuales el supuesto estándar de la prueba de Mann-Whitney ($\sigma_x^2 = \sigma_y^2$) no se cumple. Usando simulación Monte Carlo se calculó el nivel de significancia real de la prueba de MW para muestras pequeñas ($5 \leq n_1, n_2 \leq 10$) teniendo como objetivo identificar el tipo de distribuciones que producen diferencias "grandes" entre el nivel de significancia real y el nominal.

2. LA FAMILIA DE DISTRIBUCIONES λ GENERALIZADA

La distribución de probabilidad de una variable aleatoria continua usualmente se especifica mediante su función de densidad o por su función de distribución de probabilidad acumulativa. Alternativamente también se puede especificar en base a la función de distribución acumulativa inversa usualmente llamada la función percentil: $R(p)$, $0 \leq p \leq 1$. Este concepto es particularmente útil en estudios de simulación Monte Carlo debido al siguiente

resultado: Si X es una variable aleatoria uniforme sobre el intervalo $(0,1)$, entonces la transformación $R(U)$ produce una variable aleatoria con función percentil R . Un ejemplo específico es la función percentil lambda de Tukey (1960):

$$R(p) = \{p^\lambda - (1-p)^\lambda\} / \lambda, \quad (0 \leq p \leq 1) \quad (2)$$

Ramberg y Schmeiser (1972) demostraron como ésta distribución podría emplearse para aproximar muchas de las distribuciones simétricas más conocidas, y exploraron su aplicación en estudios de simulación Monte Carlo. En un trabajo posterior Ramberg y Schmeiser (1974) generalizaron (2) a una familia de distribuciones con 4 parámetros definida mediante la siguiente función percentil:

$$R(p) = \lambda_1 + [p^{\lambda_3} - (1-p)^{\lambda_4}] / \lambda_2, \quad (0 \leq p \leq 1) \quad (3)$$

en donde λ_1 es un parámetro de tendencia central, λ_2 es un parámetro de escala y, λ_3 y λ_4 son parámetros de forma. Esta distribución incluye la función de distribución lambda original y también incluye distribuciones asimétricas. A esta clase de distribuciones se le conoce como a la familia de distribuciones λ Generalizada. Esta distribución con cuatro parámetros incluye una amplia variedad de formas de funciones de densidad. Estas densidades se caracterizan por su media μ , varianza σ^2 , asimetría $\phi_3 = E(x-\mu)^3/\sigma^3$ y por su Kurtosis $\phi_4 = E(x-\mu)^4/\sigma^4$. Schmeiser (1977) probó que la distribución (3) tiende a la distribución exponencial (θ) cuando $\lambda_4 \rightarrow 0$, $\lambda_1 = 0$ y $\lambda_2 = \lambda_4/\theta$. La distribución en (3) también nos permite obtener buenas aproximaciones a otras densidades bien conocidas. Por ejemplo la distribución con $\lambda_1 = 0$, $\lambda_2 = 0.1975$, y $\lambda_3 = \lambda_4 = 0.1349$ resulta en una aproximación a la distribución normal $|\Phi(x) - R^{-1}(x)| = 0.001$.

Como ya se dijo la familia λ Generalizada tiene la interesante propiedad de que la generación de números aleatorios con éste tipo de distribución de probabilidad puede realizarse de manera muy eficiente. Todas las variables aleatorias reportadas en éste trabajo se generaron usando ésta propiedad de la distribución λ Generalizada. Las variables aleatorias uniformes se generaron empleando el método de Tzuka y Lécuyer (1991).

3. DISTRIBUCIONES GENERADAS

El nivel de significancia de la prueba de Mann-Whitney se estimó por simulación para 120 parejas de distribuciones $\{F_X(x; \mu_1, \sigma_1, (\Phi_3)_1, (\Phi_4)_1), F_Y(y; \mu_2, \sigma_2, (\Phi_3)_2, (\Phi_4)_2)\}$ que se obtuvieron al combinar diferentes valores de las medias, desviaciones estándar, coeficiente de asimetría y de kurtosis así como el tamaño de las muestras de ambas poblaciones. Los valores que se seleccionaron para éstos parámetros son todas las combinaciones posibles de los siguientes conjuntos: $(n_1, n_2) \in \{(5,5), (5,10), (10,5), (10,10)\}$, $(\sigma_1, \sigma_2) \in \{(1,2), (1,0.5)\}$, $(\phi_3)_1 = (\phi_3)_2 \in \{0,1,2\}$, $(\phi_4)_1 = (\phi_4)_2 \in \{1.8, 3, 4, 6, 9\}$ y en todos los casos $\mu_1 = \mu_2 = 0$, ver apéndice 1. Para cada pareja de distribuciones $F_X(x)$ y $F_Y(y)$ se generaron 100,000 pares de muestras independientes de

