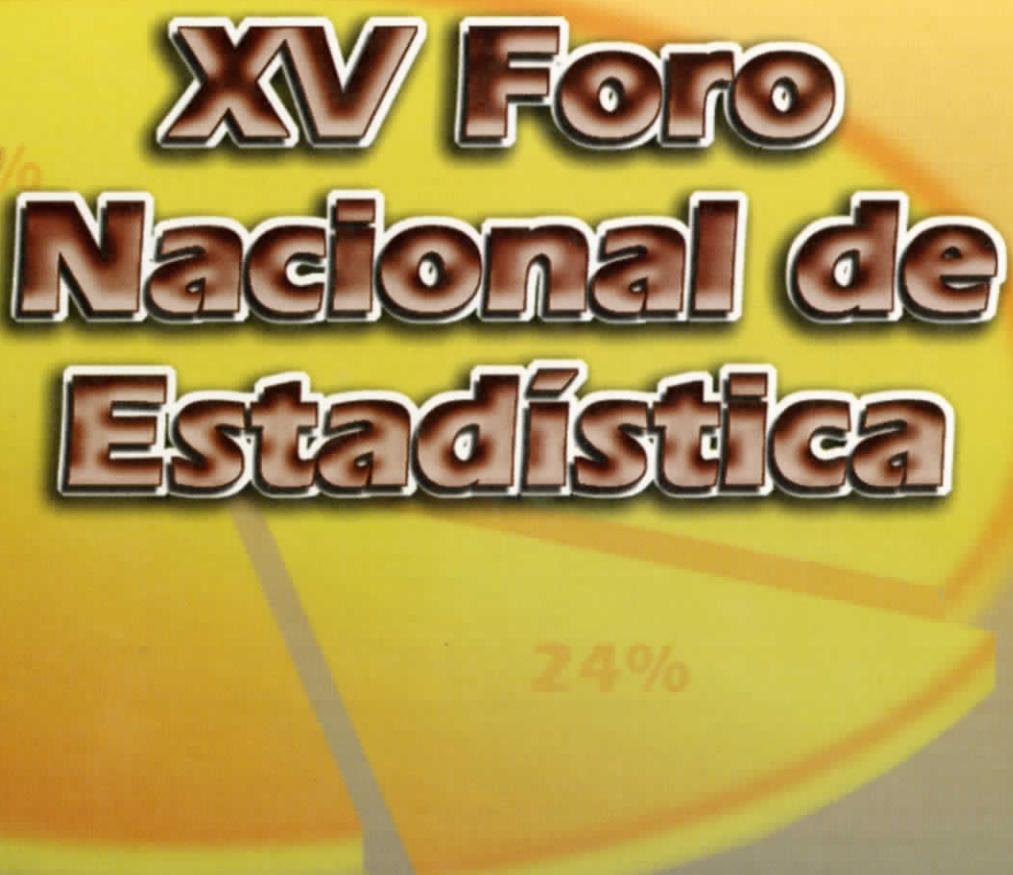




# **MEMORIAS**



## **XV Foro Nacional de Estadística**

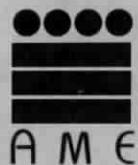
**UNIVERSIDAD AUTONOMA METROPOLITANA, UNIDAD AZCAPOTZALCO, MEXICO, D.F.  
16 AL 20 DE OCTUBRE DE 2000.**



# **MEMORIAS**

# **XV Foro Nacional de Estadística**

**UNIVERSIDAD AUTONOMA METROPOLITANA, UNIDAD AZCAPOTZALCO, MEXICO, D.F.  
16 AL 20 DE OCTUBRE DE 2000.**



DR © 2001, **Instituto Nacional de Estadística,  
Geografía e Informática**  
**Edificio Sede**  
**Av. Héroe de Nacozari Núm. 2301 Sur**  
**Fracc. Jardines del Parque, CP 20270**  
**Aguascalientes, Ags.**

[www.inegi.gob.mx](http://www.inegi.gob.mx)  
[atencion.usuarios@inegi.gob.mx](mailto:atencion.usuarios@inegi.gob.mx)

**Memorias**  
**XV Foro Nacional de Estadística**

**Impreso en México**  
**ISBN 970-13-3632-1**

## **Presentación**

El XV Foro Nacional de Estadística se llevó a cabo del 16 al 20 de octubre de 2000, en la Casa del Tiempo de la Universidad Autónoma Metropolitana y fue organizado por la Unidad Azcapotzalco, en la Ciudad de México.

Entre otras actividades se presentaron 48 contribuciones libres y 6 conferencias magistrales y un curso de consultoría. En estas memorias se presentan resúmenes de dichas contribuciones. Todos los resúmenes recibidos fueron incluidos sin un proceso de arbitraje, aunque con una detallada revisión.

La Asociación Mexicana de Estadística, agradece a la Universidad Autónoma Metropolitana Unidad Xochimilco y a la Casa del Tiempo de la misma universidad su apoyo para la realización de este foro y al Instituto Nacional de Estadística Geografía e Informática el apoyo para la edición de estas memorias.

*El Comité Editorial*  
**Eduardo Castaño Tostado**  
**José M. González Barrios M.**  
**Alberto Molina Escobar**

# Contenido

<b>Presentación</b>	<b>iii</b>
<b>Inferencia en distribuciones estables</b>	<b>1</b>
<i>Alegria, A. y Alvarez, E.</i>	
<b>Statistical inference for mixtures of distributions for censored data with partial identification</b>	<b>7</b>
<i>Contreras, A., O'Reilly, F. y Gutiérrez, E.</i>	
<b>Optimización estadística del proceso de densificación mécanica</b>	<b>15</b>
<i>Domínguez, J., González, M. y Muñoz, G.</i>	
<b>Modelación estocástica, específicamente como cadena de Markov, aplicada a dos etapas de la producción del hongo Seta: Pleurotus Ostreatus</b>	<b>21</b>
<i>Lara, J. y Ayala, N.</i>	
<b>Cómputo de probabilidades de error en el conteo rápido</b>	<b>27</b>
<i>O'Reilly, F. y Rueda, R.</i>	
<b>Uso de componentes principales y correlación canónica para dasometría de Cirián</b>	<b>33</b>
<i>Padrón, E., Méndez, I., Hernández, N. y Olivares, E.</i>	
<b>Estimación de los parámetros de la curva de lactancia de vacas Holstein Friesian sometidas a un programa de somatotropina bovina recombinante</b>	<b>39</b>
<i>Rivero, G., Rosas, G. y Avila, R.</i>	
<b>Representación gráfica de los resultados de las elecciones presidenciales de 1994 y 2000 en México</b>	<b>47</b>
<i>Romero P., Eslava, G. y Méndez, I.</i>	

<b>Análisis de datos de mediciones repetidas utilizando metodología de modelos mixtos</b>	<b>53</b>
<i>Rosas, G., Avila, R., Rivero, G. y Avila, B.</i>	
<b>Casandra, software de apoyo didáctico para la enseñanza de la estadística</b>	<b>61</b>
<i>Sánchez, F. y Martínez, A.</i>	
<b>Métodos estadísticos en la normatividad en metrología</b>	<b>67</b>
<i>Segura, C. y Castaño, E.</i>	
<b>Estadística y metrología</b>	<b>75</b>
<i>Villa, E.</i>	

# Inferencia en Distribuciones Estables

Alejandro Alegría

Elia Alvarez

*Instituto Tecnológico Autónomo de México*

## 1 Introducción

En el estudio del comportamiento estocástico del cambio en el precio de las acciones de un mercado financiero, se han propuesto modelos basados en la evidencia empírica de que las distribuciones de los rendimientos de las cotizaciones bursátiles presentan cierto grado de asimetría y una curtosis mayor a la esperada bajo el supuesto de normalidad. Estos modelos, entre ellos el de Fama y Roll (1968), proponen la conveniencia de usar familias de distribuciones que consideran estas características de las distribuciones empíricas. El uso de las distribuciones llamadas *G - H* es una solución alternativa que permite inferir sobre los parámetros de una distribución con el comportamiento anteriormente citado.

Por otro lado, el uso de distribuciones estables asimétricas es otra solución que permite modelar distribuciones con colas más pesadas respecto a la distribución normal.

## 2 Distribuciones *G - H*

Las distribuciones *G - H*, introducidas por Tukey (1977) y Hoaglin (1985), constituyen una familia de distribuciones con 2 parámetros,  $g$  y  $h$ , que están definidos en términos de cuantiles de la distribución normal estándar. El parámetro  $g$  determina la asimetría, mientras que el parámetro  $h$  controla la elongación o extensión (curtosis) de las colas de la distribución.

Consideremos la siguiente función

$$Q_{g,h}(z) = \frac{e^{gz} - 1}{g} e^{hz^2/2}, \quad z, g, h \in \mathbb{R}.$$

Ahora se define la variable aleatoria  $Y = Q_{g,h}(Z)$ , donde  $Z \sim N(0, 1)$ . Si  $y_p$  y  $z_p$  son los cuantiles de orden  $p$  de  $Y$  y  $Z$ , respectivamente, se desea que  $y_p = Q_{g,h}(z_p)$ . La distribución

de la variable  $Y$  se denomina distribución  $G-H$ , y los parámetros son  $g$  y  $h$ . Para tomar en cuenta parámetros de localización y escala, se propone la transformación

$$X = A + BY = A + BQ_{g,h}(Z).$$

### 3 Subfamilias

Cuando  $h = 0$ ,

$$\frac{e^{gz} - 1}{g},$$

y se dice que  $Y$  tiene una  $G$  – *distribución*. La distribución presenta asimetría a la derecha cuando  $g > 0$ , asimetría a la izquierda si  $g < 0$ , y es simétrica si  $g = 0$ .

Si ahora  $g = 0$ ,

$$Y = Z e^{hZ^2/2},$$

y  $Y$  tiene una  $H$  – *distribución*, cuyas colas son más pesadas o menos pesadas con respecto a una distribución normal, dependiendo si  $h > 0$  ó  $h < 0$ , respectivamente.

### 4 Estimación

A menos de que el parámetro  $g$  sea cero, la estimación de los parámetros inicia con  $g$ , Mills (1995). Ya que los cuantiles  $x_p$  están relacionados con los cuantiles  $y_p$ , es fácil ver que

$$x_{0.5} = A + BQ_{g,h}(z_{0.5}) = A + BQ_{g,h}(0) = A.$$

Para  $p \in (0, 0.5)$ , se tiene que  $x_p = x_{0.5} + BQ_{g,h}(z_p)$  y  $x_{1-p} = x_{0.5} + BQ_{g,h}(z_{1-p})$ , de donde es sencillo demostrar que

$$g = \frac{1}{z_p} \log \left( \frac{x_{0.5} - x_p}{x_{1-p} - x_{0.5}} \right)$$

Usando diversos valores de  $p$ , un posible estimador de  $g$  es la mediana de los valores de  $g$  obtenidos para cada  $p$ . La estimación de  $h$  es condicional al valor estimado de  $g$ , ya que para  $g \neq 0$ , se obtiene

$$\log \left( (x_p - x_{1-p}) \frac{g}{e^{gz_p} - e^{-gz_p}} \right) = \log(B) + h z_p^2 / 2,$$

mientras que para  $g = 0$ ,

$$\log((x_p - x_{1-p})/2) = \log(B) + h z_p^2/2 .$$

Los parámetros  $B$  y  $h$  se pueden estimar al notar que  $\log(B)$  y  $h$  son la ordenada al origen y la pendiente, respectivamente, de una relación lineal en  $z_p^2$ . Si en lugar de usar los cuantiles  $x_p$  y  $x_{1-p}$  se usan  $x_{0.5}$  y  $x_{1-p}$ , es posible construir estimadores para la mitad superior de los datos, y con  $x_p$  y  $x_{0.5}$ , estimadores para la otra mitad.

Para tener mayor flexibilidad se puede generalizar lo anterior suponiendo que los parámetros  $g$  y  $h$  no son constantes, por ejemplo,

$$g = g_0 + g_1 Z^2 + \dots , \quad h = h_0 + h_1 Z^2 + \dots$$

## 5 Ejemplo

A manera de ejemplo, se trabajó con 1155 observaciones del Indice de Precios y Cotizaciones de la Bolsa Mexicana de Valores (IPC), correspondientes al periodo 3/01/1996 a 6/09/2000. El histograma de la figura 1 muestra frecuencias observadas y esperadas del cambio en el IPC, éstas últimas suponiendo normalidad. Lo que se observa es un pequeño sesgo a la derecha.

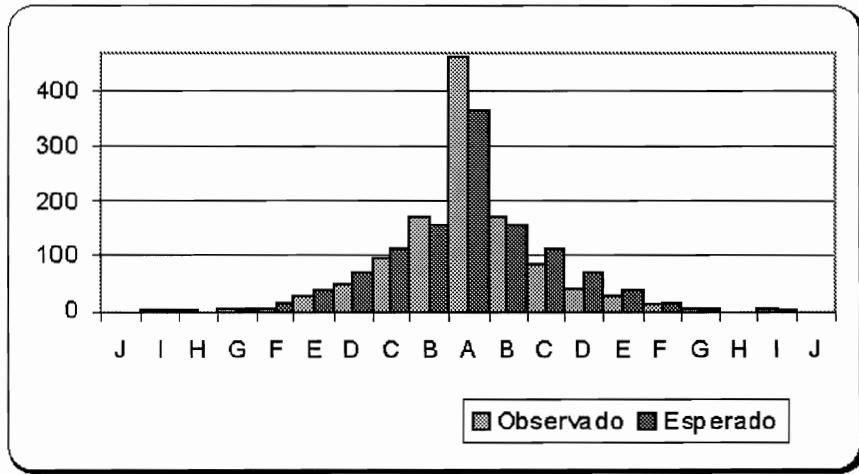


Figura 1

El ajuste de una distribución  $G - H$  a los datos del IPC dió como resultado lo siguiente,

$$A = 0.00029$$

$$B = \begin{cases} 0.01411, & p > 0.5 \\ 0.01365, & p < 0.5 \end{cases} , \quad g(z_p) = 0.22354 - 0.2334z_p^2 + 0.082z_p^4 - 0.0083z_p^6$$

$$h = \begin{cases} 0.17046 - 0.0105z_p^2 + 0.0012z_p^4, & p > 0.5 \\ 0.31544 - 0.0715z_p^2 + 0.0077z_p^4, & p < 0.5 \end{cases}, \quad SC_e = 5.89888E - 05$$

La bondad del ajuste se puede apreciar en la figura 2 en donde aparecen valores cuantiles observados y ajustados.

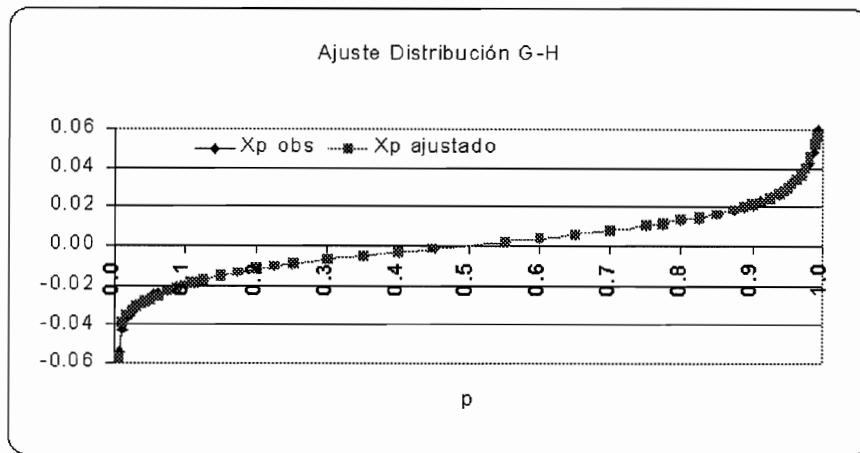


Figura 2

## 6 Otros modelos

La familia de distribuciones estables ofrece otra alternativa para analizar variables que presentan ciertas características de asimetría y curtosis. Teóricamente esta familia ya ha sido estudiada desde 1924 por Levy, y también por Gnedenko y Kolmogorov en 1954. Su utilización en inferencia estadística ha presentado dificultades por que no se tiene una expresión analítica para la función de densidad de una distribución estable (salvo casos particulares como la normal o Cauchy). Algunos trabajos son: Fama y Roll (1968), Fielitz y Smith (1972), Leitch y Paulson (1975), Press (1972). Un resultado interesante es el que presentan Lambert y Lindsey (1999). Ellos utilizan la transformada de Fourier para obtener una expresión de la función de densidad de una distribución estable, y a partir de ella construyen la función de verosimilitud correspondiente. La verosimilitud es una función complicada, así que ahora el problema es básicamente numérico. La ventaja es que se pueden construir intervalos de confianza y tambien incluir covariables en el análisis.

En 1995, Buckle propone realizar inferencia Bayesiana sobre distribuciones estables. Debido a la complejidad numérica del problema, se aplica simulación Monte Carlo utilizando cadenas de Markov.

## 7 Conclusiones

La distribuciones  $G - H$  proporcionan un método, relativamente sencillo, de modelar el comportamiento de datos que presentan cierta asimetría y curtosis. En general, el uso que se ha hecho hasta ahora de estas distribuciones ha sido más bien de tipo descriptivo.

Otra alternativa es la familia de distribuciones estables. No obstante, hace falta mayor análisis numérico.

## Referencias

- Buckle, D.J. (1995). Bayesian Inference for Stable Distributions. *Journal of the American Statistical Association*, **90**, 605-613.
- Fama, E.F. y Roll,R. (1968). Some Propieties of Symmetric Stable Distributions. *Journal of the American Statistical Association*, **63**, 817-836.
- Fielitz, B.D. y Smith E. W. (1972). Asymmetric Stable Distributions of Stock Price Changes. *Journal of the American Statistical Association*, **67**, 813-814.
- Gnedenko,B.V. y Kolmogorov, A.N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Adisson-Wesley, Cambridge.
- Hoaglin, D.C. (1985). Summarizing shape numerically: the g-and-h distributions. In *Exploring Data Tables, Trends and Shapes*. (Eds. D.C. Hoaglin, F. Mosteller and J.W. Tukey) pp. 461-513. Wiley, New York.
- Lambert, P. y Lindsey, J.K. (1999). Analysing Financial Returns by Using Regression Models Based on Non-Symmetric Stable Distributions. *Appl. Statist*, **48**, 409-424.
- Leitch, R.A. y Paulson, A.S. (1975). Estimation of Estable Law Parameters: Stock Price Behavior Application. *Journal of the American Statistical Association*, **70**, 690-697.
- Lévy, P. (1924). Théorie des erreurs. La loi de Gauss et les Lois Exceptionnelles. *Bulletin-Société Mathématique de France*, **52**, 49-85.
- Press, S. J. (1972) Estimation in Univariate and Multivariate Stable Distributions. *Journal of the American Statistical Association*, **67**, 842-846.
- Tukey, J.W. (1977). *Exploratory Data Analysis*. Adisson-Wesley Reading.



# Statistical Inference for Mixtures of Distributions for Censored Data with Partial Identification

**Alberto Contreras**

**Federico O'Reilly**

**Eduardo Gutiérrez**

*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,  
Universidad Nacional Autónoma de México*

## 1 Introduction

Consider the mixture of  $k$  distributions

$$G(\cdot; \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{j=1}^k \pi_j F_j(\cdot; \theta_j),$$

where  $\{\pi_j : \pi_j > 0, j = 1, 2, \dots, k; \sum_{j=1}^k \pi_j = 1\}$  and  $\boldsymbol{\theta}' = \{\theta_j : j = 1, 2, \dots, k\}$  are unknown.

The information obtained from  $N$  independent observations from  $G$  can be summarized as

$$\{x_{1i}\}_{i=1}^{r_1}, \{x_{2i}\}_{i=1}^{r_2}, \dots, \{x_{ki}\}_{i=1}^{r_k} \quad \text{and } N - \sum_{j=1}^k r_j \equiv N - r,$$

where, for each  $j = 1, 2, \dots, k$ ,  $x_{j1}, \dots, x_{jr_j}$  denote the  $r_j$  observations which are less than or equal to  $C$  and identified as coming from distribution  $F_j$ . In the notation above,  $N - r$  stands for the number of observations that exceed the value  $C$ .

The likelihood function  $L(\boldsymbol{\theta}, \boldsymbol{\pi}; \mathbf{x})$  is proportional to

$$\left\{ \prod_{j=1}^k \prod_{i=1}^{r_j} \frac{f_j(x_{ji}; \theta_j)}{F_j(C; \theta_j)} \right\} \prod_{j=1}^k \{\pi_j F_j(C; \theta_j)\}^{r_j} \times \{1 - G(C; \boldsymbol{\theta}, \boldsymbol{\pi})\}^{N-r}. \quad (1)$$

Mendenhall and Hader (1958) and Díaz-Francés (1998) use the above scheme for  $k = 2$ . In the latter work a profile likelihood analysis (PLA) is used, and the functions  $F_j$ ,  $j = 1, 2$ , are taken to be Weibull with different shape and scale parameters.

## 2 Profile likelihood analysis via latent variables

For each  $j = 1, 2, \dots, k$ , let  $r_j^*$  be the number of unobserved values exceeding the limit  $C$  and corresponding to distribution  $F_j$ . Then  $r_j^* \geq 0$  and  $\sum_{j=1}^k r_j^* = N - r$ .

Assuming that the  $r_j^*$ ;  $j = 1, 2, \dots, k$ , are known, then the weights  $\pi_1, \dots, \pi_k$  are irrelevant and  $L(\boldsymbol{\theta})$  can be written as  $L(\boldsymbol{\theta}) \propto \prod_{j=1}^k L_j(\theta_j)$ , where

$$L_j(\theta_j) \propto \left\{ \prod_{i=1}^{r_j} \frac{f_j(x_{ji})}{F_j(C; \theta_j)} \right\} \binom{r_j + r_j^*}{r_j} \{F_j(C; \theta_j)\}^{r_j} \{1 - F_j(C; \theta_j)\}^{r_j^*}.$$

Let  $\hat{\theta}_j(r_j^*)$  be the maximum likelihood estimate for  $\theta_j$  for  $r_j^*$  fixed. The corresponding profile likelihood function for  $r_1^*, r_2^*, \dots, r_k^*$ , is

$$\prod_{j=1}^k L_j(\hat{\theta}_j(r_j^*)). \quad (2)$$

We search for the values  $\hat{r}_j^*$ ,  $j = 1, 2, \dots, k$ , which maximize (2).

The *overall maximum likelihood* estimates are given by

$$\hat{\theta}_j(\hat{r}_j^*), \quad j = 1, 2, \dots, k. \quad (3)$$

We refer to this scheme as the latent variable profile likelihood analysis (LVPLA).

The computational advantages over the PLA are that it is easier to evaluate the profile likelihood function for each parameter and the Normal approximations are easier to be obtained.

The advantages of the LVPLA become greater as the number of populations  $k$  increases. Here we analyze a simulated data set from a mixture of  $k = 3$  Weibull distributions with different shape and scale parameters. Our data were generated using  $\pi_1 = 0.2954$  and  $\pi_2 = 0.55$ . The values for the other parameters are in Table 1 under the heading “True value”. Here  $N = 500$  and  $r_1 = 136$ ,  $r_2 = 238$ ,  $r_3 = 81$ . Correspondingly,  $N - r = 45$ .

Table 1: Comparison of m.l. estimates and true values,  $k = 3$ .

	m.l.e. via LVPLA	True value
$\mu_1$	-1.056	-1.06
$\mu_2$	-0.55779	-0.5758
$\mu_3$	-0.79648	-0.7078
$\log \sigma_1$	-0.30151	-0.2525
$\log \sigma_2$	-0.16281	-0.1161
$\log \sigma_3$	-0.68543	-0.5013

Table 2: Comparison of results for LVPLA and BALV,  $k = 3$ .

Upper and lower limits for 95 % intervals and highest posterior density regions.

	LVPLA	BALV
$\mu_1$	(-1.1868,-0.9257)	(-1.1976,-0.8664)
$\mu_2$	(-0.6665,-0.4490)	(-0.7020,-0.4597)
$\mu_3$	(-0.91147,-0.6815)	(-0.9083,-0.5959)
$\log \sigma_1$	(-0.4365,-0.1665)	(-0.4039,-0.1426)
$\log \sigma_2$	(-0.2799,-0.0456)	(-0.2609,-0.0823)
$\log \sigma_3$	(-0.8606,-0.5102)	(-0.7868,-0.4211)

Figure 1 shows the standarized profile likelihood function for  $(r_1^*, r_2^*)$ . The values that maximize this function are  $\hat{r}_1^* = 2$  and  $\hat{r}_2^* = 43$ , which in turn implies that  $\hat{r}_3^* = 0$ .

Figure 2 displays the profile likelihood functions for the location parameters obtained when we consider  $r_1^* = 2$  and  $r_2^* = 43$ . The same figure includes the corresponding normal approximations. Regarding the log-scale parameters, the respective plot is in Figure 3. From Table 1, we see that the maximum likelihood estimates are close to the true values of the parameters used to generate the mixture of  $k = 3$  populations.

Table 2 reports limits for the confidence intervals obtained via the normal approximations.

### 3 Statistical modelling from a Bayesian perspective

Diebolt and Robert (1994) discuss a sampling-based approach to the Bayesian analysis of finite mixture distributions. Our approach is similar, but takes into account the additional information provided by the partial identification of the populations and allows for censoring.

Equation (1) states the likelihood function for  $\Theta' \equiv (\boldsymbol{\theta}', \boldsymbol{\pi}')$ . Thus, if  $p(\Theta)$  is a prior

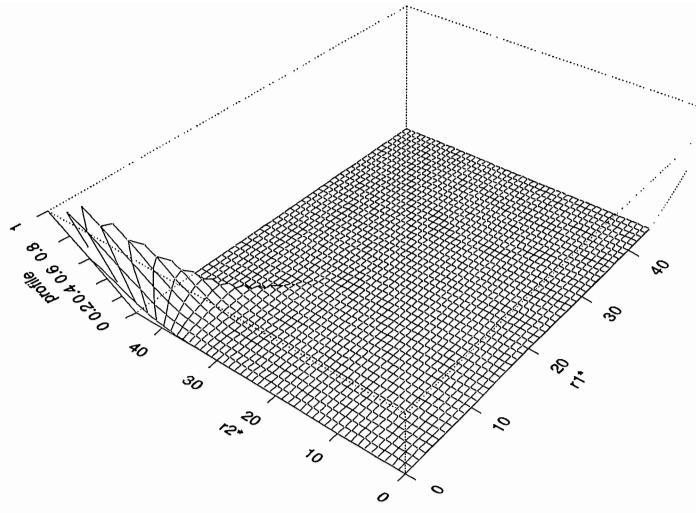


Figure 1: Standardized profile likelihood function of  $r_2^*$  and  $r_2^*$ .

distribution for  $\Theta$ , we have that  $p(\Theta|\mathbf{x}) \propto p(\Theta)L(\Theta; \mathbf{x})$ . We shall consider vague priors of the form  $p(\Theta) \propto 1$ , so that

$$p(\Theta|\mathbf{x}) \propto L(\Theta; \mathbf{x}). \quad (4)$$

## 4 The use of latent variables

Let  $\tilde{\mathbf{x}}'_j \equiv \{\tilde{x}_{ji}\}_{i=r_j+1}^{r_j+r_j^*}$  be the additional data that would have been observed from the  $j$ -th population had the samples not been censored. Consider a Gibbs sampler scheme alternating between  $\Theta$  and  $\tilde{\mathbf{x}}$ . The corresponding full conditionals

$$p(\Theta|\mathbf{x}, \tilde{\mathbf{x}}) \quad (5)$$

and

$$p(\tilde{\mathbf{x}}|\Theta, \mathbf{x}) = p(\tilde{\mathbf{x}}|\Theta). \quad (6)$$

For equation (5), suppose that there are no censored data. Then

$$L(\Theta; \mathbf{x}, \tilde{\mathbf{x}}) \propto \prod_{j=1}^k \left\{ \pi_j^{r_j} \prod_{i=1}^{r_j} f_j(x_{ji}; \theta_j) \right\} \prod_{j=1}^k \left\{ \pi_j^{r_j^*} \prod_{i=r_j+1}^{r_j+r_j^*} f_j(\tilde{x}_{ji}; \theta_j) \right\}. \quad (7)$$

From (4), the posterior distribution  $p(\Theta|\mathbf{x}, \tilde{\mathbf{x}})$  is proportional to the right hand side of (7).

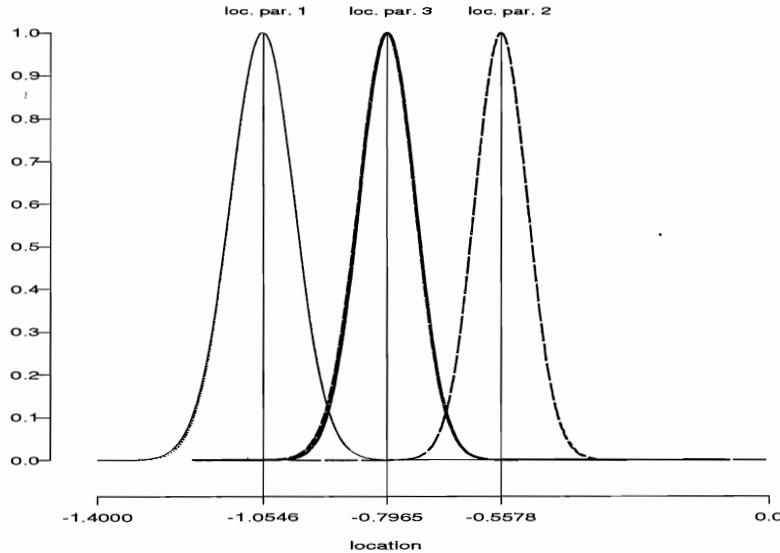


Figure 2: Profile likelihood functions for location parameters.

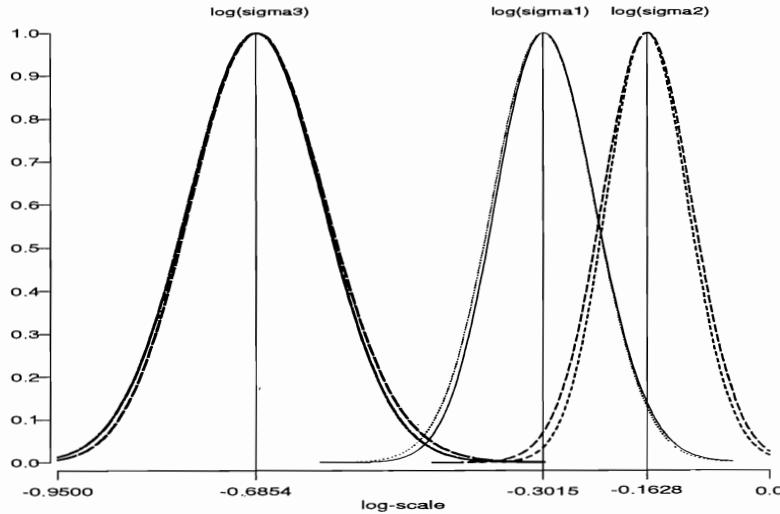


Figure 3: Profile likelihood functions for log-scale parameters .

Given  $\boldsymbol{\theta}$  and  $\boldsymbol{\pi}$  and by the conditional independence of  $\mathbf{x}$  and  $\tilde{\mathbf{x}}$  (see 6), we have

$$p(\tilde{\mathbf{x}}|\Theta) = \prod_{j=1}^k \prod_{i=r_j+1}^{r_j+r_j^*} g^*(\tilde{x}_{ji}|\Theta),$$

where the density  $g^*(\cdot|\Theta)$  of the censored observations for each of the  $k$  populations is given by

$$g^*(\tilde{x}_{ji}|\Theta) = \frac{g(\tilde{x}_{ji}|\Theta)}{1 - G(C; \Theta)} \quad (8)$$

We will refer to this scheme as the Bayesian analysis using latent variables (BALV). For references on Gibbs sampler see Gamerman (1997) and Gelman et al (1995).

Let  $Y'_j = (\mathbf{x}'_j, \tilde{\mathbf{x}}'_j)$ ,  $j = 1, 2, 3$ . Each  $F_j$ ,  $j = 1, 2, 3$ , is modeled as a Weibull distribution with different shape and scale parameters. The following probability density functions are the full conditional densities found from  $p(\Theta|\mathbf{x}, \tilde{\mathbf{x}})$ :

$$\begin{aligned} p(\pi_j | \boldsymbol{\pi}_{-j}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}, \tilde{\mathbf{x}}) &= p(\pi_j | Y_1, Y_2, Y_3) \\ &\propto \pi_j^{r_j + r_j^*} (1 - \pi_j)^{\rho_j - 1}, \end{aligned}$$

where  $\rho_j = \sum_{i=1}^3 (r_i + r_i^* + 1) - (r_j + r_j^* + 1)$ ,  $j = 1, 2, 3$ .

$$\begin{aligned} p(\mu_j | \boldsymbol{\pi}, \boldsymbol{\sigma}, \boldsymbol{\mu}_{-j}, \mathbf{x}, \tilde{\mathbf{x}}) &= p(\mu_j | \sigma_j, Y_j) \\ &\propto e^{-(r_j + r_j^*)\mu_j/\sigma_j} \cdot \exp \{-e^{-(\mu_j - s_j)/\sigma_j}\}, \end{aligned}$$

where  $s_j = \sigma_j \cdot \log \left( \sum_{i=1}^{r_j + r_j^*} \exp \{Y_{j,i}/\sigma_j\} \right)$ ,  $j = 1, 2, 3$ .

$$\begin{aligned} p(\sigma_j | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}_{-j}, \mathbf{x}, \tilde{\mathbf{x}}) &= p(\sigma_j | \mu_j, Y_j) \\ &\propto \frac{e^{\frac{(r_j + r_j^*)(\bar{Y}_j - \mu_j)}{\sigma_j}}}{\sigma_j^{r_j + r_j^*}} \cdot \exp \left\{ - \sum_{i=1}^{r_j + r_j^*} e^{(Y_{j,i} - \mu_j)/\sigma_j} \right\}, \end{aligned}$$

where  $\bar{Y}_j = \sum_{i=1}^{r_j + r_j^*} Y_{j,i}/(r_j + r_j^*)$ ,  $j = 1, 2, 3$ .

A Metropolis-Hastings step based on a normal approximation is used to produce samples of the full conditional density for  $\sigma_j$ ,  $j = 1, 2, 3$ . Regarding the samples from  $p(\tilde{\mathbf{x}}|\Theta)$ , note that the density function in (8) is a mixture of Gumbel densities, with weights  $\pi_1, \pi_2$ , and  $1 - (\pi_1 + \pi_2)$ , which is truncated on the left at zero.

Figure 4 displays the corresponding plots of the density estimates for each of the location and log-scale parameters. Dotted vertical lines indicate the limits of a highest posterior density regions ( $p=0.95$ ) for each case.

## 5 Concluding remarks

The LVPLA and BALV methods produce similar results and for both schemes there are computational advantages over the PLA scheme. However, the Bayesian approach takes full

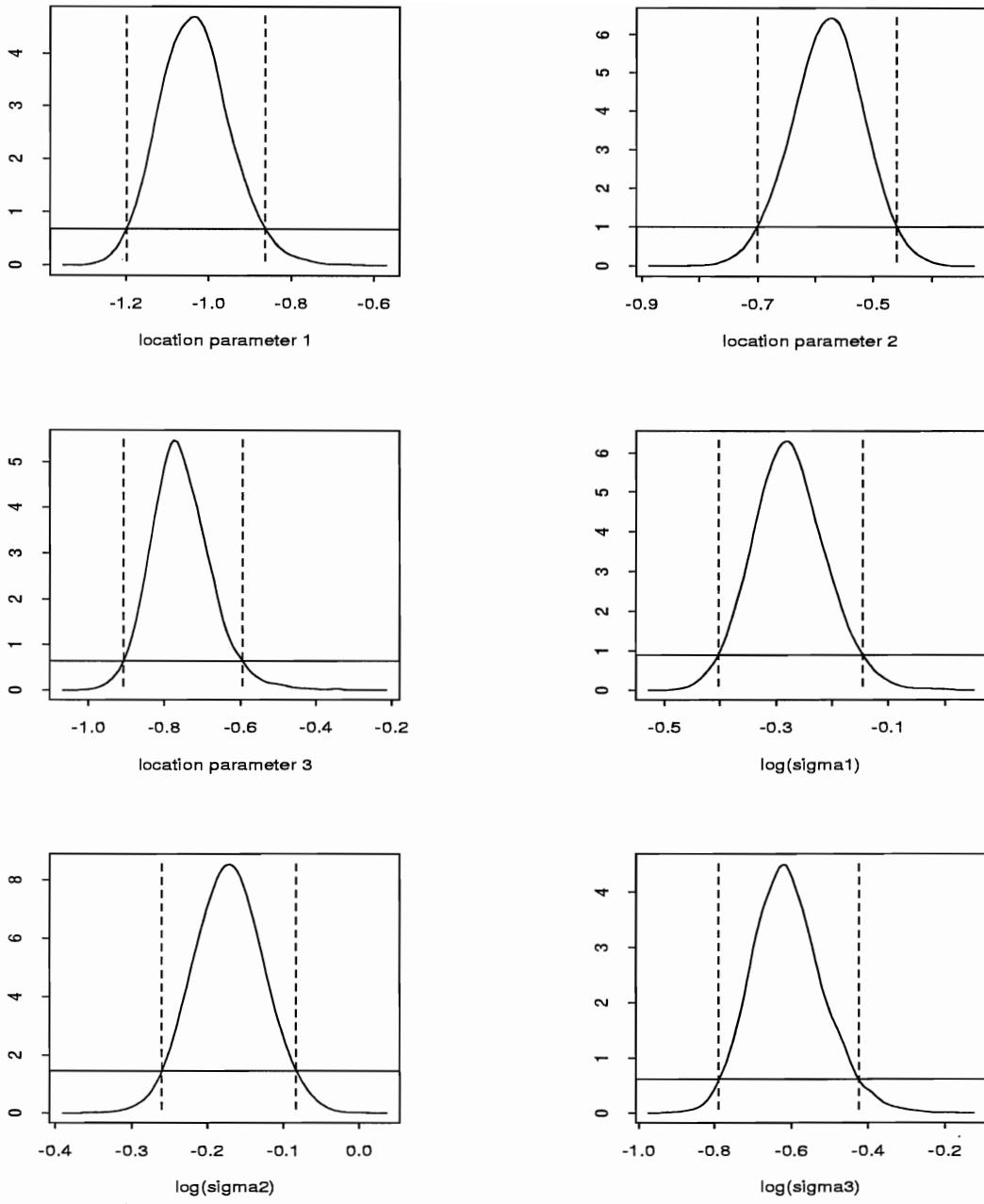


Figure 4: Density estimates from posterior samples,  $k = 3$  populations.

account of the additional uncertainty due to the lack of full identification of the populations. Moreover, the Bayesian analysis based on the Gibbs sampler is more efficient than the classical LVPLA, which relies on a systematic search of the maximum value of the likelihood over all possible combinations of the values of  $r_1^*, r_2^*, \dots$ . Such a gain in efficiency is more evident as the number of populations in the mixture increases.