tamaño n_1 y n_2 respectivamente, y en seguida se calculó el nivel de significancia estimado de la prueba de MW en forma empírica. Es decir se calculó el porcentaje de muestras en que la prueba rechazó la hipótesis nula. Este nivel de significancia empírico es un estimador del nivel de significancia real de la prueba. El nivel de significancia nominal α empleado corresponde al nivel de la prueba de MW más cercano a 0.05.

Para simplificar la interpretación de los resultados, se calculó para cada combinación de los parámetros (n_1 , n_2 , σ_1 , σ_2 , $(\phi_3)_1$, $(\phi_4)_1$), la desviación media ($d = \sum |\alpha_i - \alpha_n|/k$) del nivel real estimado (α_1) con respecto al nivel nominal de significancia (α_n), donde k es el número de diferentes valores de kurtosis considerados. También se calculó el Porcentaje de Error Relativo (PER) calculado por el cociente d/α_n .

4. RESULTADOS

De la tabla 1 resulta claro que los casos con un alto PER son en primer término aquellos en donde $n_1 = n_2$ y $\phi_3 = 2$ y en segundo lugar aquellos en donde $n_1 \neq n_2$, y la población con mayor desviación estandar es la que tiene menor tamaño de muestra. Por ejemplo en el caso $n_1 = 5$, $n_2 = 10$, $\alpha = 0.03996$, $\sigma_x = 1$ y $\sigma_y = 0.5$, $PER = 185.9\%$ lo que implica un nivel de significancia real igual a 11.4%, valor que resulta demasiado alto para ser utilizado en aplicaciones prácticas. Para todos los demás casos se obtiene un $PER \leq 40\%$ lo cual pudiera ser un error tolerable ya que si el nivel de significancia nominal es del 5% lo peor que podría pasar es que el nivel de significancia real alcanzara un valor del 7%.

REFERENCIAS

- Lehmann, E.L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.
- Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.
- Mann H. and Whitney, D. (1947). On a test whether one of two random variables is stochastically larger than the other. *Ann. Math. Statistics.*, 18, 50-60
- Ramberg, J. et al. (1979). A probability distribution and its uses in fitting data. *Technometrics*. 21, 201-214.
- Ramberg, J. S., and Schmeiser, B.W. (1972). An approximate method for generating symmetric random variables. *Comm. ACM*, 15, 987-990.
- Ramberg, J.S. and Schmeiser, B.W. (1974). An approximate method for generating asymmetric random variables. *Comm. ACM*, 17, 78-82
- Randles, R. and Wolfe, D. (1979). *Introduction to the Theory of Nonparametric Statistics*. John Wiley and Sons. New York.
- Schmeiser, B.W. (1977). Methods for modelling and generating probabilistic components in digital computer simulation when the standard distributions are not adequate: a survey. *Proceedings of the Winter Simulation Conference*, pp. 51-57.
- Tzuka, S. and Lécuyer, P. (1991). Efficient and portable combined tausworth random number generator. *ACM Transactions on Modeling and Computer Simulation*. 1, 99-112.

APENDICE 1

Tamaño de muestra: $n_1 = 5, n_2=5$		Alfa Nominal = 0.03174 (θ_N)
$\sigma_x=1, \sigma_y = 2$		$\sigma_x=1, \sigma_y = 0.5$
$(\Phi_3)_1=(\Phi_3)_2 = 0$ y	$(\Phi_4)_1=(\Phi_4)_2 \in \{1.8, 3, 4, 6, 9\}$	
d	0.0073	0.0073
PER	22.9%	22.9%
$(\Phi_3)_1=(\Phi_3)_2 = 1$ y	$(\Phi_4)_1=(\Phi_4)_2 \in \{4, 6, 9\}$	
d	0.0103	0.0102
PER	32.4%	32.19%
$(\Phi_3)_1=(\Phi_3)_2 = 2$ y	$(\Phi_4)_1=(\Phi_4)_2 \in \{8, 6, 9\}$	
d	0.0302	0.0298
PER	95.1%	93.8%
Tamaño de muestra: $n_1 = 5, n_2=10$ y		Alfa Nominal = 0.03996
$\sigma_x=1, \sigma_y = 2$		$\sigma_x=1, \sigma_y = 0.$
$(\Phi_3)_1=(\Phi_3)_2 = 0$ y	$(\Phi_4)_1=(\Phi_4)_2 \in \{1.8, 3, 4, 6, 9\}$	
d	0.0147	0.0129
PER	22.9%	32.2%
$(\Phi_3)_1=(\Phi_3)_2 = 1$ y	$(\Phi_4)_1=(\Phi_4)_2 \in \{4, 6, 9\}$	
d	0.0108	0.0352
PER	27.0%	88.0%
$(\Phi_3)_1=(\Phi_3)_2 = 2$ y	$(\Phi_4)_1=(\Phi_4)_2 \in \{8, 6, 9\}$	
d	0.0103	0.0743
PER	25.7%	185.9%

Tabla 1. Resumen de los niveles de significancia estimados.

$d = \sum |\theta_i - \theta_N|/k$ donde θ_N es el nivel de significancia nominal (0.03996) y las $\theta_1, \theta_2, \dots, \theta_k$ son los niveles de significancia estimados correspondientes a los diferentes valores dados a los parámetros de Kurtosis $(\phi_4)_1$ y $(\phi_4)_2$ indicados en la tabla. Por ejemplo en el caso $(\phi_3)_1 = (\phi_3)_2 = 0$ los valores dados a los parámetros de Kurtosis son $\{1.8, 3, 4, 6, 9\}$. PER = Porcentaje de Error Relativo.

APENDICE 1 (continuación)

Tamaño e Muestra: $n_1 = 10, n_2=5$ y $\sigma_x=1, \sigma_y = 2$		Alfa Nominal = 0.03996 (θ_N) $\sigma_x=1, \sigma_y = 0.5$
$(\Phi_3)_1=(\Phi_3)_2 = 0$ y		$(\Phi_4)_1=(\Phi_4)_2 \in \{1.8, 3, 4, 6, 9\}$
d	0.0299	0.0151
PER	74.8%	27.5%
$(\Phi_3)_1=(\Phi_3)_2 = 1$ y		$(\Phi_4)_1=(\Phi_4)_2 \in \{4, 6, 9\}$
d	0.0345	0.0110
PER	86.3%	27.5%
$(\Phi_3)_1=(\Phi_3)_2 = 2$ y		$(\Phi_4)_1=(\Phi_4)_2 \in \{8, 6, 9\}$
d	0.0723	0.0298
PER	180.9%	22.5%
Tamaño de Muestra: $n_1 = 10, n_2 = 10$ y $\sigma_x=1, \sigma_y = 2$		Alfa Nominal = 0.04324 $\sigma_x=1, \sigma_y = 0.5$
$(\Phi_3)_1=(\Phi_3)_2 = 0$ y		$(\Phi_4)_1=(\Phi_4)_2 \in \{1.8, 3, 4, 6, 9\}$
d	0.0073	0.0079
PER	16.8%	18.2%
$(\Phi_3)_1=(\Phi_3)_2 = 1$ y		$(\Phi_4)_1=(\Phi_4)_2 \in \{4, 6, 9\}$
d	0.0173	0.0352
PER	40.0%	38.0%
$(\Phi_3)_1=(\Phi_3)_2 = 2$ y		$(\Phi_4)_1=(\Phi_4)_2 \in \{8, 6, 9\}$
d	0.0682	0.0689
PER	157.7%	159.3%

Tabla 1. Resumen de los niveles de significancia estimados.