## References

- Díaz Francés, E. (1998). *Scientific application of maximum likelihood in multiparametric problems*. Ph.D, Dissertation. CIMAT, México.
- Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 363-375.
- Gamerman, D. (1997). *Markov Chain Monte Carlo. Stochastic simulation for Bayesian inference*. Chapman and Hall: London.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman and Hall: London.
- Mendenhall, W. and R.J. Hader (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, **45**, 504-520.

# Optimización Estadística del Proceso de Densificación Mécanica

**Jorge Domínguez Domínguez**

*Centro de Investigación en Matemáticas. A.C., Guanajuato.*

**Miguel González Valadez**

*Centro de Investigación y Asistencia Técnica del Estado de Querétaro.*

**Guillermo Muñoz Hernández**

*Centro de Investigación y Asistencia Técnica del Estado de Querétaro.*

## 1 Introducción

La densificación de la biomasa significa el uso de alguna forma de presión mecánica para reducir el volumen de la materia vegetal y su conversión a una forma sólida, la cual es más fácil de manejar y almacenar que el material original (Erickson y Prior, 1990).

En la actualidad, el proceso de densificación ha causado un gran interés en los países desarrollados de todo el mundo, como una técnica de aprovechamiento de los residuos para su utilización como una fuente de energía (Batacharya, 1989). Sin embargo, el proceso no es nuevo, en uso existen al menos 4 métodos de densificación empleados en las máquinas comerciales: empacado, cubicado, peletizado y briqueteado; ya sea por prensas de pistón, tornillos de extrusión o por prensas de rodillos. Estos equipos pueden ser incluidos en dos grandes categorías en que se dividen las técnicas de densificación: tipo A, o densificación caliente y alta presión; y tipo B, o densificación fría y baja presión (Batacharya, 1989).

A partir de una extensa revisión bibliográfica, se encontró que en el proceso de densificación para diferentes materiales fibrosos están presentes los factores: contenido de humedad, temperatura, presión, tamaño de partículas, contenido de aglutinante y tiempo de relajación. Estos factores son los que se consideraron en este estudio. Por razones de aceptación de los residuos como alimento, se seleccionó la elaboración de alimentos para rumiantes, donde las variables de respuesta apropiadas son: la densidad, la durabilidad y el consumo de energía específico. Una densidad adecuada de los cubos para alimentación de bovinos debe ser mayor

a 650 kg/m<sup>3</sup> (Brunh, 1957), con una durabilidad de 90% como mínimo para garantizar una buena resistencia del cubo al manejo mecánico (Macarthur, 1981) y el menor consumo de energía. El óptimo en este proceso debe satisfacer esas restricciones.

El objetivo de este trabajo es llevar a cabo una estrategia experimental para establecer las condiciones que determinen un óptimo común para: la densidad, durabilidad y consumos de energía en la densificación de residuos agrícolas. Con ese propósito, es necesario tener modelos que se ajusten de manera adecuada.

## 2 Planeación experimental

El proceso comercial para la densificación conocido como cubicado, fue el que se empleó en el presente estudio. Las etapas identificadas que suceden en el proceso de cubicado son: compresión, relajación de esfuerzos y expansión. También, se determinaron y calcularon los parámetros suficientes que sirvieron como condiciones para el diseño del equipo de pruebas, éste incluyó un dispositivo hidráulico (presa y unidad hidráulica).

En la fase inicial de esta investigación se realizó un experimento para reconocer los factores que, bajo el rango de valores establecidos en la (Tabla 1), resultan significativos estadísticamente. El esquema experimental que se aplicó con el propósito de alcanzar esta meta, es el conocido como factorial fraccionado en dos niveles, en este caso fue el  $2^{6-1}$ . Este diseño consta de 32 tratamientos y se le agregaron 5 en el nivel intermedio de cada factor para evaluar la falta de ajuste del modelo. Se llevó a cabo el experimento aleatorizando los 37 tratamientos.

Factores y Niveles	Nivel uno	Nivel dos
A: Humedad (%)	10 %	22 %
B: Temperatura (°C)	Ambiente	100 (°C)
C: Presión (MPa)	30 MPa	90 MPa
D: Tamaño de partícula (pulg)	1/8 Pulg:pg	$\frac{3}{4}$ Pulg:pg
E: Aglutinante (%)	0 %	10 %
F: Tiempo de relajación (seg)	0 seg	20 seg

Tabla 1. Niveles de los factores utilizados.

Las muestras que sirvieron como la unidad experimental se prepararon mediante una dieta para alimentación de bovinos; la cual se formuló a base de residuos de maíz, combinada con un porcentaje bajo de pollinaza y melaza. El experimento se realizó simulando las etapas

del proceso de densificación.

Tres de los factores se controlaron durante la preparación de las muestras: tamaño de partículas, contenido de humedad y contenido de aglutinante. El material proveniente del campo se picó en un molino de martillos, utilizando mallas de  $1/8$  pg,  $\frac{1}{2}$  pg y  $\frac{3}{4}$  pg, con los cuales se obtuvieron los niveles del tamaño de partículas. Para obtener los niveles de la humedad, se adoptó la norma ASAE S358.2 y se ocupó un horno de secado con una precisión  $\pm 2^{\circ}C$ . Los niveles de humedad (Cuadro 1) se alcanzaron secando o añadiendo agua atomizada según sea el contenido de humedad original. La melaza se utilizó como aglutinante, ésta se diluyó en agua y se atomizó, ya que su alta viscosidad hace imposible una aplicación homogénea en las muestras de residuos.

Los tres factores adicionales se controlaron al momento de la experimentación: presión, tiempo de relajación y la temperatura. Para el control de la presión y el tiempo de relajación se contó con una tarjeta de adquisición de datos con tres señales, dos señales provenientes de dos sensores de presión: uno para la presión total del sistema y otra para registrar la presión efectiva, y un sensor de desplazamiento. El tiempo de relajación es un periodo en el cual, una vez alcanzada la presión máxima, el émbolo del cilindro activado se detiene a deformación constante. La temperatura se registró por medio de un termopar tipo K, colocado en un orificio hecho en el dado de compresión y un indicador de temperatura.

A continuación se describe la forma que se empleó para medir las tres variables de respuesta. La densidad se calculó dividiendo la masa entre el volumen del cubo. La masa se obtuvo en una balanza granataria de una precisión de 1g, luego se midieron las dimensiones del producto densificado por medio de un calibrador de carátula, Área por la longitud, para estipular su volumen. La durabilidad se determinó en función de la norma ASAE S269.3. Las pruebas de durabilidad requieren 10 cubos, por lo que en cada tratamiento se emplearon 9 cubos de madera, similares en masas y dimensiones. Los cubos se pesan antes y después de la prueba, el material retenido entre el material inicial es el porcentaje de durabilidad. Por último, el consumo específico de energía se calculó por medio de los registros de sensores; dos de presión y uno de desplazamiento. El producto de la fuerza por el desplazamiento es la energía gastada, ésta dividida entre la masa de la muestra proporciona el consumo específico de energía.

Considerando la información del modelo generado con los factores que resultaron sig-

nificativos del experimento, y además de contrastar la viabilidad de las respuestas con el conocimiento del proceso: se decidió detallar y ampliar la región experimental con el fin de establecer condiciones óptimas de operación en el proceso de densificación. En la Tabla 2 se muestran los factores y sus nuevos valores reales, en el primer renglón se describen los valores codificados.

Factores y Niveles	-2	-1	0	1	2
A: Humedad (%)	8	11	14	17	20
B: Temperatura (°C)	20	45	65	85	107
C: Presión (MPa)	30	45	60	75	90
D: Tamaño de partícula (pulg)	1/8	5/16	1/2	11/16	7/8

Tabla 2. Niveles reales y codificados de los factores en el segundo experimento.

Como el objetivo es optimizar el proceso: un modelo de segundo orden es apropiado para tal fin, en ese sentido es necesario utilizar diseños que permitan ajustar ese tipo de modelos. Así en la segunda etapa experimental, se seleccionó el diseño central compuesto (dcc) (Box y Draper, 1987). En este caso se tienen los valores codificados  $-2$  y  $2$  y el número de repeticiones en el nivel intermedio de los factores,  $n_c = 7$

### 3 El procedimiento de análisis

El procedimiento que se siguió para analizar los datos generados al efectuar el experimento, consistió:

1. En hacer un análisis de correlación entre las respuestas. En la circunstancia que resulte significativa la correlación entre alguna de las variables, sólo se considera una de ellas, posteriormente, ésta se usará para predecir la que no se contempló.
2. A continuación, las respuestas no correlacionadas se ajustan mediante el método de mínimos cuadrados. Por medio del análisis de residuales se verificó si existen factores que afecten la variabilidad. Dado que esto ocurrió así, se reajustaron los modelos usando mínimos cuadrados ponderados. Este procedimiento expresa en forma de algoritmo como sigue:
  - (a) Estimar el modelo para cada respuesta por mínimos cuadrados:  $\hat{Y}_u$ ; donde  $u$  representa alguna de las tres respuestas.

- (b) Modelar  $W_u = \log(\text{abs}(Y_u - \hat{Y}_u)) = X\alpha + \varepsilon_u$  (donde  $\varepsilon_u$  es una variable aleatoria); en función de los factores. Este modelo indicará qué factores influyen en la variabilidad del proceso.
- (c) Se estiman los pesos  $P_i = (\exp(\hat{W}_{ui}))^2$ ,  $i = 1, \dots, n$  =tratamientos.
- (d) Se ajusta el modelo por mínimos cuadrados ponderados, con los  $P_i$  como pesos; éste se denota por  $\hat{Y}_u(P)$ .

Dado que existe más de una respuesta, se aplicó el método gráfico de multirrespuesta, el cual consiste en la superposición de curvas de nivel con el fin de obtener un óptimo común para las respuestas presentes en el proceso. El procedimiento es como sigue:

- Utilizar las curvas de nivel para encontrar un óptimo de cada  $\hat{Y}_u(P)$ .
- Se sobreponen las curvas de nivel de las respuestas para determinar el óptimo común.

## 4 Resultados y conclusiones

Los resultados del primer experimento mostraron que los factores E: porcentaje de aglutinante y F: tiempo de relajación, fueron significativos en la densidad, no así para el consumo de energía. Dado que esta última respuesta es la más importante, estos factores se fijaron en un nivel donde el consumo de energía disminuye, es decir: cercano al 0% en el aglutinante y 20s en el tiempo de relajación. Por otro lado la falta de ajuste es altamente significativa. Por ello y en función del conocimiento técnico del proceso, se detalló la región experimental tal como se describe en el Cuadro 2. El detalle de los datos experimentales obtenidos, el análisis estadístico y la modelación se puede consultar con alguno de los autores.

Aplicando el algoritmo descrito en el apartado anterior, se obtuvo el óptimo común en el punto  $A = -1$ ,  $B = 0$ ,  $C = -1$  y  $D = -2$ , y los valores respectivos para las tres respuestas son: (884.4, 94.84, 13.03). Para verificar este resultado se llevaron a cabo 8 pruebas experimentales extras en esos valores de  $A, B, C$  y  $D$ . Los resultados en promedio y su error estándar se escriben en la Tabla 3. En resumen, se puede concluir que el procedimiento propuesto permitió optimizar el proceso con una mínima variabilidad. Con estos valores se alcanzaron las condiciones adecuadas que se requerían para este proceso

	Densidad(Kg/m <sup>3</sup> )	Durabilidad(%)	C. de energía(J/g)
Promedio	930.63	97.3	13.36
Error estándar	6.73	0.13	0.17

Tabla 3. Promedio y error estándar de las pruebas confirmatorias

## Referencias

- Batacharya, S.C. (1989). State of the Art of Biomass Densification. *Division of Energy Technology. Energy Sources, N. Y.*, Eds. Taylor and Francis, **11**, No. 3, 161-186.
- Box, G.E.P. y N.R. Draper. (1987). *Empirical Model-Building and Response Surface*. John Wiley & Sons, New York.
- Brunh, H.D. (1957). Engineering Problems in Pelletized Feeds. *Agricultural Engineering of ASAE*. Julio, 522-525.
- Erickson S. y M. Prior. (1990). The briquetting of agricultural wastes for fuel. *Food and Agricultural Organization of the United Nations*.
- Macarthur, D. S. (1981). Pelleting Behaviour of Bagasse. *Proceedings of Australian Society of Sugar Cane Technologist*. 215-223.

# **Modelación Estocástica, Específicamente como Cadena de Markov, Aplicada a Dos Etapas de la Producción del Hongo Seta: *Pleurotus Ostreatus***

**José de Jesús Lara Tejeda**

**Nahara Ayala Sánchez**

*Universidad Autónoma de Baja California, Facultad de Ciencias*

## **1 Introducción**

La necesidad de triplicar la producción alimentaria mundial (Ramos, 1987) y la declaración de los biólogos expertos en el tema, ha presentado a los hongos comestibles como una alternativa para mejorar las necesidades nutricionales de la población humana que habita en los países subdesarrollados (Jiménez, 1996). Lo anterior se aúna a otras alternativas como el cultivo masivo de peces e invertebrados marinos, así como de algas, insectos comestibles y la creación de razas híbridas de arroz, trigo, maíz y otros cereales.

La alternativa del hongo es basada en el bajo costo de producción, con grandes cantidades de hongo fresco en lapsos cortos y con alto contenido de proteínas (Martínez et al., 1984); el contenido de proteínas es el doble de lo que presentan los espárragos y la col, cuatro veces más de lo que contienen las naranjas y veinte con respecto a las manzanas. El contenido proteínico, así como vitaminas y minerales, hace que su valor nutritivo sea equiparable al del huevo, además de que contiene poca grasa y buen sabor.

En particular *Pleurotus Ostreatus*, se cultiva sobre residuos vegetales tales como papa de arroz, hojas de plátano disecado, lirio acuático, desechos de algodón y de caña de azúcar, de la estructura fibrosa que recubre el coco, fragmento de papel, olores de maíz, etc. La posibilidad de cultivar esta especie sobre desechos agrícolas abre las posibilidades de obtener a corto plazo y en poco espacio, cantidades significativas de proteínas para el hombre y un buen forraje, no convencional y de bajo costo, así como un fertilizante para el suelo (Guzmán, 1986; Gastón, 1992).

Ciertas condiciones de extensión territorial, diversidad de climas y tipos de vegetación que presenta México, da cabida a considerar la alternativa del Pleurotus Ostreatus. De lo anterior su industrialización juega un papel importante (Lizárraga, 1993).

## 2 Modelo de líneas de espera para las etapas: micelio secundario y fructificación

El modelo de Líneas de Espera concibe la existencia de un servicio que se le proporciona a un elemento del sistema real al que podemos llamar “Unidad Receptora de Servicio”; el servicio lo proporciona otro elemento que se llama Servidor. Se considera que las unidades receptoras de servicio originalmente se encuentran dentro de un primer conjunto que se denomina Población y, como es factible que la Unidad Receptora de Servicio tenga que esperar (puede que no) para ser atendida, entonces el modelo plantea estar formado por tres conjuntos: la Población, la Línea de Espera y el Canal de Servicio (Prawda, 1980).

El resultado de aplicar este modelo a algún sistema real tiene como resultado, para las unidades de receptoras de servicio, la estimación probabilística de tiempos de espera, tiempos estancia, cantidad de elementos en la espera y en la estancia del sistema de Líneas de Espera que podrían darse durante la operación del sistema real; esto como resultado de trabajar como una Cadena de Markov (Saaty, 1961). Estos resultados son atractivos para el caso del crecimiento de hongos para efectos de diseñar una producción.

De las etapas de la producción, hasta obtener los carpóforos, se consideraron la del Crecimiento Micelial Secundario y el de la Inducción a la Fructificación. En ambos casos, por involucrarse el crecimiento de una especie de hongo (*Pleurotus Ostreatus*), se concibe el componente de la espera, formado por los individuos de la especie en cuestión y por realizarse el crecimiento de ella misma, el servicio se produce por el mismo hongo que es también la unidad receptora de servicio (el servicio es el crecimiento), es decir, se modela como un autoservicio.

## 3 Micelio secundario

En la etapa del crecimiento del micelio secundario, se inocula 1 mm<sup>2</sup> cuyo crecimiento se concibe bajo el esquema de la Fig. 1(a), a una tasa de crecimiento. En esta etapa no se

considera tasa de salida ya que cuando el micelio termina con los recursos nutricionales que le son suministrados, se le pasa a la siguiente etapa para obtener los carpóforos.

En la Fig. 1(b), se muestra el proceso estocástico correspondiente exhibiendo que no habrá tasa de salida. Segundo el enésimo estado en que se encuentre el micelio, se va a tener el crecer o no crecer, es decir, la probabilidad de crecer es  $\lambda\Delta t$  y  $1 - \lambda\Delta t$  la de no crecer. Del tiempo  $t$  al  $t + \Delta t$ :

Localización en el nodo	con probabilidad	crecer	no crecer
$n - 1$	$P_{n-1}(t)$	$\lambda\Delta t$	$P_{n-1}(t) \times \lambda\Delta t$
$n$	$P_n(t)$	$1 - \lambda\Delta t$	$P_n(t) \times (1 - \lambda\Delta t)$

Estableciendo el sistema de ecuaciones diferenciales correspondiente.

$$P_{n-1}(t) \times \lambda\Delta t + P_n(t) \times (1 - \lambda\Delta t) = P_n(t + \Delta t)$$

$$\frac{dP_n(t)}{dt} = \lambda P_{n-1}(t) - \lambda P_n(t)$$

que resolviendo para  $n=0, n=1, \dots$  se obtiene la expresión general

$$P_i(t) = \frac{(\lambda t)^i e^{-\lambda t}}{i!} \quad (1)$$

para  $i$  unidades, del tamaño del inóculo, que se han producido.

## 4 Fructificación

En la etapa de fructificación para la producción de carpóforos en cierto momento se provoca una cosecha lo cual involucra una tasa  $\mu$  de salida, figura 2. Con el proceso estocástico para esta etapa es factible llegar a:

$$P_i = P_0 \frac{\lambda_0 \lambda_1 \dots \lambda_{i-1}}{\mu_i \mu_2 \dots \mu_i} \quad (2)$$

Con  $\lambda_i = \lambda(m - i)$  y  $\mu_i = i\mu$ ,  $\mu$  y  $\lambda$  promedio, tenemos entonces:

$$P_i = P_0 C_i^m \left(\frac{\lambda}{\mu}\right)^i \quad (3)$$

que sería la probabilidad de tener  $i$  unidades de carpóforos producidos. Como

$$\sum P_i = 1 \rightarrow \sum P_0 C_i^m \left(\frac{\lambda}{\mu}\right)^i = 1,$$

de aquí

$$P_0 = \frac{1}{\sum C_i^m \left(\frac{\lambda}{\mu}\right)^i}.$$

Sea  $W$  la cantidad promedio de carpóforos que se encuentran en producción hasta antes de la cosecha,

$$W = \sum i P_i = \sum i (P_0 C_i^m \left(\frac{\lambda}{\mu}\right)^i) \quad (4)$$

y definiendo una tasa efectiva como  $\lambda_{efe} = \sum \lambda_i P_i$ . La estimación del tiempo en que los carpóforos permanecen en el proceso y son cosechados es

$$Tw = \frac{W}{\lambda_{efe}} \quad (5)$$

tal que el tiempo en que el micelio espera con crecimiento de carpóforos para ser luego cosechados se estimaría como  $T_l = Tw - T_s$  donde  $T_s = \frac{1}{\mu}$ .