$d = \sum |\theta_i - \theta_N|/k$ donde θ_N es el nivel de significancia nominal (0.03996) y las $\theta_1, \theta_2, \dots, \theta_k$ son los niveles de significancia estimados correspondientes a los diferentes valores dados a los parámetros de Kurtosis $(\phi_4)_1$ y $(\phi_4)_2$ indicados en la tabla. Por ejemplo en el caso $(\phi_3)_1 = (\phi_3)_2 = 0$ los valores dados a los parámetros de Kurtosis son $\{1.8, 3, 4, 6, 9\}$. PER = Porcentaje de Error Relativo.

Un Sistema de Análisis Multivariado para Grandes Volúmenes de Información

JOSÉ VENCES RIVERA, IGNACIO RAMÍREZ MARTÍNEZ

y

MARCO ANTONIO FLORES NÁJERA

INEGI - Aguascalientes

1. INTRODUCCIÓN

En la actualidad existe una demanda creciente de los métodos estadísticos multivariados aplicados para describir e interpretar la información contenida en un conjunto de datos multidimensionales. Si bien es cierto que muchos de esos métodos no son de origen reciente, su aplicación en cambio, para grandes volúmenes de información, inicia con el surgimiento de la computadora digital comercial, y constituyen una poderosa herramienta para resolver problemas en las más diversas disciplinas científicas. Para el caso de México, en los últimos años las técnicas de análisis estadístico multivariado han cobrado cada vez mayor importancia; sin embargo, la experiencia muestra que hace falta mayor difusión en esta materia bajo esquemas que estén acorde a las necesidades de países como el nuestro.

El INEGI constituye la mayor fuente de información oficial del país, en donde se genera información estadística básica y derivada para niveles geográficos altamente desagregados, lo cual permite identificar características de la población más cercanas a la realidad y por ende tomar decisiones de manera directa. La cantidad de variables disponibles, producto de censos y encuestas por muestreo, es enorme así como el número de unidades de análisis. Los paquetes estadísticos que existen en el mercado (como el SAS, SPSS, BMDP y otros) presentan ciertas limitaciones en el formato de salida de resultados y en el tiempo de proceso porque, al ser de propósito general, obtienen una gran cantidad de resultados que para muchos fines no son de utilidad; además no incluyen algunos métodos de reciente creación.

Por otro lado, el costo que implica la compra de un paquete estadístico para utilizar solamente unos cuantos procedimientos, y la capacitación de los recursos humanos para su operación, resulta una carga económica de consideración.

Lo anterior dio origen al desarrollo de un *sistema de análisis multivariado*, denominado SAM, el cual en su etapa inicial contempla seis procedimientos básicos y un paquete de utilerías. El objetivo principal de este trabajo es ofrecer a la comunidad investigadora un sistema interactivo de fácil operación para analizar grandes volúmenes de datos multidimensionales mediante procedimientos relativamente novedosos o poco conocidos.

2. MÉTODOS ESTADÍSTICOS

A continuación se describen brevemente los métodos estadísticos multivariados que se incluyen en el sistema en su primera versión. No se presentan fórmulas, ya que el enfoque está dirigido a resaltar la importancia que tiene cada uno de ellos.

2.1. Métodos Para Probar Normalidad Multivariada

La hipótesis de normalidad multivariada es fundamental para muchas técnicas estadísticas de análisis de observaciones multidimensionales. Desafortunadamente son pocos los métodos formales disponibles para validar esta hipótesis, entre los más importantes que reporta la literatura estadística son el de *Shapiro-Wilk* y el de *Cramér-von Mises*. Con el objeto de probar multinormalidad, Vences y Cossío (1989) hacen una comparación de estos procedimientos mediante simulación Monte Carlo. Bajo diferentes alternativas de prueba se encuentra que la estadística generalizada W de Shapiro-Wilk proporciona una medida ligeramente superior de no normalidad respecto a la estadística generalizada C de Cramér-von Mises. En el artículo de Malcovich y Afifi (1973) se recomienda que la estadística W puede ser utilizada para tamaños de muestra dentro de la amplitud $7 \leq n \leq 2000$. Por su parte, la estadística C puede ser utilizada para tamaños de muestra $n \geq 30$. El nivel de significancia observado es aproximado en la amplitud $.01 \leq \alpha \leq .15$; fuera de esta amplitud la aproximación puede perder precisión.