## 5 Resultados de micelio secundario

De datos experimentales:

	día	$mm^2$	$\lambda$
dato I	1	1 (inoculado)	
dato II	2	6.25 (crecido)	5.25
dato III	7	21.16 (crecido)	2.982
dato IV	10	30.25 (crecido)	3.0

$$\lambda_{efe} = \sum \lambda_I P_i + \sum \lambda_{II} P_{ii} + \sum \lambda_{III} P_{iii}.$$

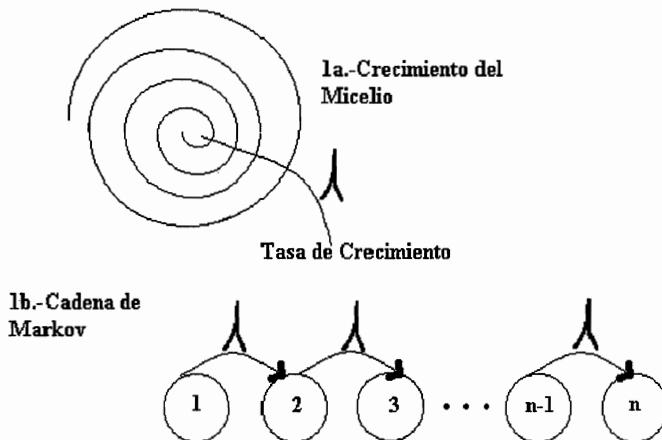
Utilizando (1), de donde se obtiene  $\lambda_{efe} = 2.982$ ,  $W = 15.038$ ,  $Tw = 5.04$  días. Con apoyo en la distribución de probabilidad  $P_i$ ... (1), con un intervalo de confianza del 95% el crecimiento estaría entre 7.7 y 20.429 unidades de inóculo en un lapso de 2.6 a 6.9 días.

## 6 Etapa de fructificación

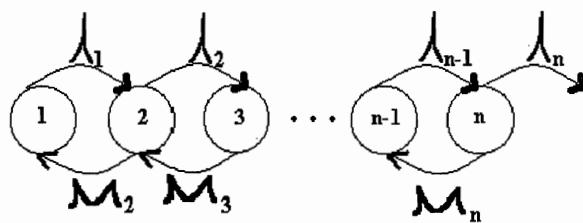
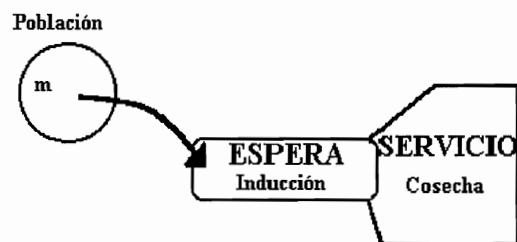
Con una tasa de llegada de micelios para fructificación de  $\lambda = 2.84$  y una tasa de cosecha de  $\mu = 0.798$ . La cantidad de carpóforos que permanecen en la etapa correspondiente hasta antes de la cosecha es  $W = 48.389$  con (4) y considerando la distribución de probabilidad para esta etapa (3), con un intervalo de confianza de 95% se estimaría  $41.21 \leq W \leq 53.97$ .

El tiempo de estancia hasta antes de la cosecha en esta etapa, sería  $T_w = 17.065$  con (5), que utilizando la distribución de probabilidad con un intervalo de confianza del 95%,  $14.53 \leq T_w \leq 19.03$ .

**Fig. 1 Micelio Secundario**



**Fig. 2 Fructificación del Carpóforo.**



## Referencias

- Gastón, G., Mata, G., Salmones, D., Conrado, S., y Guzmán, D. (1992). *El cultivo de hongos comestibles con especial atención a especies tropicales y residuos agroindustriales*. IPN, SEP.
- Guzmán, G.; y Martínez, D. (1986). Planta productora de hongos comestibles sobre pulpa de café. *Revista Ciencia y Desarrollo* **11 (65)**, 41-48.
- Jiménez, A.(1996). Proyecto producción de hongos comestibles (*Pleurotus spp.*) del grupo mujeres de Coajumulco, Morelos.
- Lizárraga, M. (1993). *Eficiencia Biológica, Biométrica y Análisis Nutricional de Pleurotus Ostreatus, (Jacquin ex Fr.) Kummer, (cepa sin esporas), sobre desechos agrícolas de Baja California*. Tesis profesional. Universidad Autonoma de Baja California, Ensenada.
- Martinez, D., M. Quirarte, C. Soto, D. Salmones y G. Guzmán (1984). Perspectivas sobre el cultivo de hongos comestibles en residuos agroindustriales en México. *Bol. Soc. Mex. Mic.* **19**, 207-219.
- Prawda, J. (1980). *Métodos y modelos de investigación de operaciones*. Limusa, México.
- Ramos, E.J. (1987). *Los insectos como fuente de proteínas en el futuro*. LIMUSA. 2a Ed. México.
- Saaty, L.T. (1961). *Elements of queueing theory*. Dover publications Inc.

# Cómputo de Probabilidades de Error en el Conteo Rápido

Federico O'Reilly y Raúl Rueda <sup>1</sup>

*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,  
Universidad Nacional Autónoma de México*

## 1 Introducción

El país está dividido, para efectos electorales federales, en 300 distritos electorales. Cada distrito está constituido por secciones electorales y éstas por casillas. El objetivo en un conteo rápido es pronosticar los resultados de una elección, federal en este caso, con base en una muestra de secciones o casillas.

El esquema de muestreo que se utilizó para el conteo rápido que el IFE realizó, fue el siguiente: se tomaron 3 muestras de 850 secciones cada una y se encomendaron a tres empresas. Dichas secciones se seleccionaron con muestreo aleatorio simple para cada distrito de manera proporcional al número de secciones totales en cada uno de ellos.

Al cierre de las casillas, la información recopilada consistió en las votaciones de cada uno de los partidos/coaliciones que participaron, además de los votos emitidos a candidatos no registrados y los votos anulados. Por otra parte, se tenía el tamaño de la lista nominal de cada sección en muestra.

Así, para cada sección en muestra, al finalizar la jornada, se tendría como información a  $\{x_{hk1}, \dots, x_{hkr}, v_{hk}\}$ , en donde  $x_{hkl}$  es el número de votos emitidos para cada uno de los  $r$  partidos (incluyendo como “partidos” a los candidatos no registrados y los “anulados”) y  $v_{hk}$  es al total de votos emitidos en la  $k$ -ésima sección del distrito  $h$ .

Si se denota con  $n_{hk}$  al tamaño de la lista nominal de la sección  $k$  del distrito  $h$ , entonces la proporción de votos emitidos a favor del partido  $l$  con respecto a la lista nominal de la sección es

$$\theta_{hkl} = \frac{x_{hkl}}{n_{hk}}.$$

---

<sup>1</sup>Con agradecimiento a Karim Anaya

Sea  $M_h$  el total de las secciones que conforman al distrito  $h$  y  $m_h$  el número de secciones de este total, que aparecen en la muestra. La proporción  $\theta_{hkl}$  es un dato a nivel sección, por lo que si hubiéramos hecho un censo, la proporción

$$\theta_{h.l} = \sum_{k=1}^{M_h} \left( \frac{n_{hk}}{n_{h.}} \right) \theta_{hkl}$$

sería la proporción de votos para el partido  $l$  en el distrito  $h$  relativo a la lista nominal del distrito, en donde el punto significa que se ha sumado sobre el índice correspondiente. De la misma manera,

$$\theta_{..l} = \sum_{h=1}^{300} \left( \frac{n_{h.}}{n_{..}} \right) \theta_{h.l} = \frac{1}{n_{..}} \sum_{h=1}^{300} \sum_{k=1}^{M_h} x_{hkl}$$

sería la proporción nacional respecto a la lista nominal total.

Pero no se tienen todas las secciones, sino sólo una muestra  $m_h$  de ellas en cada distrito, por lo que deben encontrarse estimadores de estas proporciones.

## 2 Proceso de estimación

El estimador que se propone es el que se utiliza en un muestreo por conglomerados cuando el tamaño de ellos es conocido, como en este caso. Específicamente, en cada estrato se tiene que

$$\hat{\theta}_{h.l} = \frac{\sum_{k=1}^{m_h} x_{hkl}}{n_{h.} \cdot \frac{m_h}{M_h}},$$

y de manera natural se tendrá que el estimador a nivel nacional resulta en

$$\hat{\theta}_{..l} = \sum_{h=1}^{300} \left( \frac{n_{h.}}{n_{..}} \right) \hat{\theta}_{h.l},$$

cuya varianza se estima con

$$\hat{V}(\hat{\theta}_{..l}) = \frac{1}{n_{..l}^2} \sum_{h=1}^{300} \frac{M_h^2}{m_h} (1 - f_h) \left\{ \frac{\sum_{k=1}^{m_h} (x_{hkl} - \frac{n_{h.}}{M_h} \hat{\theta}_{h.l})^2}{m_h - 1} \right\},$$

en donde  $f_h = m_h/M_h$  es la fracción de muestreo.

Esto se hace para cada uno de los  $r$  partidos e inclusive para los votos emitidos. Obsérvese que cada estimador es la suma de 300 variables aleatorias, por lo que se tendrá muy buena aproximación normal para el vector  $(\hat{\theta}_{..1}, \dots, \hat{\theta}_{..r}, \hat{\theta}_{..e})$ , en donde  $\hat{\theta}_{..e}$  se refiere a la proporción de votos emitidos. La estimación de las covarianzas se hace generalizando la expresión de la estimación de la varianza utilizando los productos cruzados.

Los reportes usuales sobre resultados de las elecciones, se limitan a exhibir los valores de los  $\hat{\theta}_{..m}, \hat{\theta}_{..m'}, \hat{\theta}_{..m''}$  correspondientes a los tres partidos que quedaron a la cabeza, incluyendo la estimación de las correspondientes varianzas, (o sus raíces, las desviaciones) y/o se exhiben los intervalos de confianza (usualmente al 95 %) para los correspondientes valores poblacionales  $\theta_{..m}, \theta_{..m'} y \theta_{..m''}$ .

Pero la normatividad acordada por el IFE para los reportes de las votaciones, pide que las proporciones de votos a favor de un partido, se hagan respecto al total de votos emitidos. En la notación utilizada aquí, el reporte debería hacerse sobre las estimaciones para los  $\theta_{..l}/\theta_{..ve}$ ; con  $\theta_{..e}$  la contraparte poblacional de  $\hat{\theta}_{..e}$ . Como veremos a continuación, se puede replantear el verdadero problema de inferencia y desde luego atender también este detalle de normatividad utilizando un análisis sencillo que utiliza la aproximación normal.

¿Qué es realmente lo que se quiere inferir? Opinamos que es el poder “asegurar” con un alto grado de confianza, si al haber observado

$$\hat{\theta}_{..m} > \hat{\theta}_{..m'},$$

en efecto el partido  $m$  puede ser declarado vencedor, *i.e.*, si  $\hat{D} = \hat{\theta}_{..m} - \hat{\theta}_{..m'} < 0$ , que es una observación de la diferencia verdadera  $D$ , permite asegurar que  $D < 0$ .

### 3 Estimación de D

Se approximó con una normal a la distribución conjunta de los estimadores  $(\hat{\theta}_{..m}, \hat{\theta}_{..m'}, \hat{\theta}_{..e})$ . La matriz de varianzas covarianzas fue estimada como ya se indicó anteriormente y es importante observar que el vector de medias tiene precisamente a los ingredientes para inferir sobre  $D$ . De hecho  $\hat{D}$  es el estimador máximo verosímil para  $D$  y su varianza se estima a partir de la matriz de varianza-covarianza referida.

Con lo anterior puede inferirse “fiducialmente” sobre  $D$ , simplemente utilizando como su distribución, una normal centrada en  $\hat{D}$  y con la varianza estimada de  $\hat{D}$ . La exhibición de esta distribución permite evaluar directamente la probabilidad fiducial del “evento”  $D < 0$ .

Sin embargo, dado el acuerdo para reportar todas las proporciones relativas al número de votos emitidos y considerando que el número de votos emitidos, en sí, no es un parámetro central en este análisis, se convino en calcular la distribución condicional de  $\hat{D}$  dado  $\hat{\theta}_{..e}$  y con ella evaluar la correspondiente probabilidad fiducial del “evento”  $D < 0$ . Se observa que en esta distribución fiducial la varianza es necesariamente menor que con el análisis incondicional y el sesgo que incorpora en la esperanza condicional, que se puede acotar, resulta despreciable.

## 4 La muestra no está completa

En esta sección abordamos un problema que en nuestra opinión, es generalmente ignorado por otros análisis. El problema a que nos referimos es el que a la hora de tener que hacer la inferencia, no toda la muestra ha llegado.

Existe una opinión bastante generalizada en el sentido de que la muestra que ya se tiene, es en sí una muestra pero más chica en número. En una situación tan delicada como lo son las elecciones presidenciales, el supuesto de que “una parte de la muestra, es muestra” es un supuesto arriesgado.

Antes de haber aplicado el procedimiento de inferencia descrita en las secciones 1 y 2, se procedió a llevar a cabo un análisis en el que se hicieron supuestos sobre la parte de la muestra que aún no se tenía.

El ejercicio de inferencia contó la noche del 2 de julio con el 60% aproximadamente, de la muestra total de 2,550 secciones electorales.

Con ese 60%, tomado como una muestra completa pero de menor tamaño, se obtuvieron los estimadores referidos en la sección 2 y se propusieron en un programa interactivo (en línea) distintas votaciones para el restante 40%. Estas votaciones supuestas intentaban medir, al combinarse con los estimadores, qué tanto podrían variar las estimaciones ya hechas.

La inclusión de esta variabilidad adicional por lo no observado y bajo varios supuestos, llevaron a incorporar esta incertidumbre adicional, en forma de un incremento en la varianza estimada.

Este procedimiento combina lo observado con escenarios tan extremos como uno quiera plantear y produce una apreciación más conservadora. Si el porcentaje observado de la muestra es muy alto, los escenarios, por muy extremos que se elijan, tienen poco o ningún efecto, pero si el porcentaje es bajo, los escenarios pueden llevar a concluir que aún no se tiene suficiente evidencia para declarar un ganador.

A continuación se muestra una “salida” de la evaluación de la distribución fiducial para  $D$ , estandarizada (y condicionada) a  $\hat{\theta}_{..e}$ . Dicha salida resultó de haber incorporado hasta un 6% de ventaja del segundo partido a la cabeza respecto al primer partido a la cabeza en la parte de la muestra que aún no había arribado a las 10:30 p.m. del día 2 de julio del 2000.

La distribución fiducial de  $D$  condicional al total de votos emitidos, se approximó con una normal centrada en 0.0637 con varianza igual a 0.0003, cuya gráfica se muestra en la siguiente figura.

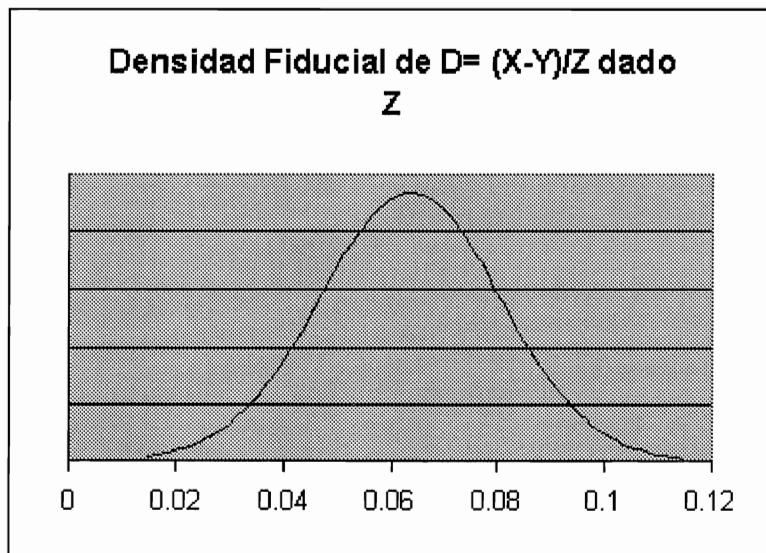


Figura 1

En la figura anterior, la escala en la que se grafica es respecto a los votos emitidos, es decir, la gráfica es de la densidad fiducial de  $D/\hat{\theta}_{..e}$  condicional a  $\hat{\theta}_{..e}$ , debido a la normatividad acordada por el IFE.

A partir de esta distribución, es fácil calcular la probabilidad fiducial de que  $D < 0$  resultando en  $9.49 \times 10^{-5}$ , concluyendo que los resultados apuntaban a una victoria del primer partido, aún en un escenario adverso.



# **Uso de Componentes Principales y Correlación Canónica para Dasometría de Cirián**

**Emilio Padrón Corral**

*Universidad Autónoma de Coahuila*

**Ignacio Méndez Ramírez**

*Universidad Nacional Autónoma de México*

**Nidia Hernández Pérez**

*Universidad Autónoma de Coahuila*

**Emilio Olivares Sáenz**

*Universidad Autónoma de Nuevo León*

## **1 Introducción**

En este trabajo se utilizaron los datos referentes a mediciones de dasometría del árbol y del fruto, en el cultivo del Cirián (*Crescentia alata* H.B.K.) usado como recurso forrajero en la localidad “El Llano” municipio de Coahuayana, Michoacan, México. Avila(1999). En un muestreo probabilístico en tres etapas, se aplicó la técnica de los componentes principales y de correlación canónica para el análisis de los datos. El método usado en la muestra se aplicó a una muestra mayor y se encontró que las tendencias de asociación entre las variables fueron muy similares; esto implica que la estimación en la muestra caracteriza adecuadamente a la población. Para el muestreo trietápico las unidades de la primera etapa fueron cuadrantes, las de la segunda etapa los árboles y los de la tercera fueron los frutos, aquí las unidades de muestreo se seleccionaron bajo un muestreo aleatorio irrestricto sin reemplazo en las diferentes etapas, todos con igual probabilidad (Sukhatme, 1971). Los datos se obtuvieron de 30 cuadrantes en una superficie de 120 ha.; cada 25 m. de distancia se monitoreó un cuadrante de 25x25 m. con orientación este-oeste, del cual se determinó la densidad de árboles y se extrajo una muestra del 15 porciento de árboles para obtener la dasometría del árbol y del fruto. Para el análisis se tomaron frutos por árbol, para el árbol se consideraron

las variables: altura, número de ramas, diámetro de ramas, cobertura, número de frutos y árboles por cuadrante; para el fruto se consideraron las variables: peso, diámetros ecuatorial y polar. El objetivo de este trabajo es explorar variables que influyan en el rendimiento y sus asociaciones.