2.2. Método Para Seleccionar Variables

Una vez definido el problema de interés, el primer paso es seleccionar las variables que serán medidas sobre los individuos u objetos de alguna población determinada, de tal manera que sean representativas para explicar el comportamiento del fenómeno estudiado. Con frecuencia esto no es fácil, pues generalmente el investigador inicia proponiendo una serie de variables con base en el conocimiento que posee y posteriormente las somete a discusión con especialistas afines, llegando así a conformar un primer conjunto de variables. Esto es un tanto subjetivo, porque el número de variables puede crecer desmesuradamente, complicando posteriormente el análisis de resultados y a veces distorsionando la realidad, debido a que algunas variables pueden ser redundantes, o bien, no aportar de manera sustancial a la explicación del fenómeno. El sistema contempla tres procedimientos para la selección de variables mediante la técnica de componentes principales. Los dos primeros son bajo los criterios de la *varianza generalizada* y de la *traza* (ver McCabe, 1984), mientras que el tercero es por *análisis de factores*. Una discusión de estos criterios y los supuestos de programación pueden verse en Vences (1994).

2.3. Método Para Generar Indicadores Compuestos

En algunos problemas se presentan situaciones en donde se mide una gran cantidad de variables sobre un mismo individuo u objeto. Con frecuencia es de interés encontrar un indicador multivariado que resuma la información contenida en ese conjunto de variables a fin de facilitar el análisis y la interpretación. Para tal efecto, uno de los métodos más utilizados es el de *componentes principales*, el cual consiste en encontrar un número pequeño de nuevas variables (componentes principales) independientes, de manera que absorban en la mayor medida posible la varianza de las variables originales. La influencia de las variables originales sobre cada componente es determinada por el mismo método. De

esas nuevas variables, la primera componente principal es la que se utiliza para generar el índice, ya que explica la mayor cantidad de información inherente en las variables de insumo. Así, es un indicador de las características generales del fenómeno (Reyment, 1984). Dicho método se presenta bajo dos variantes: *componentes principales convencionales* y *componentes principales robustas*. El primero de ellos se utiliza cuando se tiene un conjunto "bondadoso" de datos, mientras que el segundo suele utilizarse cuando existen observaciones que se alejan notablemente de las demás (outliers). Una discusión del análisis de componentes principales robustas puede verse en Rodríguez (1995).

2.4. Método Para Detectar Puntos Extremos

Dado un conjunto de datos, en ocasiones suelen presentarse observaciones que se alejan notablemente del resto, siendo muy pequeñas o muy grandes comparadas con las demás, es decir, no siguen el patrón que presenta la mayoría de los datos. A este tipo de observaciones se les llama *puntos extremos* (outliers), los cuales pueden ser reales, o sea, son datos correctos, o bien son errores debido a las diferentes fuentes durante el proceso de captación. Como quiera que sea, hay situaciones en que estos puntos deben ser analizados para depurar y validar la información antes de ser utilizada, y así evitar posibles distorsiones en los resultados finales. Se consideran dos procedimientos para detectar observaciones atípicas desde el punto de vista multivariado, mediante la *distancia de Mahalanobis* y por *componentes principales robustas de Campbell (1980)*. Este último tiene la ventaja de detectar observaciones que se encuentran poco más allá de la periferia del resto. Sin embargo, para un gran número de variables el procedimiento es relativamente lento.

2.5. Método de Análisis de Factores

El análisis de factores (AF) es una técnica estadística multivariada, cuya necesidad surge debido a que algunas ideas fundamentales están inmersas en un conjunto de tres o más variables de insumo. Por ejemplo, *la clase social* es una variable socioeconómica no observable que sirve para implementar programas de acción, y se deriva de otras susceptibles de ser medidas como escolaridad, ingreso, ocupación y posición en el trabajo. Las soluciones factoriales pueden ser *ortogonales* u *oblicuas*. Las soluciones ortogonales son matemáticamente más simples de manejar, y son utilizadas cuando el investigador tiene como propósito reducir el número de variables originales, prescindiendo en cierta forma del significado de los factores derivados. Por otro lado, las soluciones oblicuas son más flexibles y realistas porque no se impone la restricción de que los factores resultantes sean no correlacionados, y se utilizan generalmente cuando la meta del investigador es obtener los factores teóricos subyacentes en las variables.

2.6. Método Para Estratificación de Unidades

La estratificación es una técnica utilizada generalmente con fines de muestreo para satisfacer los requerimientos de varianza mínima que permiten hacer estimaciones más precisas; no obstante, también se utiliza como técnica de análisis de conglomerados, para la

formación de grupos de unidades homogéneas de acuerdo con las características de interés, que sirven como base para una mejor planeación administrativa. Jarque (1981) propone un método de estratificación óptima aproximada para el caso multiparamétrico. Para tal efecto se minimiza una función de las varianzas de los estimadores, asumiendo una asignación proporcional al tamaño de los estratos. En el sistema se incluyen dos opciones de estratificación multivariada; *medición aleatoria* y *medición en la misma dirección*. En el primer caso, por lo general la estratificación se realiza para fines de muestreo; los estratos obtenidos son identificados por un número que no indica ningún orden, es simplemente una etiqueta. Para tal efecto, algunas de las variables utilizadas pueden estar medidas en la misma dirección y otras al contrario, dependiendo de la naturaleza del fenómeno estudiado. Con la segunda opción, la estratificación se realiza con fines de planeación administrativa; los estratos obtenidos son ordenados conforme a un número que indica el nivel de categoría correspondiente. Para esto conviene (aunque no es necesario) que las variables sean medidas en la mismo sentido, ya sea que todas ellas reflejen la parte positiva de un fenómeno o bien la parte negativa del mismo. De esta manera se facilita la interpretación de los resultados finales.

3. CARACTERÍSTICAS DEL SISTEMA

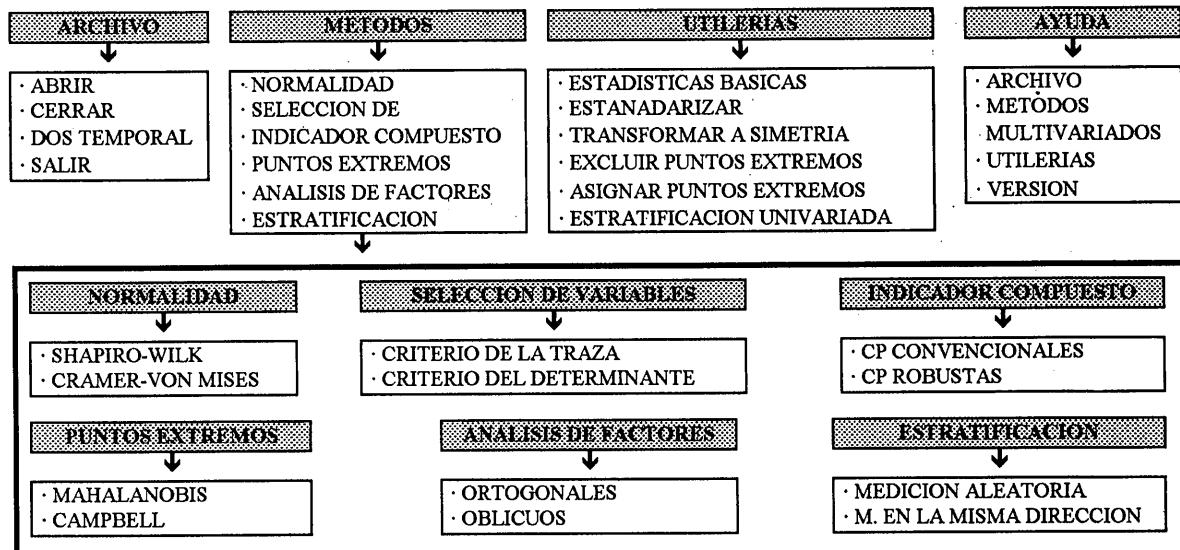
El SAM es un sistema interactivo de fácil operación programado en lenguaje C⁺⁺ a base de menús y submenús. Además de los métodos multivariados descritos en el apartado anterior, el sistema incluye una librería con diversos procedimientos de apoyo para complementar el análisis, por ejemplo, *estadísticas básicas*, *estandarización de variables*, *transformación de distribuciones marginales no simétricas a simétricas*, *exclusión de puntos extremos*, *asignación de observaciones discrepantes a estratos definidos*, *estratificación univariada* y otros; asimismo, consta de procedimientos y funciones generales para manipular archivos, memoria, monitor e impresora. El esquema básico se presenta en la tabla 1.