## 2 Metodología

Se usaron las ecuaciones respectivas del muestreo para estimar la media y la varianza de los estimadores y el límite para el error de estimación, considerando que dichos estimadores en el estudio se distribuyen aproximadamente normal y con ellos obtener los intervalos de confianza respectivos, los cuales involucran las tres etapas. Para el análisis de correlación canónica la matriz de datos originales se puede particionar por la naturaleza de sus variables en dos submatrices, cada una correspondiente a dos grupos. Por lo tanto describir y descubrir la relación entre estos dos conjuntos de variables es uno de los objetivos que se presentan en este trabajo. Los conjuntos de variables originales  $X$  y  $Y$  serán descritos por dos conjuntos de  $k$  nuevas variables  $U$  y  $V$  obtenidas como combinaciones lineales de las  $X$  y las  $Y$  respectivamente tal que para cada par de nuevas variables  $U_i$  y  $V_i$ , el coeficiente de correlación  $r_i(U_i, V_i)$  sea máximo bajo la restricción de varianzas uno (Mardia et al, 1979), esto es:

$$\text{Var}(U_i) = \text{Var}(V_i) = 1 \quad \text{para } i = 1, 2, 3, \dots, k,$$

$$r_1(U_1, V_1) \geq r_2(U_2, V_2) \geq \dots \geq r_k(U_k, V_k).$$

Al aplicar la técnica de componentes principales logramos representar con menos dimensiones un conjunto de  $n$  observaciones en  $p$  variables, esto a partir de la descomposición espectral de la matriz de covarianza muestral  $S$  o de la matriz de correlaciones muestrales  $R$  y con ellas se logra explicar la mayor variabilidad de los datos (Manly, 1986). Por lo tanto la proporción explicada por las primeras  $k$  componentes esta dada por:

$$\frac{(\sum_{i=1}^k \lambda_i)}{\sum_{j=1}^p \lambda_j}$$

donde la suma de las  $\lambda_i$  representa la varianza acumulada hasta la componente  $k$ , mientras que la suma de las  $\lambda_j$  representa la traza de  $S$  (matriz simétrica de varianzas y covarianzas de los  $n$  elementos) o la suma de las varianzas estimadas de las variables  $Y_1, Y_2, \dots, Y_p$  y si

este cociente es grande, es decir mayor que el 80%, se toman esas primeras  $k$  componentes para representar el fenómeno, por lo que es costumbre evaluar la proporción explicada por las primeras  $k$  componentes. A pesar de que el muestreo fue polietápico, no se espera que la correlación intraconglomerado cuadrantes sea grande, más bien que sea casi cero, esto hace que no se requieran introducir los llamados efectos de diseño para corregir los análisis. Esta consideración se ve muy fuertemente apoyada por el hecho de que al comparar estimadores de medias y varianzas con la muestra analizada, con los obtenidos con una muestra mucho mayor, no se encontraron diferencias importantes. Cabe aclarar que la población es una muestra de árboles (con subunidades) de los 30 cuadrantes y la muestra utilizada en este trabajo estuvo formada por las medias de las subunidades.

### 3 Resultados

Con respecto a las variables de respuesta analizadas para caracterizar el árbol y de acuerdo a las técnicas de muestreo utilizadas se obtuvieron los siguientes resultados (Tabla 1).

Variables	Estimador de la media poblacional	Varianza del estimador	Límite para el error de estimación	Intervalo de confianza
Altura del árbol m.	7.862	0.299	1.072	$6.79 \leq \mu \leq 8.93$
No. de ramas/árbol	3.965	0.160	0.786	$3.18 \leq \mu \leq 4.75$
Diámetro ramas/árbol	12.980	0.625	1.549	$11.43 \leq \mu \leq 14.53$
No. frutos/árbol	28.955	14.828	7.547	$21.41 \leq \mu \leq 36.50$
Cobertura/árbol	53.495	10.677	6.404	$47.09 \leq \mu \leq 59.90$
Peso de fruto/árbol	464.473	410.216	40.496	$423.97 \leq \mu \leq 504.96$

Tabla 1. Medias, Varianzas, Límites del Error de Estimación e Intervalos de Confianza para las medias al 95 porciento de Confianza.

Después de obtener los intervalos de confianza para las variables en estudio, se hizo un análisis por componentes principales con el fin de reducir el número de variables que expliquen mejor la producción (Tabla 2).

Aquí se puede observar que las primeras tres componentes explican el 76.986% de la varianza total. También se realizó un análisis de correlación canónica para determinar la significancia entre las variables del árbol y las del fruto (Tabla 3).

Componente	Valor Propio	% Total de Varianza	Valor Propio Acumulado	% Acumulado
1	3.043	33.819	3.043	33.819
2	2.646	29.409	5.690	63.228
3	1.238	13.757	6.928	76.986
4	0.799	8.886	7.728	85.873
5	0.509	5.661	8.238	91.534
6	0.295	3.282	8.533	94.816
7	0.232	2.578	8.765	97.395
8	0.180	2.000	8.945	99.395
9	0.054	0.604	9.000	100.000

Tabla 2. Explicación de la varianza por los componentes

Raíz Removida	R	$R^2$	$X^2$	g.l.	$X^2$ de Tablas
0	0.7095	0.5035	31.9222	18	28.8693
1	0.5418	0.2935	15.1177	10	18.3070
2	0.4960	0.2460	6.7775	4	9.4877

Tabla 3. Análisis por Correlación Canónica

En este cuadro se encontró relación entre las variables del árbol y las del fruto en la primera correlación, no así en las dos restantes donde no hubo significancia. También se presentan las ponderaciones entre las variables canónicas (Tabla 4).

Variables X	$U_1$	$U_2$	$U_3$
Árboles/cuadrante	0.3548	-0.7698	-0.0421
Altura del árbol	-0.3454	0.1355	0.5661
Número de ramas	-0.3790	-0.2611	1.0131
Diámetro de ramas	-0.6935	0.1200	0.4349
Cobertura	1.7586	0.3263	0.0930
Número de frutos	-1.1959	-0.5133	-0.4239
Variables Y	$V_1$	$V_2$	$V_3$
Peso de fruto	-1.6741	-0.5837	0.2440
Diámetro ecuatorial	0.7090	0.7729	1.1339
Diámetro polar	1.1632	-0.9299	-0.5716

Tabla 4. Ponderaciones entre las Variables Canónicas

En la interpretación del primer par de variables canónicas ( $U_1, V_1$ ), aparece un contraste de las variables cobertura y no. de frutos con respecto a las demás, esto representa un aumento en densidad y cobertura. Por otro lado, se tiene un coeficiente positivo para diámetro

polar y uno negativo para peso del fruto. Esto sugiere que el aumento en no. de frutos y cobertura está asociado con la disminución en el peso del fruto y un aumento en el diámetro polar del mismo.

## 4 Conclusiones

En la obtención de los intervalos de confianza estos coinciden en cierto porcentaje con las estimaciones que se presentan en la literatura agrícola, aunque algunas estimaciones no coinciden, esto probablemente se deba a que el trabajo se realizó en zona árida y la planta forrajera que se maneja es de región tropical húmeda. En lo correspondiente a los análisis por medio de correlación canónica y tomando de manera conjunta la interpretación de  $U_1$  y  $V_1$  se concluye que la disminución en el número de frutos y peso de los mismos está asociado con un aumento en la densidad en el número de ramas y en los diámetros polar y ecuatorial del fruto. Es decir las asociaciones se resumen en que a más densidad de árboles con menor porte se tienen frutos grandes pero ligeros. En lo que respecta al análisis mediante componentes principales se observó una asociación positiva entre las variables número de frutos y la cobertura, explicando ambos componentes un 63 % de la varianza total, cabe señalar que sólo se consideraron aquellas variables con una correlación mayor a 0.5. Estas conclusiones sugieren que para controlar mejor la producción de forraje procedente de este árbol debe tomarse en cuenta para su manejo la densidad del árbol por hectárea.

## Referencias

- Avila, R.N.A. (1999). *Ecología y Evaluación del Fruto del Cirián (Crescentia alata H.B.K.) como Recurso Forrajero en la Localidad “El Llano” Municipio de Coahuayana, Michoacán, México*: Saltillo, Coahuila. Tesis de Maestría UAAAN.
- Manly, B.F.J. (1986). *Multivariate Statistical Methods*. New York: Chapman and Hall., 59-125.
- Mardia, K.V., Kent, J.T. y Bibby, J.M. (1979). *Multivariate Analysis* Academic Press, New York, 227-243.
- Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling Theory of Surveys with Applications*. Second Edition, Iowa State University Press., Ames Iowa, 301-307.



# **Estimación de los Parámetros de la Curva de Lactancia de Vacas Holstein Friesian Sometidas a un Programa de Somatotropina Bovina Recombinante**

**G. M. A. Rivero**

*Laboratorios Schering-Plough, S.A. de C.V. Mexico.*

**G. M. E. Rosas**

**R. A. J. Avila**

*Depto de Genética y Bioestadística. Fac. de Med. Vet. Y Zoot., UNAM.*

## **1 Introducción**

La selección para producción en el ganado lechero, se ha basado principalmente en producción total de leche. La producción total puede ser debida a diferentes formas de la curva de lactancia. La curva de lactancia más adecuada, aún es cuestión de debate. Las vacas que producen moderadamente con una alta persistencia a lo largo de la lactación, usualmente estarán sometidas a menores condiciones de estrés; mientras que las vacas que son menos persistentes y tienen una gran diferencia entre el pico de producción y el final de la lactación tendrán mayores condiciones de estrés. También existe una cuestión económica acerca de los requerimientos nutricionales para una mayor persistencia de la vaca. Con el objetivo de alcanzar una eficiencia productiva, es necesario desarrollar nuevas alternativas para incrementar la eficiencia en la producción. La somatotropina bovina recombinante ha llegado a ser el producto biotecnológico de mayor impacto alrededor del mundo (Bauman, 1992). La producción de leche se incrementa del 10 al 20% con la aplicación de somatotropina bovina recombinante (Peel y Bauman, 1987). La caracterización de la forma de la curva de lactancia de vacas tratadas con somatotropina, se puede realizar a través de los coeficientes de una ecuación matemática.

Wood (1967, 1969) describió una ecuación para caracterizar la curva de lactación basada en la transformación logarítmica de la función gama incompleta no lineal. Esta ecuación

es la siguiente:  $Y = an^b \exp(-cn)$ , donde  $Y$  es el promedio de producción de leche en un periodo de tiempo  $n$ ;  $a, b$ , y son coeficientes. Trabajos posteriores apoyaron el uso de ecuaciones no lineales para describir la curva de lactancia (Cobby y Le Du, 1978; Kellog *et al.*, 1977; Schaeffer, *et al.*, 1977). El objetivo del presente estudio fue evaluar los efectos de la somatotropina bovina recombinante (rbST) sobre la producción de leche en vacas Holstein Friesian usando un modelo no lineal para caracterizar la forma de las curvas de lactación.

## 2 Material y métodos

El estudio se realizó en un hato de vacas Holstein Friesian en lactación, ubicado en el Norte de la República Mexicana. Los criterios de inclusión fueron vacas con al menos 60 días en lactación, con una condición corporal mínima de 3 y vacas clínicamente sanas (Frood y Croxton, 1978). Los animales fueron alimentados con raciones integrales para cubrir sus requerimientos nutricionales. El manejo general del hato se hizo con base en los métodos de rutina. Las vacas fueron asignadas a cada grupo de tratamiento mediante un diseño completamente aleatorizado. Cada vaca recibió dosis de 500 mg de somatotropina bovina recombinante adicionadas con 1655 UI de vitamina E (Boostin-S<sup>®</sup>, Lab. Schering Plough). La aplicación de la somatotropina se realizó por vía subcutánea en la fosa isqueorectal o en el área subescapular, cada 14 días. Las vacas se ordeñaron diariamente a las 04:00 hrs A.M. y a las 16:00 hrs P.M. La producción de leche de cada vaca lactante se midió cada semana durante quince semanas consecutivas.

## 3 Análisis estadístico

Se utilizó el siguiente modelo:

$$Y = an^b \exp(-cn), \quad (1)$$

donde  $\exp$  es la base del logaritmo natural  $\ln$ ,  $n$  = día de lactación y  $a, b, c$  son coeficientes a estimar. De estas constantes,  $a$  representa un factor de producción de leche al inicio de la lactación; mientras que  $b$  y  $-c$  representan la pendiente limitante de la curva antes y después del pico lactación, respectivamente.

Obteniendo el  $\ln$  a la ecuación (1), se linearizan los coeficientes que pueden ser estimados

mediante procedimientos de regresión lineal múltiple (Wood, 1969).

Al aplicar el  $\ln$ , se obtiene la siguiente ecuación matemática para cada vaca

$$\ln(y) = \ln(a) + b \ln(n) + cn \quad (2)$$

La ecuación (2) es una forma de un modelo de regresión lineal múltiple del tipo:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon, \quad (3)$$

Donde  $Y = \ln(y)$ ,  $b_0 = \ln(a)$ ,  $b_1 = b$ ,  $b_2 = c$ ,  $X_1 = \ln(n)$ ,  $X_2 = n$ , y  $\varepsilon$  = residuo aleatorio.

Al aplicar un modelo de regresión lineal múltiple, como la ecuación (3), se supone un término residual multiplicativo asociado con (1); además, se supone que los residuos en (3) están normal e independientemente distribuidos con media cero y varianza constante (Sahai y Ageel, 2000).

Se utilizó la ecuación (3) en cada registro de lactancia de la vaca para obtener estimadores de los tres coeficientes de regresión. Se realizó una análisis de cuadrados mínimos en cada uno de los tres coeficientes con el PROC GLM del Sistema de Análisis Estadístico SAS<sup>®</sup> (SAS, 1989).

Las constantes  $a$ ,  $b$  y  $c$  se calcularon para cada vaca junto con la persistencia ( $S$ ). La persistencia de lactación se calculó con la siguiente fórmula (Wood, 1968):

$$S = -(b + 1) \log c$$

Los valores de  $a$ ,  $b$  y  $c$  se estimaron independientemente del número y mes de parto, y los efectos de estacionalidad derivados de las desviaciones individuales se ponderaron con respecto a la media de la curva. Aunque este procedimiento reduce el sesgo, aún existe ligera confusión de la persistencia con la época de parto.

La producción máxima de leche se estimó como

$$Y_{\max} = a(b/c)b \exp(-b)$$

donde  $b/c$  es el día estimado en donde la producción máxima de leche se alcanza.

Se estimaron las estadísticas descriptivas y las curvas de lactación para las vacas Holstein Friesian tratadas con rbST se generaron utilizando valores para  $a$ ,  $b$  y  $c$  (Figura 1).

## 4 Resultados

Los coeficientes de regresión parcial para  $\log_e a$ ,  $b$  y  $c$  se presentan en el Tabla 1.

Parámetro	
$\log_e a$	2.1149
$A$	8.2892
$B$	0.3259
$C$	-0.0025
$\ln(c)$	5.9548

Tabla 1: Coeficientes para la curva de lactancia de vacas Holstein Friesian.

La curva de lactancia generada utilizando estos coeficientes se muestra en la Figura 1.

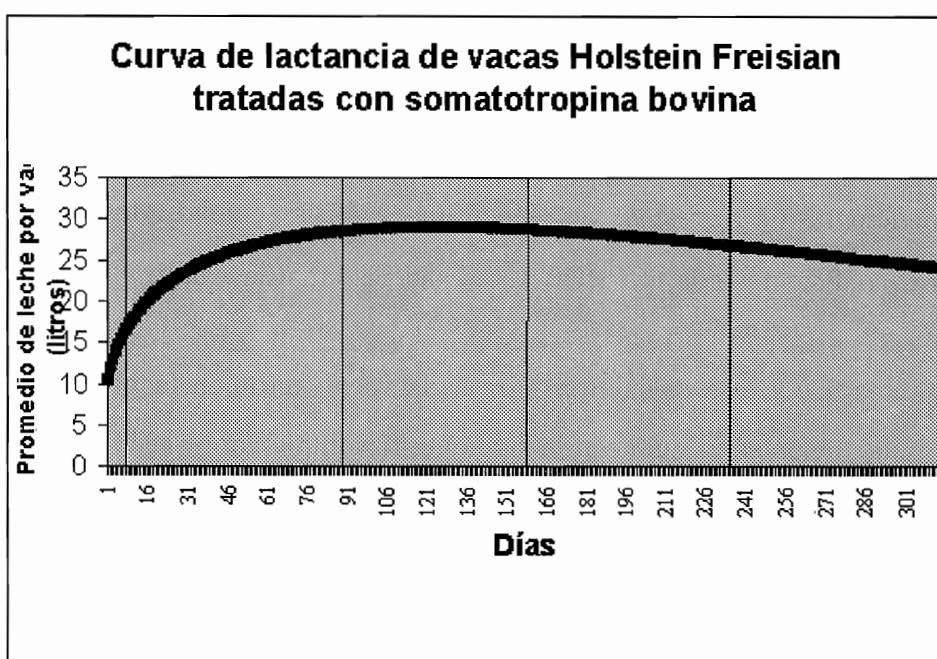


Figura 1. Curva de lactancia de vacas Holstein Friesian tratadas con somatotropina bovina recombinante

Los valores promedio para día al pico de producción, producción de leche diaria máxima ( $Y_{\max}$ ) y persistencia de la curva ( $S$ ) se presentan en el Tabla 2.

Día al pico de producción	125.681
Producción máxima diaria de leche	28.916
Persistencia de la curva	8.177

Tabla 2: Día al pico de producción, persistencia de la curva y producción de leche diaria máxima en vacas Holstein.

En el Tabla 3 se presentan algunas estadísticas descriptivas para la producción de leche.

Media	28.916
Desviación estándar	3.851
CV (%)	13.318
Skewness	-0.297 <sup>2</sup>
Kurtosis	0.803

Tabla 3: Estadísticas descriptivas para la producción de leche.

## 5 Discusión

Wood (1969) mostró que un modelo como  $Y = an^b \exp -cn$  puede dar un buen ajuste a la curva de lactancia, con al menos 73.8% , y en el mejor de los casos 91.2% con un promedio de 82.3% de ajuste a la variación en la producción semanal. La ecuación que se utilizó en el presente análisis tuvo un ajuste promedio de la variación del 82.9% .

Los estimadores de los coeficientes que caracterizan la forma de la curva de lactancia ( $b$  y  $c$ ) encontrados en este estudio se encuentran dentro de los rangos mencionados en otros estudios en ganado lechero (Wood, 1969; Wood, 1976; Wood, 1977; Cobby y Le Du, 1978; Frood y Croxton, 1978 y Rao y Sundaresan, 1979). Los resultados encontrados en este estudio son similares a los obtenidos por Ulloa (1988) en ganado Holstein Friesian en México.

Cobby y Le du (1978) obtuvieron estimadores para  $a$ ,  $b$  y  $c$  por estimación de cuadrados mínimos no lineales y por estimación de cuadrados mínimos para el logaritmo común. Ellos mencionan que la curva no describe con precisión la lactancia completa; sin embargo, mencionan que la estimación por cuadrados mínimos no lineal fue la más satisfactoria. Los resultados acerca de los coeficientes que caracterizan la forma de la curva de lactancia en vacas tratadas con somatotropina bovina recombinante coinciden con los obtenidos por Schaeffer *et al.* (1977). Los valores estimados para el día al pico de producción y persistencia son similares a los obtenidos por otros autores (Schaeffer *et al.*, 1977; Wood, 1968; Ulloa, 1988).

Al examinar la curva de lactancia de la Figura 1, se concluye que la alta producción de leche se manifestó en casi todas las etapas de la lactancia en respuesta al tratamiento con somatotropina bovina recombinante. Se sabe que algunas vacas son más persistentes o que tienen una tasa de declinación más lenta en la producción de leche que otras vacas. Con base

en lo anterior, si se puede estimar la persistencia para cada vaca con técnicas no lineales, entonces el tratamiento con somatotropina bovina recombinante podría ser evaluado para mantener la persistencia durante la lactancia completa.

## Referencias

- Bauman, D.E. (1992). Bovine somatotropin: review of an emerging animal technology. *J. Dairy Sci.*, **75**, 3432-3451.
- Cobby, J.M. and Le Du, Y.P.L. (1978). On fitting curves to lactation data. *Anim. Prod.*, **26**, 127-133.
- Frood, M.J. y Croxton, D. (1978). The use of condition-scoring in dairy cows and its relationship with milk yield and live weight. *Anim. Prod.*, **27**, 285-291.
- Kellog, D.W., Urquhart, N.S. y Ortega, A.J. (1977). Estimating Holstein curve of lactancies with a gamma curve *J. Dairy Sci.* **60**, 1308.
- Peel, C.J., Bauman, D.E.. (1987). Somatotropin and lactation. *J. Dairy Sci.*, **70**, 474-486.
- Rao, M.K. y Sundaresan, D. (1979). Influence of environment and heredity on the shape of curve of lactancies in Sahiwal cows. *J. Agric. Sci.*, **92**, 393-401.
- Sahai, H. y Ageel, M. (2000). *The Analysis of Variance Fixed, Random y Mixed Models*, Birkhäuser, New York.
- SAS. (1989). *SAS/STAT user's Guide (Version 6, 4<sup>th</sup> Ed.)* SAS Inst, Inc, Cary, North Carolina.
- Schaeffer, L.R., Minder, C.E., McMillan, I. and Burnside, E.B. (1977). Nonlinear techniques for predicting 305-day lactation production of Holsteins and Jerseys. *Can. J. Anim. Sci.*, **56**, 1636-1644.
- Ulloa, A.R. (1988). Utilización de registros parciales de producción de leche para estimar lactancias a 305 días en vacas Holstein de primer parto en México. *Universidad Autónoma de Chapingo*, México.
- Wood, P.D.P. (1967). Algebraic model of the curve of lactation in cattle *Nature*, **216**, 164-165.
- Wood, P.D.P. (1968). Factors affecting persistency of lactation in cattle. *Nature* **218**, 894.
- Wood, P.D.P. (1969). Factors affecting the shape of the curve of lactation in cattle *Anim. Prod.*, **11**, 307-316.

- Wood, P.D.P. (1976). Algebraic models of the curve of lactancies for milk, fat and protein production, with estimates of seasonal variation. *Anim. Prod.*, **22**, 35-40.
- Wood, P.D.P. (1977). The biometry of lactation *J. Agric. Sci.*, **88**, 333.



# Representación Gráfica de los Resultados de las Elecciones Presidenciales de 1994 y 2000 en México

**Patricia I. Romero Mares  
Guillermina Eslava Gómez  
Ignacio Méndez Ramírez**

*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,  
Universidad Nacional Autónoma de México*

## 1 Introducción

El graficar los resultados de preferencias políticas no es nuevo. Upton (1994), publicó gráficas para los cambios preferenciales en Inglaterra de 1987 a 1992. Se han presentado mapas de la República Mexicana con los resultados de las elecciones de 1994 (Calderón y Cazes, 1996), y recientemente por IFE(2000).

Los partidos políticos mayoritarios en México son PRI, PAN y PRD, considerados así en la gráfica correspondiente a 1994. Para las elecciones del 2000 se establecieron alianzas, los partidos considerados son PRI, AC (Alianza por el Cambio) representado por PAN y PVEM; y AM (Alianza por México) representado por PRD, PT, Convergencia por la Democracia, Sociedad Nacionalista y Alianza Social.

## 2 Método

Dado que los datos se componen de tres proporciones (una para cada uno de los partidos considerados), que se escalan para que sumen uno, se pueden trabajar como datos composicionales, y usar gráficas de coordenadas de área (Aitchison, 1986).

Se calculan los porcentajes de votos a cada partido, escalados de tal manera que sumen uno. Es decir,  $\% \text{ pan} + \% \text{ pri} + \% \text{ prd} = 1$ , y se calculan los puntos  $(x, y)$  como:

$$x = \frac{\% pan - \% pri}{\sqrt{3/2}}$$

(ver Gráfica 1).

Los vértices del triángulo representan los casos extremos en que uno de los partidos tiene 100% de los votos. El triángulo está dividido en tercios de áreas iguales según un partido tenga más de la tercera parte de los votos. Cada tercio de área se divide a su vez en dos. Cada una de estas seis partes representa al partido que quedó en primer lugar representado por el tercio mayor donde cayó el punto y el partido que quedó en segundo lugar, que es la dirección del sextante donde está el punto.

El IFE hace una clasificación de cada distrito en urbano y rural, considerando un distrito como urbano cuando la suma del número de secciones urbanas y mixtas representa un porcentaje mayor del 50% del total de secciones en el distrito, y considerándolo rural en caso contrario; esta clasificación se utilizó para las Gráficas 1 y 2.

Se hizo además una comparación por estado de los cambios que hubo en las preferencias electorales de 1994 a 2000. (Gráfica 4).

### 3 Resultados

En la Tabla 1 se presentan los resultados electorales para 1994 y 2000.

Los resultados por sextantes para 1994 y 2000 se dan en la Tabla 2.

1994		2000		Tabla 2
Orden de preferencias (sextante)	No. de distritos electorales	Orden de preferencias (sextante)	No. de distritos electorales	
PAN, PRI, PRD	17	AC, PRI, AM	156	
PAN, PRD, PRI	0	AC, AM, PRI	21	
PRI, PAN, PRD	196	PRI, AC, AM	75	
PRI, PRD, PAN	80	PRI, AM, AC	36	
PRD, PAN, PRI	0	AM, AC, PRI	4	
PRD, PRI, PAN	7	AM, PRI, AC	8	

### 4 Gráficas y comentarios

Gráfica 2 presenta los resultados de las elecciones de 1994 para los 300 distritos electorales, divididos en urbano, rural y tamaño relativo del distrito. El tamaño relativo del distrito se

definió como: pequeño si su % de lista nominal es menor o igual al primer cuartil, promedio, si su % de lista nominal está en el intervalo definido entre el primer y el tercer cuartiles; y grande si su % de lista nominal es mayor que el tercer cuartil. En 1994 el 92% de los distritos votaron por el PRI en primer lugar, el 5.7% por el PAN y solo el 2.3% por el PRD. Se nota que a excepción de un distrito rural que está en el tercio del PRD, todos los distritos rurales son priistas. De los urbanos, son unos cuantos los que votaron por PAN y PRD en primer lugar.

En el 2000 (Gráfica 4) se observa que además de que el 59% de los distritos son ahora panistas y solo 4% son perredistas, la gran mayoría de los distritos rurales permanecieron en el PRI. El PRI obtuvo el 37% de distritos y la mayoría de ellos son rurales. La gráfica más interesante, desde el punto de vista de comparación, es la Gráfica 3 que presenta la forma en que se movieron los porcentajes de votos de 1994 a 2000 por estado. Todos los estados se movieron de izquierda a derecha, a excepción del estado de Sinaloa que se movió de derecha a izquierda, es decir, su porcentaje de votos al PRI aumentó en 2000, en comparación al porcentaje que obtuvo en 1994. Una pendiente positiva de las rectas indica un aumento en el % de votos a la Alianza por México, y una pendiente negativa una disminución. El estado de Sinaloa fue el único estado con un aumento en el porcentaje de votos al PRI, una disminución del % votos a AM y también a AC. Se nota que el estado de Michoacán cambia el orden de sus votos de (PRI,PRD,PAN) a (PRD,PRI,PAN). El D.F. cambia de (PRI,PAN,PRD) a (PAN,PRD,PRI). El aumento más grande a % de votos para PRD se ubica en el estado 3 (Baja California Sur), aun cuando el orden de sus preferencias fue (PAN,PRI,PRD).

Ver Gráficas 2,3,4.

## 5 Comentarios finales

Los métodos gráficos, como los presentados en este trabajo, dan un panorama de la forma en que se distribuyen los tres porcentajes de voto, y ayudan a la toma de decisiones. Además, permiten comparar los porcentajes de votación para diversas regionalizaciones y tamaños de unidades simultáneamente. Se recomienda un mayor uso de estas técnicas gráficas para procesos electorales o algunos otros con particiones en tres categorías.

Tabla 1: Población<sup>1</sup>, lista nominal y resultados para las elecciones presidenciales de 1994 y 2000 en México.

Región Electoral	Población %	Lista Nominal %	Distritos Electorales	Total Secciones	Participación %	PAN % (PAN)	AC % (PAN)	PRI %	PRD %	AM PRD %
I Noroeste	21	20	62	62	14,341	81	65	35.8	50.1	49.5
II Norte	20	20	60	59	14,243	76	62	31.1	48.5	53.5
III Sureste	20	19	55	61	12,205	72	61	16.9	36.1	52.7
IV Centro	19	21	66	59	11,304	79	67	25.8	43.1	48.1
V Sur	20	20	57	59	11,327	77	64	22.1	38.4	47.3
Nacional	100	100	300	300	63,420	77	64	26.7	43.4	50.1

(1) Fuente: Conteo de Población y vivienda 1995. INEGI 1996. Resultados Preliminares del XII Censo General de Población, INEGI, México, 2000.

(2) Fuente: Instituto Federal Electoral, IFE.

PAN: Partido Acción Nacional

PRI: Partido Revolucionario Institucional

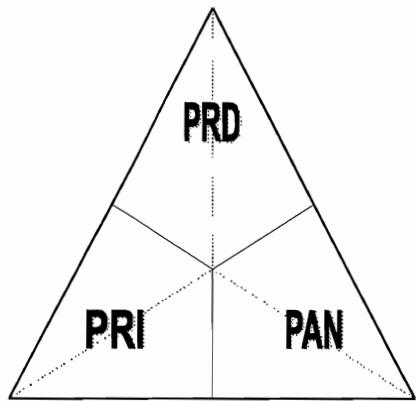
PRD: Partido de la Revolución Democrática

AC: Alianza por el Cambio (PAN, PVEM)

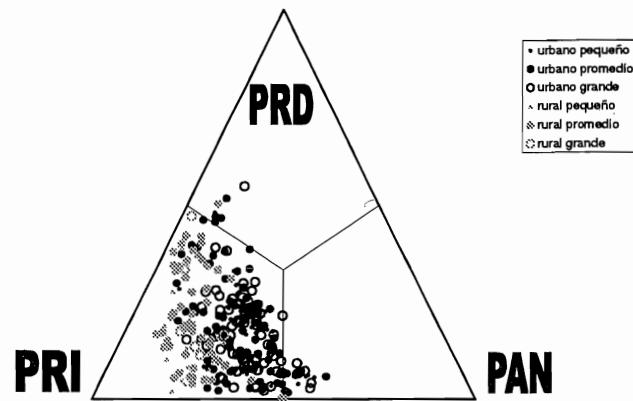
AM: Alianza por México (PRD, PT, Convergencia por la Democracia, Sociedad Nacionalista, Alianza Social).

\*Población total en 1994: 91,158,290

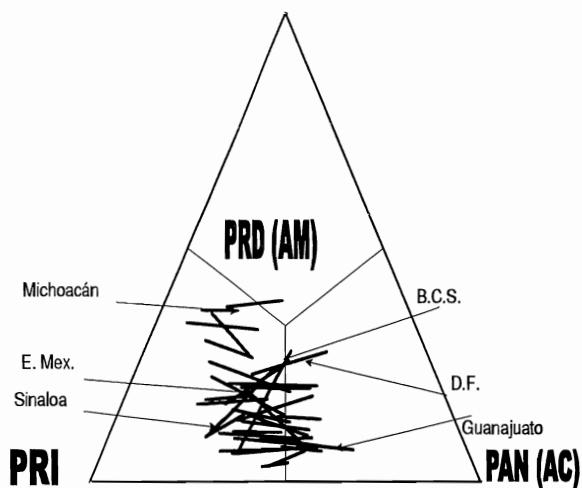
\*Población total en 2000: 97,361,711



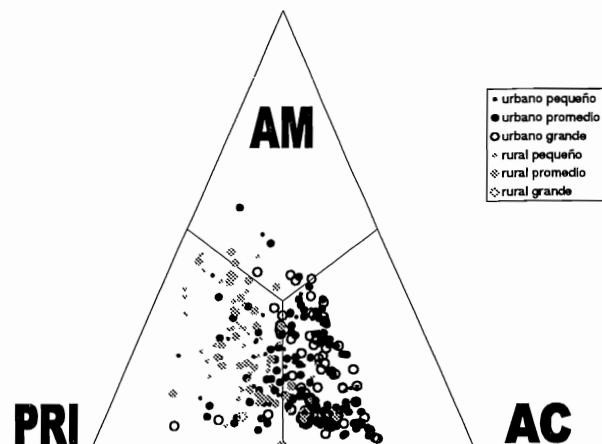
Gráfica 1



Gráfica 2: Elecciones 1994, Tipo de Distrito (300 Distritos)



Gráfica 3: Mov. de votos 1994-2000



Gráfica 4: Elecciones 2000, Tipo de Distrito (300 Distritos)

## Bibliografía

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data* London: Chapman and Hall.
- Calderón, A.E. y Cazes, D. (1996). *Las elecciones presidenciales de 1994*. México: UNAM, Centro de Investigaciones Interdisciplinarias en Humanidades.
- IFE (2000). Sistema de Consulta *Estadística de las Elecciones Federales de México 2000*. Versión 1.0, México, 2000.
- Upton, G.J.G. (1994). Picturing the 1992 British General Election. *J. R. Statist. Soc. A*, 231-252.

# Análisis de Datos de Mediciones Repetidas Utilizando Metodología de Modelos Mixtos

**G. M. E. Rosas**

**R. A. J. Avila**

*Dept. de Genética y Bioestadística. Fac. de Med. Vet. y Zoot., UNAM.*

**G. M. A. Rivero**

*Laboratorio Schering-Plough, S. A. de C. V.*

**B. R. Avila**

*Esc. de Med. Vet. y Zoot., Benemérita Universidad Autónoma de Puebla.*

## 1 Introducción

Los modelos lineales mixtos fueron desarrollados para evaluar el potencial genético de los sementales lecheros (Henderson, 1984); sin embargo, la aplicación de los modelos mixtos se ha extendido a todas las áreas de la investigación. Anteriormente, el análisis de los modelos mixtos se implementó adaptando métodos de efectos fijos a modelos con efectos aleatorios, lo que impone limitaciones en la aplicabilidad debido a que la estructura de la covarianza no se modela. La metodología de modelos mixtos para el análisis de datos de mediciones repetidas permite dirigir directamente la estructura de covarianza para obtener errores estándar validos y pruebas estadísticas eficientes. Las mediciones repetidas son una secuencia de múltiples respuestas, usualmente a través del tiempo, en la misma unidad experimental (Vgr.: un animal). En el caso de la producción de leche, la respuesta se da a través del tiempo (curva de lactancia). Los experimentos con mediciones repetidas son un tipo de experimento factorial, con el tratamiento y el tiempo como los dos factores. Los datos de mediciones repetidas usualmente se analizan con un análisis de varianza univariado como parcelas divididas con datos en el tiempo, tratando la unidad experimental como la parcela grande y la unidad experimental en un tiempo específico como la subparcela (parcela chica). Esta aproximación puede ser inválida debido a las fallas para cumplir los supuestos concernientes a las varianzas y correlaciones. Los objetivos del análisis de datos de mediciones

repetidas son evaluar y comparar las tendencias en la respuesta a través del tiempo. Esto puede involucrar comparaciones de tratamientos en tiempos específicos o comparaciones de los promedios en todo el tiempo. También puede involucrar comparaciones de tiempos dentro de un tratamiento. Estos son objetivos comunes de cualquier experimento factorial; sin embargo, la característica de los experimentos de mediciones repetidas que requiere especial atención en el análisis de los datos es el tipo de correlación entre las respuestas sobre la misma unidad experimental a través del tiempo.

Existen varios métodos estadísticos utilizados para el análisis de datos de mediciones repetidas. Estos van desde los más básicos hasta los más sofisticados. Estos incluyen a) análisis separados en cada punto del tiempo, b) análisis de varianza univariado, c) análisis univariado y multivariado de contrastes de variables en el tiempo, y d) metodología de modelos mixtos. El análisis separado en cada punto del tiempo no requiere de métodos especiales para medidas repetidas; sin embargo, no cumple con el objetivo de examinar y comparar tendencias a través del tiempo. Las otros tres métodos requieren de software y una metodología especial. El objetivo del estudio fue utilizar la metodología de modelos mixtos para analizar datos de mediciones repetidas como son los datos de producción de leche de vacas Holstein-Friesian sometidas a un programa de somatotropina bovina recombinante.

## 2 Materiales y métodos

Los datos utilizados para la aplicación de la metodología de modelos lineales mixtos provinieron de un experimento realizado en un establo lechero ubicado en el estado de Chihuahua, México. El objetivo del experimento fue comparar el efecto de la aplicación de somatotropina bovina recombinante en ganado lechero. Se formaron dos grupos de tratamiento con 24 vacas cada uno. Los animales se asignaron de manera aleatoria a cada grupo de tratamiento. Cada vaca es una unidad experimental. A los animales del grupo experimental (*A*) se les aplicaron dosis de 500 mg de somatotropina bovina recombinante, adicionada con 1655 UI de vitamina E y 166.5 mg de lecitina (Boostin-S®, Lab. Schering-Plough) por vía subcutánea en la fosa isqueo-rectal, cada 14 días. A los animales del grupo control (*B*) se les aplicó solución salina fisiológica. La producción de leche se midió semanalmente en cada vaca durante 8 semanas consecutivas, tiempo de duración del experimento. El análisis estadístico se enfocó a comparaciones de tratamientos en tiempos específicos, com-

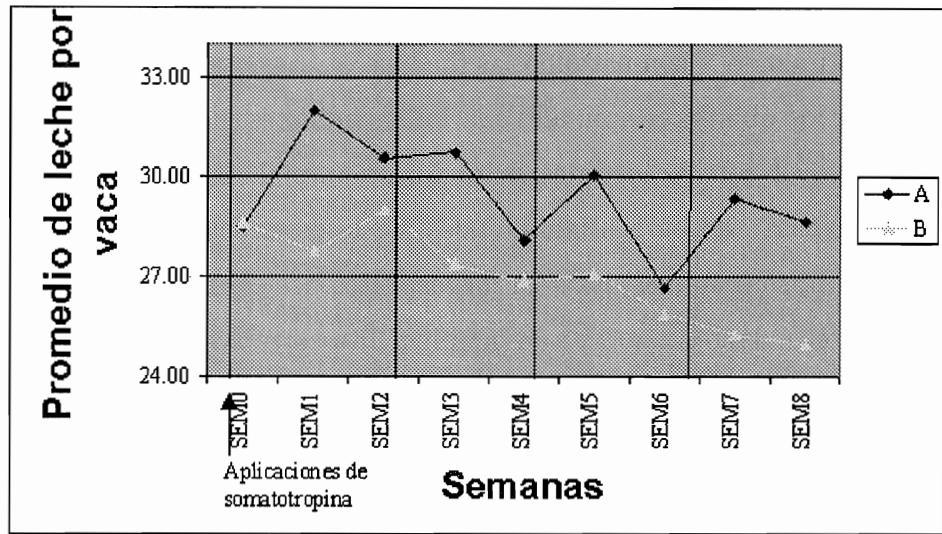


Figura 1. Medias de producción de leche por vaca por semana en los grupos de tratamiento.

paraciones de tratamiento promediados en todos los tiempos y a cambios a través del tiempo en tratamientos específicos. Las diferencias entre los tratamientos *A* y *B* se calcularon en tiempos individuales y promediados a través del tiempo. Los errores estándar se calcularon con base en el método de análisis. Para el análisis de la información, los datos se organizaron en “modo univariado”; es decir, un renglón por unidad experimental en cada tiempo, con todas las medidas de producción de leche como valores de la variable de respuesta llamada prod (producción) de leche. Para modelar la estructura de covarianza se utilizó el PROC MIXED del Sistema de Análisis Estadístico (SAS, 1999), donde la variación entre los animales se especificó con la instrucción RANDOM, y la covariación dentro de los animales se especificó con la instrucción REPEATED.

### 3 Resultados y discusión

El análisis de datos de mediciones repetidas requiere de especial atención sobre la estructura de covarianza debido a la naturaleza secuencial de los datos en cada unidad experimental (animal). Ignorar la estructura de la covarianza puede resultar en conclusiones incorrectas de los análisis estadísticos. Evitar la estructura de covarianza puede resultar en análisis ineficientes, lo cual es equivalente a echar a perder los datos. El modelo mixto lineal general permite modelar la estructura de covarianza. Existen dos pasos básicos para llevar a cabo un análisis de mediciones repetidas utilizando la metodología de modelos mixtos. El primero es

modelar la estructura de covarianza. El segundo es analizar las tendencias del tiempo para los tratamientos estimando y comparando las medias.