Los requerimientos mínimos son los siguientes:

Sistema operativo MS-DOS 3.0 o posterior	640 Kb o más en RAM
Procesador 80386 o posterior	Monitor VGA (color)
2 Mb de memoria extendida, sobre todo si los procesos se realizan con archivos grandes	
Disco duro con capacidad para instalar el sistema y área de trabajo	

Se cuenta con disco de instalación que contiene el archivo INSTALA.EXE, basta con teclear la palabra "INSTALA" para crear automáticamente los subdirectorios necesarios y sus correspondientes archivos. Se requiere de al menos 3 Mb de espacio libre en disco duro para instalar el SAM en su totalidad. Una vez que el sistema haya sido instalado, para su activación, se teclea la palabra "SAM"; así aparecerá una pantalla de presentación seguida del menú principal. Los archivos de entrada deben ser del tipo ASCII delimitado por blancos y estructurados en forma de matriz de datos. Los registros pueden llevar consigo un primer campo identificador de longitud fija y un máximo de 40 caracteres alfanuméricos. Los archivos de resultados pueden verse en pantalla, o bien almacenarse en disco para su posterior impresión según el formato requerido.

Tabla 1. S A M (Versión 1.0)



5. COMENTARIOS FINALES

El sistema presentado en este trabajo ha sido de gran utilidad en el desarrollo de proyectos nacionales encaminados a conocer aquellos grupos de población “pobre” y “vulnerable”. El SAM es un producto que alienta y resalta las investigaciones nacionales puestas en práctica para resolver problemas que en un principio afectan a nuestros conciudadanos, y al mismo tiempo es un intento modesto por salir de la dependencia tecnológica en ese campo. Su fácil operación mediante una comunicación máquina-usuario en español, así como el diseño de los formatos de salida con la mínima información requerida, permite fomentar la cultura estadística entre los investigadores de otras áreas. En caso de requerir alguno de los procedimientos que incluye el sistema, es posible amortiguar el alto costo económico que implica la compra de un paquete estadístico comercial y la correspondiente capacitación para operarlo.

REFERENCIAS

- Campbell, N.A.(1980). Robust procedures in multivariate analysis. I : Robust covariance estimation. *J. of Appl. Statist.*, **29**, 231-237.
- Jarque, C. M. (1981). A solution to the problem of optimum stratification in multivariate sampling. *J. of the Roy. Statist. Soc. C*. **30**, 163-169.
- Malkovich, J. F. and Afifi, A.A. (1973). On test for multivariate normality. *J. of the Am. Statist. Assoc.*, **68**, 176-179.
- McCabe, G.P. (1984). Principal variable. *Technometrics*, **26**, 137-144.
- Reyment, R. A., Blackith, R. E., and Campbell, N. A. (1984). *Multivariate Morphometry*.(2nd ed.) Academic Press, London.

- Rodríguez, S. (1995). Análisis de componentes principales robustos. *Tesis de Licenciatura*, Facultad de Matemáticas, Universidad de Guanajuato.
- Vences, J. y Cossío, F.V. (1989). Comparación de los procedimientos de Shapiro-Wilk y Cramér-von Mises para probar normalidad multivariada. *Agrociencia*, 75, 63-75.
- Vences, J. (1994). Un procedimiento para la selección de variables por componentes principales. *Memoria del IX Foro Nacional de Estadística*, INEGI-AME. pp. 109-112.
- Vences, J. (1996). Estadística Multivariada, Análisis de Factores. Libro que consta de 210 pp. (en proceso de publicación).

Esta publicación consta de 410 ejemplares y se terminó de imprimir en el mes de agosto de 1997 en los talleres gráficos del **Instituto Nacional de Estadística, Geografía e Informática**
Av. Héroe de Nacozari Núm. 2301 Sur, Acceso 11, P.B.
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.
México