Las medidas en los diferentes animales son independientes, de tal manera que la covarianza sólo tiene que ver con las medidas en el mismo animal. La estructura de covarianza se refiere a las varianzas en tiempos individuales y a la correlación entre medidas en diferentes tiempos en el mismo animal. Existen dos aspectos básicos de la correlación. Primero, dos medidas sobre el mismo animal están simplemente correlacionadas debido a que ellas comparten contribuciones comunes del animal. Esto es debido a la variación entre animales. Segundo, las medidas en el mismo animal más cercanas en el tiempo están frecuentemente más altamente correlacionadas que las medidas mas apartadas en el tiempo. Esto es la covariación (covarianza) dentro de los animales.

En el presente trabajo se ilustran tres diferentes estructuras para los datos de producción de leche y se eligió una como la mejor entre las tres. La primer estructura conocida como simetría compuesta (CS) especifica que las medidas en todos los tiempos tienen la misma varianza, y que todos los pares de medidas en el mismo animal tiene la misma correlación. La implicación es que el único aspecto de la covarianza entre las medidas repetidas es debido a la contribución animal, sin importar la proximidad del tiempo.

Utilizando los parámetros estimados de covarianza de los dos componentes de varianza, la correlación entre dos mediciones en el mismo animal, asumiendo estructura de simetría compuesta es:  $r_{CS} = 5.679/(5.679 + 6.803) = 0.454$ .

La segunda estructura general se indica como "UN" (no estructurado). Esta estructura no hace supuestos en cuanto a igualdad de varianzas o correlaciones. Existen dos grandes problemas potenciales al usar la estructura de covarianza no estructurada. 1) Se requiere de la estimación de un gran número de parámetros de varianza y covarianza (48 para este ejemplo) y puede llevar a severos problemas de computo, especialmente con datos desbalanceados. 2) No se explota la existencia de las tendencias en las varianzas y covarianza a través del tiempo, lo que a menudo resulta en modelos erróneos de estimados de errores estándar. La tendencia en las correlaciones observadas, puede ser modelada utilizando una combinación de la estructura autoregresiva dentro de animales y un efecto aleatorio entre animales. Esta estructura combinada especifica un efecto aleatorio inter-animal de diferencias entre animales, y una estructura de correlación dentro de animales (intra-animal) que disminuye con el incremento

en el intervalo de tiempo entre las mediciones. La función de correlación para AR(1) más la estructura del efecto aleatorio es:  $r_{AR(1)} + RE \text{ (lag)} = (5.480 + 6.991).106\text{lag})/(5.480 + 6.991)$ , donde lag es la longitud del intervalo de tiempo entre las mediciones. La concordancia entre  $r_{AR(1)} + RE \text{ (lag)}$  y  $r_{UN} \text{ (lag)}$  es buena.

Las estructuras de covarianza se pueden comparar de manera objetiva utilizando un criterio de bondad de ajuste como el logaritmo de máxima verosimilitud restringida (REML logL), el criterio de información de Akaike (AIC) y el criterio bayesiano de Schwarz (SBC). El AIC y SBC son versiones ajustadas del REML logL que imponen una penalización de acuerdo al número de parámetros estimados. La penalización impuesta por SBC es más severa que la impuesta por AIC. En el presente trabajo se utiliza el criterio SBC para elegir la mejor estructura de covarianza (Tabla 1).

Estructura de covarianza	No. de parámetros	Criterio Bayesiano de Schwarz
Simetría compuesta (CS)	2	-1048.55
No estructurado (UN)	48	-1144.00
Autoregresivo de orden 1 + RE (AR(1)+RE)	3	-1050.11

Tabla 1: Valores del Criterio Bayesiano de Schwarz para cada estructura de covarianza.

El SBC es negativo en este ejemplo. El valor más grande de SBC es la mejor estructura. Los valores de SBC para CS y AR(1) + RE son semejantes. De tal manera que, los valores de AR(1) + RE se utilizaron como la estructura de covarianza para este ejemplo.

## 4 Pruebas de efectos fijos para diferentes estructuras de covarianza

El procedimiento MIXED imprime pruebas para todos los efectos fijos listados en la instrucción MODEL. Estas pruebas son similares a las pruebas F del análisis de varianza univariado. Las pruebas para los efectos fijos de la instrucción MODEL con las tres estructuras de covarianza utilizadas se muestran en el Tabla 2. Los resultados de la estructura de covarianza CS serían válidos si los supuestos de la estructura de covarianza de simetría compuesta o de Huynh-Feldt (H-F) se cumplen. La condición requerida para validar las

pruebas del análisis de varianza univariado es la condición llamada Huynh-Feldt (Huynh y Feldt, 1970), la cual matemáticamente es menos rigurosa que la igualdad de varianzas y covarianzas.

Estructura de covarianza				
Fuente de variación	g.l.	Simetría compuesta	No estructurado	AR + Efecto aleatorio
Tratamiento	1	10.22 (0.0025)	9.72 (0.0031)	10.25 (0.0025)
Semana	8	12.24 (0.0001)	12.72 (0.0001)	10.98 (0.0001)
Tratamiento * Semana	8	0.33 (0.0001)	5.66 (0.0001)	4.80 (0.0001)

Tabla 2: Valores de F y significancia para pruebas de efectos fijos utilizando diferentes estructuras de covarianza en el PROC MIXED.

## 5 Estimación y comparación de medias

Se utilizan dos tipos de comparaciones para ilustrar los efectos de la estructura de covarianza sobre los estimadores y los errores estándar de los estimadores.

La primer comparación es la diferencia entre las medias para los tratamientos A y B promediados a través de las nueve semanas. La segunda comparación son las diferencias entre las medias para el tratamiento A y B en cada semana. En el Tabla 3 se presentan los estimados y errores estándar de las diferencias entre medias para los tratamientos A y B promediados a través de todos las semanas y en semanas individuales utilizando modelos mixtos con estructuras de covarianza CS, UN, y AR(1)+RE y se grafican en la Figura 1. Diferentes estructuras de covarianza producen diferentes errores estándar de los estimadores. La estructura de covarianza que proporciona el mejor ajuste es el apropiado para usarse, aunque puede no resultar en los errores estándar más pequeños. Líneas arriba se mencionó que la estructura AR(1)+RE fue la mejor entre las estructuras de covarianza con base en el criterio SBC. De tal manera que los errores estándar resultantes de AR(1)+RE se consideran los más apropiados.

## 6 Conclusiones

Los modelos lineales mixtos pueden ser implementados con el procedimiento MIXED del

SAS. El procedimiento MIXED es un procedimiento de modelos mixtos que hace cálculos válidos para las pruebas de hipótesis y estimados de los errores estándar. Una aplicación importante de los modelos lineales mixtos es en el análisis de datos de mediciones repetidas. La metodología del modelo mixto implementada en el procedimiento MIXED hace posible analizar datos de mediciones repetidas correcta y eficientemente al modelar la estructura de varianza y la correlación de las mediciones repetidas. La estructura de covarianza estimada se utiliza para obtener estimados de cuadrados mínimos generalizados de tratamiento y diferencias de tiempo. No elegir la estructura de covarianza apropiada puede afectar el cálculo de los estimados, particularmente con datos desbalanceados.

	Estructura de covarianza		
	Simetría compuesta	No estructurado	AR + Efecto aleatorio
Tratamiento A	29.321 ± 0.519	29.274 ± 0.521	29.323 ± 0.519
Tratamiento B	26.976 ± 0.517	26.976 ± 0.520	26.976 ± 0.517
Tratamiento A semana 1	28.500 ± 0.721	28.500 ± 0.772	28.500 ± 0.720
Tratamiento A semana 2	32.000 ± 0.721	32.000 ± 0.816	32.000 ± 0.720
Tratamiento A semana 3	30.583 ± 0.721	30.583 ± 0.733	30.583 ± 0.720
Tratamiento A semana 4	30.750 ± 0.721	30.750 ± 0.695	30.750 ± 0.720
Tratamiento A semana 5	28.083 ± 0.721	28.083 ± 0.574	28.083 ± 0.720
Tratamiento A semana 6	30.041 ± 0.721	30.041 ± 0.766	30.041 ± 0.720
Tratamiento A semana 7	26.384 ± 0.741	26.385 ± 0.769	26.390 ± 0.741
Tratamiento A semana 8	29.111 ± 0.741	29.020 ± 0.659	29.126 ± 0.741
Tratamiento B semana 1	28.625 ± 0.721	28.000 ± 0.772	28.625 ± 0.720
Tratamiento B semana 2	27.750 ± 0.721	27.000 ± 0.816	27.750 ± 0.720
Tratamiento B semana 3	29.000 ± 0.721	29.000 ± 0.733	29.000 ± 0.720
Tratamiento B semana 4	27.375 ± 0.721	27.000 ± 0.695	27.375 ± 0.720
Tratamiento B semana 5	26.875 ± 0.721	26.000 ± 0.574	26.875 ± 0.720
Tratamiento B semana 6	27.083 ± 0.721	27.333 ± 0.766	27.083 ± 0.720
Tratamiento B semana 7	25.875 ± 0.721	25.000 ± 0.747	25.875 ± 0.720
Tratamiento B semana 8	25.250 ± 0.721	25.000 ± 0.645	25.250 ± 0.720
Trat A - Trat B semana 1	2.344 ± 0.733	2.297 ± 0.736	2.346 ± 0.732
Trat A - Trat B semana 2	-0.125 ± 1.019	-0.125 ± 1.093	-0.125 ± 1.019
Trat A - Trat B semana 3	4.250 ± 1.019	4.250 ± 1.155	4.250 ± 1.019
Trat A - Trat B semana 4	1.583 ± 1.019	1.583 ± 1.037	1.583 ± 1.019
Trat A - Trat B semana 5	3.375 ± 1.019	3.375 ± 0.983	3.375 ± 1.019
Trat A - Trat B semana 6	1.208 ± 1.019	1.208 ± 0.812	1.208 ± 1.019
Trat A - Trat B semana 7	2.958 ± 1.019	2.958 ± 1.084	2.958 ± 1.019
Trat A - Trat B semana 8	0.509 ± 1.034	0.510 ± 1.072	0.515 ± 1.033

Tabla 3: Medias ± error estándar y estimados de diferencias ± error estándar entre las medias del tratamiento A y el tratamiento B en todo el experimento y en cada semana.

## Referencias

- Henderson, C.R. (1984). Applications of linear models in animal breeding. *University of Guelph*, Guelph, Ontario.
- Huynh, H. And Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measures designs have exact F-distributions. *J. Amer. Stat. Assoc.*, **65**, 1582-1589.
- SAS. (1989). *SAS/STAT user's Guide (Version 6)*, 4<sup>th</sup> Ed. SAS Inst., Inc., Cary, NC.
- Searle, S.R. (1971). *Linear Models*. John Wiley & Sons. New York.

# Cassandra, Software de Apoyo Didáctico para la Enseñanza de la Estadística

**Francisco Sánchez Villarreal**

*Facultad de Ciencias, UNAM*

**Aarón Martínez Rangel**

*PRAGMA S.A. de C.V.*

La práctica moderna de la Estadística guarda fuerte vinculación con el uso de computadoras personales y software para cómputo estadístico. Paquetes como SPSSPC, SAS, BMDP, RATS, ECONOMETRICS VIEW. STATGRAPHICS, STATA, etc., se cuentan entre los más populares por reunir amplios menús de procedimientos estadísticos y formas de operación accesibles al usuario. Sin embargo se tiene la desventaja del costo de las licencias de uso por ser productos extranjeros y el que los manuales de operación suelen omitir detalles metodológicos. Esto limita su presencia en las aulas.

Esta situación y el interés por investigar las técnicas de cómputo y algoritmos de análisis numérico requeridos por las diferentes metodologías de análisis estadístico, nos ha llevado al desarrollo de un paquete adaptado a los programas de las materias de Muestreo, Estadística I y Estadística II que se incluyen en el plan de estudios de la Carrera de Actuaría y que se imparten en la Facultad de Ciencias de la Universidad Nacional Autónoma de México.

Presentamos la versión 2 del Paquete de Análisis Estadístico, que hemos denominado **Cassandra**, como recuerdo del personaje mitológico Casandra, hija del Rey Príamo y a la que se le atribuía el poder de ver el futuro y lo que ocurría en lugares lejanos. Los Métodos Inferenciales y de Pronóstico que proporciona la Estadística actual, cumplen en un contexto técnico y científico tales funciones, sin el velo de romanticismo y esoterismo de la antigüedad clásica.

**Cassandra** se desarrolló en Delphi IV para Windows y cuenta con las siguientes opciones generales:

- Manejo de Archivos de Datos
- Procedimientos de Estadística Descriptiva
- Inferencia Estadística para medias y proporciones
- Cálculo de probabilidades para las principales funciones de distribución
- Análisis de Regresión Múltiple
- Modelos de Regresión Linealizables Bivariados
- Análisis de Varianza para los Principales Modelos Lineales de Diseño de Experimentos
- Cálculos Básicos de Muestreo Aleatorio Simple y Estratificado
- Módulo de Ayuda Local al Usuario y Vía Correo Electrónico

Para explicar mejor la forma de operación de **Cassandra**, se procederá con un ejemplo sencillo:

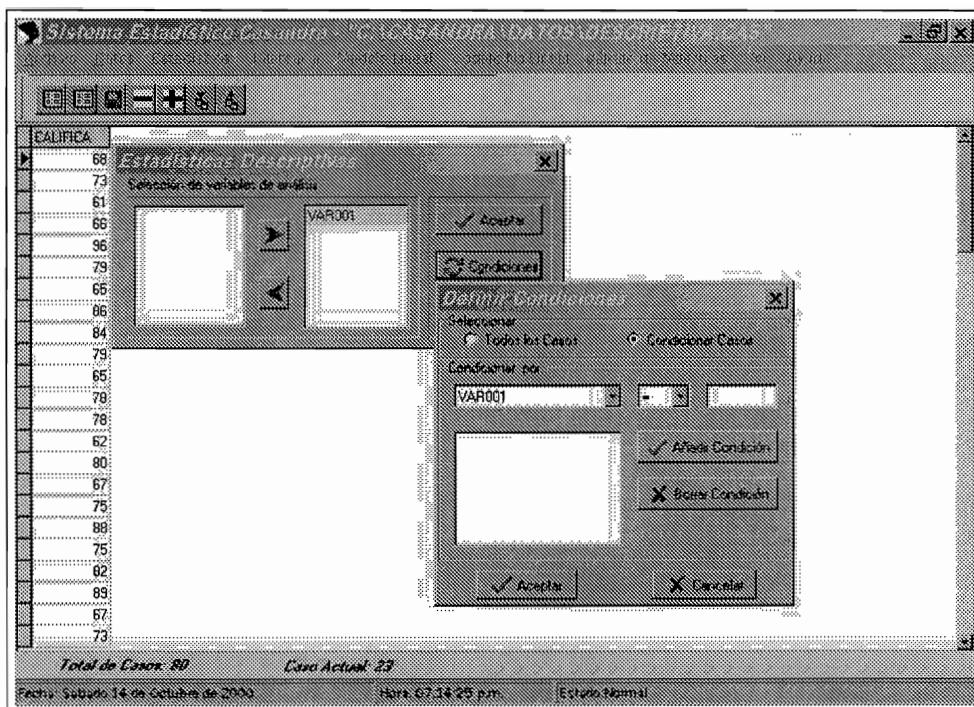
El usuario procederá a incorporar sus datos en la columna correspondiente a una variable que **Cassandra** identifica inicialmente como VAR001.

A continuación se graba el conjunto de datos como un archivo de nombre DESCRIPTIVA.CAS que es la estructura propia de **Cassandra**.

Una vez grabado el archivo, se procede a calcular las **Estadísticas Básicas** mediante la selección de la opción **Estadísticas**, que al abrir la ventana de diálogo dispone de dos opciones: **Descriptiva y Frecuencias**.

Posteriormente el usuario selecciona la variable VAR001 (CALIFICA) para efectuar el proceso de estadísticas básicas. Si el usuario hubiera tenido un archivo con más variables para calcular estadísticas descriptivas, en este punto podría haber seleccionado las otras variables. También es posible imponer condiciones de cálculo en esta opción, por ejemplo excluir valores fuera de cierto rango.

Los resultados del cálculo de estadísticas básicas son reportados inmediatamente en la forma siguiente:



Estadísticas Básicas	
Estadísticas Básicas	
Variable	
Número de datos	80
Media Aritmética	75.2500
Varianza n	106.2875
Varianza n-1	107.6329
Error Estándar	1.1599
Máximo	97.0000
Coef. de Asimetría	0.1712
Coef. de Curtosis	-0.5554
Coef. de Variación	0.1379
Límite Inferior	74.5442
Suma de Datos	6,020.0000
Mediana	75.0000
Desv. Estándar n	10.3095
Desv. Estándar n-1	10.3745
Rango	44.0000
Mínimo	53.0000
Límite Superior	75.9558

El reporte puede ser impreso directamente o ser exportado en diferentes formatos para que el usuario pueda someterlo a edición. Los formatos de exportación son: Excel, DBF (DBASE III) y ASCII (TXT).

Como ejemplo de otro procedimiento muy popular se presenta **Análisis de Regresión Simple** del menú: **Modelo Lineal**, que incluye además regresión múltiple, polinomial y

Analisis de Regresión				
Análisis de Regresión Simple				
Media:	EDAD (X)			
Var:	52.50000			
Desviación Estándar:	203.21053			
Pendiente	136.20000			
Error Estándar de la Pendiente	401.11579			
Intercepto	14.46411			
Error Estándar del Intercepto	20.02789			
r (Coeficiente de Correlación)	1.37107			
r <sup>2</sup> (Coeficiente de Determinación)	0.04561			
Error Estándar de Estimación	2.47360			
	0.99018			
	0.98047			
	2.87589			
Tabla de Análisis de Varianza				
Fuente de Variación	Sumas de Cuadrados	Grados de Libertad	Cuadrados Medios	F Calculada
Regresión	7.472.32704	1	7.472.3270	903.4676
Residual	148.67296	18	8.2707	
Total	7.621.20000	19		
Probabilidad Asociada	0.00000			

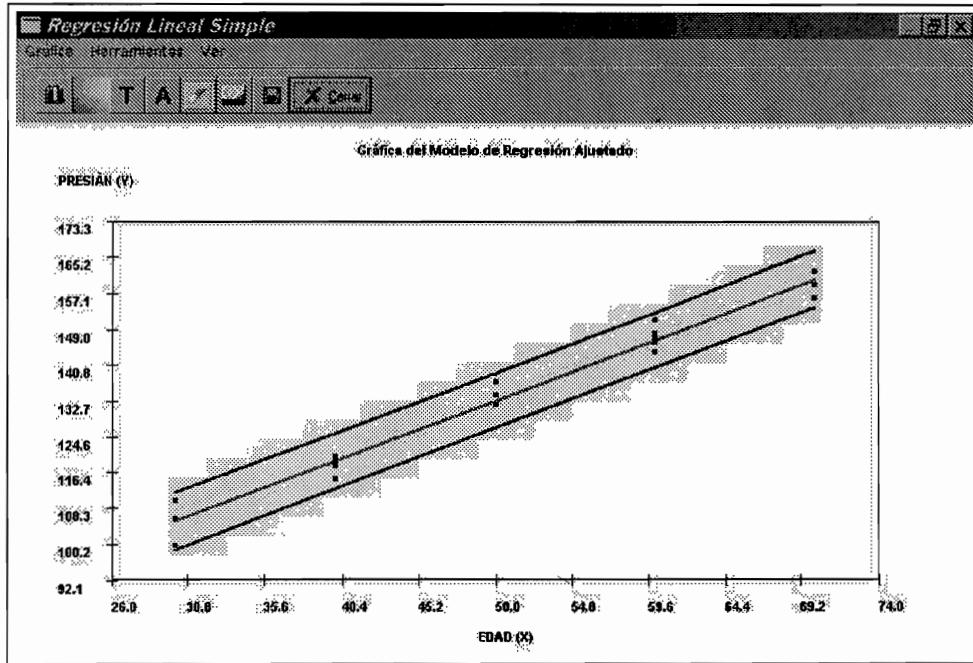
forzada al origen. En un menú aparte se tiene la posibilidad de ajustar modelos no lineales pero linealizables mediante transformaciones logarítmicas y recíprocas.

El usuario procede a integrar un archivo con al menos dos variables: la variable dependiente y la independiente . En el ejemplo se tienen registros de edad de un grupo de personas (X) y su presión sistólica (Y).

Una vez integrado el archivo, se selecciona Análisis de Regresión Simple, inmediatamente se seleccionan la variable dependiente e independiente.

El reporte emitido con estimaciones de parámetros, estadísticas básicas y tabla de análisis de la varianza.

Además del reporte, el usuario puede optar por obtener la gráfica de dispersión de los puntos (X,Y), la recta ajustada y las bandas de 95% de confianza para el valor puntual de Y y el valor medio de Y. También una gráfica de los residuales en función de X.



## Bibliografía

- Carnahan, H.A. Luther *Applied Numerical Methods*, John Wiley & Sons Inc. USA, 1969.
- Cochran, W. G. *Técnicas de Muestreo*, CECSA; México 1971.
- Conover, W.J. *Practical Nonparametric Statistics*, John Wiley & Sons Inc. USA, 1971.
- Freund, J.E, *Mathematical Statistics*, Prentice-Hall, Inc. Englewood Cliffs, N.J., 1971.
- Harris, B. *Theory of Probability*, Addison Wesley, USA, 1966.
- Johnstohn, J. *Econometric Methods*, Mc Graw-Hill, New York, 1960.
- Nie, H., Norman *SPSS Statistical Package for the Social Sciences*. Second Edition, Mc Graw-Hill, New York, 1975.
- Zar, J. H. *Biostatistical Analysis*, Prentice-Hall, Inc.; Englewood Cliffs, N.J. 1974.



# Métodos Estadísticos en la Normatividad en Metrología

**Cristina Segura Cabrera**

*Centro Nacional de Metrología*

**Eduardo Castaño Tostado**

*Universidad Autónoma de Querétaro*

## 1 Introducción

Dentro del ámbito metrológico es de vital importancia conocer el valor de los patrones de referencia. El valor de un patrón se puede obtener a través de:

- La materialización de la unidad y transferencia del valor de la unidad al patrón de referencia: para laboratorios primarios, es más económico y fácil de utilizar un patrón de referencia que aplicar la materialización de una unidad, por lo cual se calibran dichos patrones en intervalos regulares.
- Realizar calibraciones periódicas: cuando no se cuenta con patrones primarios, los patrones de referencia de laboratorios se envían a calibrar en intervalos periódicos.

En ambos casos es necesario aplicar una metodología normalizada que permita monitorear que el valor que mantienen los patrones se encuentra dentro de los parámetros de calibración e incertidumbre, entre una calibración y otra, así mismo se requiere monitorear que los patrones se encuentran dentro de control estadístico para detectar cualquier condición de operación o de medición especial. A este monitoreo se le llama mantenimiento por intercomparación de patrones, en el cual se supone que el promedio de los valores de calibración del grupo se mantiene constante a través del tiempo entre una calibración y otra. El proceso de intercomparación se refiere a obtener diferencias al compararlos entre sí y por medio de un modelo de regresión determinar valores estimados.

## 2 Planteamiento del problema

Denote por  $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_k$  a los valores de calibración de  $k$  patrones. La intercomparación es necesaria para verificar en un punto en el tiempo si tales valores de calibración no han cambiado. La intercomparación intenta estimar tal cambio por medio de un puente de medición con dos posiciones, derecha (D) e izquierda (I), como se muestra en la Figura 1 (Eicke y Cameron, 1967).

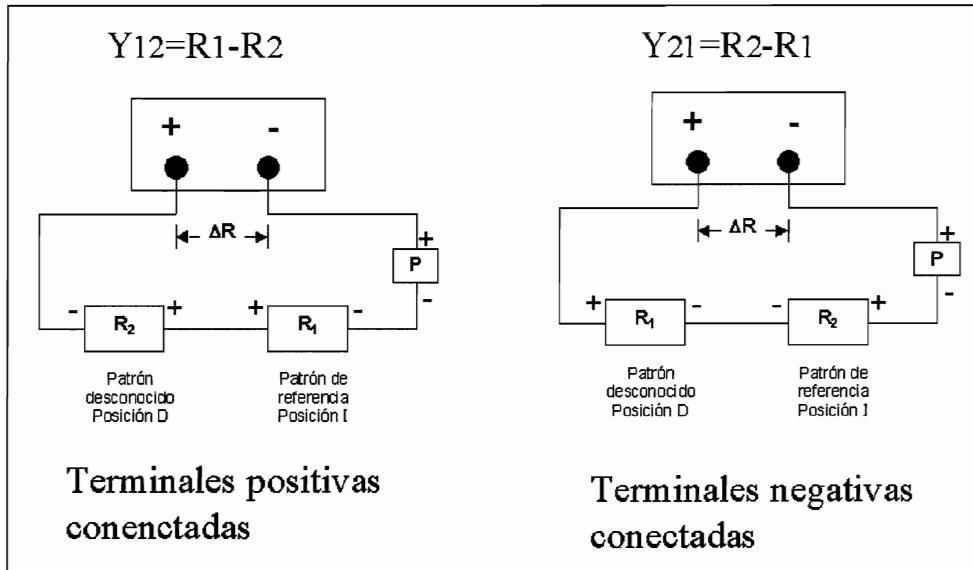


Figura 1

Todo puente de medición implica un error sistemático no medible, denotado por  $P$ . Si  $Y_{12}$  representa la lectura de la diferencia de medición entre dos equipos uno en la posición D y otro en la posición I, entonces se puede pensar que

$$Y_{12} = P + R_1 - R_2 \quad (1)$$

donde  $R_1$  y  $R_2$  denotan los valores de los equipos respectivamente que no son observables sino sólo de manera indirecta a través de su diferencia contaminada por  $P$ . Al realizar una segunda medición con los patrones en posición inversa se tendrá

$$Y_{21} = P + R_2 - R_1; \quad (2)$$

sumando (1) y (2), se obtiene un estimado de  $P$ , dado por

$$2P = Y_{21} + Y_{12},$$

y al obtener la diferencia de (1) y (2) se tiene que

$$Y_{12} - Y_{21} = 2(R_1 - R_2),$$

pudiendo así obtener un valor estimado para  $R_1 - R_2$  libre de  $P$ . Sea  $\bar{M}$  el promedio de los valores resultado de la calibración inicial, es decir,

$$\bar{M} = \frac{\mathcal{R}_1 + \mathcal{R}_2 + \dots + \mathcal{R}_k}{k};$$

denotemos por  $r_j = R_j - \bar{M}$ , las respectivas desviaciones del promedio de los valores de la calibración actual.

Para cada comparación medición  $Y_{jj'}$  se supone que se involucran pequeños errores aleatorios  $\varepsilon_{jj'}$ , tales que

$$Y_{jj'} = P + R_j - R_{j'} + \varepsilon_{jj'}$$

de los que se supone, dada una estandarización adecuada del proceso de intercomparación, que

$$E[\varepsilon_{jj'}] = 0 \quad \text{y} \quad V[\varepsilon_{jj'}] = \sigma^2, \quad \text{Cov}(\varepsilon_{jj'}, \varepsilon_{lk}) = 0, (j, j') \neq (l, k).$$

En este sentido, realizar un experimento de intercomparación resulta en estimar los parámetros del modelo

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$$

donde  $\mathbf{Y} = \{Y_{jj'}\}$ , la matriz  $X$  es una matriz que en la fila correspondiente a  $Y_{jj'}$  tendrá un 1 en la primera posición, un 1 en la posición  $j$  y un  $-1$  en la posición  $j'$ ; el vector de parámetros es  $\boldsymbol{\beta} = (P, r_1, \dots, r_k)'$ . Toda esta estructura que corresponde a un experimento de intercomparación, implica que el objetivo es estimar  $\boldsymbol{\beta}$ ; una opción es entonces, utilizar como criterio de optimización la minimización de  $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ , la suma de cuadrados de los errores al cuadrado.

Sin embargo, dada la estructura del experimento de intercomparación, la matriz de diseño  $X$  es de rango  $r < k + 1$ , incompleto, entonces el problema de estimación de  $\boldsymbol{\beta}$  por mínimos cuadrados no tiene una solución única, por lo que se pueden seleccionar restricciones de la

forma  $H\beta = 0$ , tales que las filas de  $H$  forman un conjunto de  $k+1-r$  vectores linealmente independientes de las filas de  $X$ . Así (Seber, 1977),

$$\begin{pmatrix} X \\ H \end{pmatrix} \beta = G\beta$$

con  $G$  de rango completo. Entonces  $G'G = X'X + H'H$  será de rango  $k+1$ , y así

$$G'G\hat{\beta} = X'\mathbf{Y},$$

con lo que

$$\hat{\beta}_H = (G'G)^{-1}X'\mathbf{Y},$$

y

$$\hat{V}(\hat{\beta}_H) = \hat{\sigma}^2 C$$

donde  $C = (G'G)^{-1}X'X(G'G)^{-1}$ . Si  $L$  denota una combinación lineal de  $r'_1$  s,  $L = a_0P + a_1r_1 + \dots + a_kr_k$ , entonces la desviación estándar de

$$\hat{L} = a_0\hat{P} + a_1\hat{r}_1 + \dots + a_k\hat{r}_k,$$

se sabe que es

$$\hat{\sigma}\sqrt{\mathbf{l}'C\mathbf{l}},$$

donde  $\mathbf{l} = (a_0, a_1, \dots, a_k)'$  es el vector de coeficientes de  $L$ ,  $\hat{\sigma}$  es la suma de cuadrados medios de la regresión correspondiente; con lo cual es posible estimar un intervalo de confianza, por ejemplo para  $r_1$ ,  $\mathbf{l}$  con 0 en todos sus elementos excepto el  $i+1=1$  el intervalo de confianza será

$$\hat{r}_1 \pm t_{n-k}^{\alpha/2} \sqrt{c_{ii}} \hat{\sigma}.$$

También es posible calcular los intervalos de confianza Bonferroni los cuales se obtienen de

$$\hat{r}_1 \pm t_{n-k}^{\alpha/(2m)} \sqrt{c_{ii}} \hat{\sigma},$$

donde  $m$  es el número de intervalos por construir.

### 3 Ejemplo en el mantenimiento de cuatro patrones de resistencia

Se tiene un conjunto de cuatro resistores patrón con valor nominal de 1 los cuales se comparan utilizando el sistema de medición descrito y realizando las mediciones de comparación descritas en la figura 1 y realizadas en el orden establecido en la Tabla 1.

$R_1$	$R_2$	$R_3$	$R_4$	$P$	$Y_{jj'}$
1	-1	0	0	1	$Y_{12} = 1.11 \times 10^{-6}$
1	0	-1	0	1	$Y_{13} = -4.84 \times 10^{-6}$
1	0	0	-1	1	$Y_{14} = -1.09 \times 10^{-6}$
-1	1	0	0	1	$Y_{21} = -1.20 \times 10^{-6}$
0	1	-1	0	1	$Y_{23} = -6 \times 10^{-6}$
0	1	0	-1	1	$Y_{24} = -2.24 \times 10^{-6}$
-1	0	1	0	1	$Y_{31} = 4.75 \times 10^{-6}$
0	-1	1	0	1	$Y_{32} = 5.92 \times 10^{-6}$
0	0	1	-1	1	$Y_{34} = 3.71 \times 10^{-6}$
-1	0	0	1	1	$Y_{41} = 1 \times 10^{-6}$
0	-1	0	1	1	$Y_{42} = 2.13 \times 10^{-6}$
0	0	-1	1	1	$Y_{43} = 0 \times 10^{-6}$

Tabla 1. Diseño de intercomparación de 4 patrones

En este caso, dado que el rango de  $X$  es 4, se requiere de una restricción a través de

$$H = (1, 1, 1, 1, 0).$$

El estimador por mínimos cuadrados resulta entonces

$$\hat{\beta} = \begin{bmatrix} \hat{r}_1 = -1.17125 \times 10^{-6} \\ \hat{r}_2 = -2.325 \times 10^{-6} \\ \hat{r}_3 = 3.1525 \times 10^{-6} \\ \hat{r}_4 = 0.34375 \times 10^{-6} \\ \hat{P} = 0.27083 \times 10^{-6} \end{bmatrix}.$$

Dado que

$$\bar{M} = \sum \mathcal{R}_1 = 1.00000343,$$

entonces

$$\begin{bmatrix} \hat{R}_1 = 1.00000226 \\ \hat{R}_2 = 1.00000110 \\ \hat{R}_3 = 1.00000658 \\ \hat{R}_4 = 1.00000377 \end{bmatrix}.$$

La matriz de varianzas y covarianzas de los valores estimados es:

$$\begin{bmatrix} 0.0122 \times 10^{-6} & -0.0006 \times 10^{-6} & -0.0006 \times 10^{-6} & -0.0006 \times 10^{-6} & 0 \\ -0.0006 \times 10^{-6} & 0.0122 \times 10^{-6} & \times 10^{-6} & -0.0006 \times 10^{-6} & 0 \\ -0.0006 \times 10^{-6} & -0.0006 \times 10^{-6} & 0.0122 \times 10^{-6} & -0.0006 \times 10^{-6} & 0 \\ -0.0006 \times 10^{-6} & -0.0006 \times 10^{-6} & -0.0006 \times 10^{-6} & 0.0122 \times 10^{-6} & 0 \\ 0 & 0 & 0 & 0 & 0.0183 \times 10^{-6} \end{bmatrix}.$$

Los intervalos para cada valor estimado de los patrones con una confianza individual de 95% son

$$R_1\epsilon(1.00000223, 1.00000229), R_2\epsilon(1.00000107, 1.00000113),$$

$$R_3\epsilon(1.00000651, 1.00000655), R_4\epsilon(1.00000374, 1.00000380).$$

Los intervalos de confianza Bonferroni al menos 95% de confianza conjunta, son:

$$R_1\epsilon(1.00000222, 1.00000230) R_2\epsilon(1.00000106, 1.00000115)$$

$$R_3\epsilon(1.00000654, 1.00000662) R_4\epsilon(1.00000373, 1.00000381).$$

Se puede observar que los valores de la calibración inicial  $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \mathcal{R}_4$  están contenidos en estos intervalos de confianza.

La estimación por medio del uso de modelos de regresión, en un punto en el tiempo, debe ser repetida de manera periódica durante un lapso que permita tener una idea clara de las tendencias de los valores estimados en el tiempo; para su monitoreo se usan cartas de control estadístico.

## 4 Conclusiones

Las herramientas estadísticas presentadas en esta aplicación en metrología, son sólo un ejemplo de las amplias posibilidades en la relación de la estadística y la metrología. Este tipo de herramientas resultan útiles para operacionalizar la intercomparación requerida para el

mantenimiento de un patrón. Su uso impone un orden para realizar de manera normalizada la intercomparación. Debe impulsarse su conocimiento en detalle por los metrólogos para evitar su aplicación a nivel de receta. La capacitación formal en métodos estadísticos aplicados a la metrología es esencial, obteniéndose los siguientes beneficios:

1. se apoya de manera operacional en la garantía de la trazabilidad de servicios de calibración.
2. se tendrán estimaciones de la estabilidad a largo plazo de patrones de medición, que es un componente difícilmente medible.
3. el uso de modelos impone un orden que permite detectar actividades de la intercomparación importantes de normalizar, con la consecuente reducción de la variabilidad aportada por tales actividades.

## Referencias

- Eicke, W.G. y Cameron, I.J.M. (1967). *Nota técnica 430: Diseño para el estudio del mantenimiento del volt utilizando un grupo de celdas patrón saturadas*. NIST.  
Seber, G.A.F.(1977). *Linear Regression Analysis*. Wiley. New York.



# Estadística y Metroología

Enrique Villa Diharce

*Centro de Investigación en Matemáticas, A.C.*

## 1 Introducción

La metroología y la estadística guardan entre si, una relación sinérgica, donde ambas se necesitan y se fortalecen. Por ejemplo, un requisito fundamental para que una Red de Monitoreo Ambiental genere datos confiables, es tener un programa de aseguramiento de mediciones, y para esto la herramienta estadística es determinante. Esto es, la metroología depende de la estadística. También esta relación de dependencia ocurre en sentido opuesto, ya que para un correcto análisis estadístico de datos se requiere, en principio, de un excelente sistema de medición. Esto es, la estadística depende de la metroología.

Si concebimos a la metroología como la ciencia de las mediciones, podemos entender la gran responsabilidad que ésta tiene en la generación de mediciones confiables, que se obtienen utilizando diferentes instrumentos y sistemas de medición. El largo y sinuoso camino que se recorre desde la definición de un sistema de medición hasta llegar a la determinación de una medición, tiene que pasar la mayoría de las veces por el análisis estadístico de datos que se generan dentro del proceso de medición, y que influyen en el valor final de la medición. La inevitable presencia de la estadística en la metroología se debe a que el proceso de medición es de naturaleza aleatoria dado que el resultado arrojado por un instrumento de medición siempre depende de una gran cantidad de factores que son difíciles de cuantificar con exactitud: pericia del operador que utiliza el instrumento de medición, temperatura ambiente, humedad, desgaste del instrumento de medición, entre otros.

Por exigencias del desarrollo tecnológico, los metrólogos debieron cambiar el enfoque de su trabajo: pasar de un conjunto de operaciones de medición a un sistema de medición conformado por una cadena de procesos, en donde la variación está presente en todas las etapas. Esto necesariamente le dió entrada al pensamiento estadístico (G. C. Birtz et al., 2000), que permite diseñar estrategias para perfeccionar el proceso de medición. El desarrollo de programas de mejoramiento en los sistemas de medición, se ha visto favorecido por el

progreso que han tenido las herramientas para el control estadístico de procesos en los últimos años.

Una razón más para la aplicación de la estadística en los procesos de medición, es la exigencia de una mayor exactitud de las mediciones a través del tiempo por necesidad y como resultado del desarrollo tecnológico. Un ejemplo de esto lo tenemos en la medición de la longitud. En el siglo XVIII, se definió al metro como la diezmillonésima parte del cuadrante terrestre, y con esta medida se construyó el metro patrón. En aquel tiempo la incertidumbre del metro llegó a ser del orden de  $10^{-7}$  mts; en 1928, se pensó en redefinir el metro a partir de la luz y su longitud de onda. Actualmente el metro se define como la longitud de la trayectoria que recorre la luz en el vacío en un intervalo de 1/299 792 458 segundo. En este caso la precisión llegó a los  $10^{-13}$  mts. Desde la determinación original del metro la precisión aumentó un millón de veces, pero ahora para llevarla a cabo, se requiere de tecnología avanzada, mientras que en el pasado las mediciones se hacían por comparación directa. El análisis de los datos generados en este proceso de medición requiere de una herramienta más avanzada que en el pasado.

## 2 Sistema metrológico internacional

Los procesos metrológicos son de naturaleza global debido a que generalmente se dan dentro de relaciones de intercambio, en las que están involucradas organizaciones de diferentes puntos del globo terráqueo. Así por ejemplo, en un país se fabrica una parte de automóvil, que termina en una planta armadora de autos ubicada en otro país. En esta relación, las mediciones que hacen quienes producen la pieza y quienes la utilizan en el proceso de armado deben ser coherentes para que la pieza ensamble perfectamente. El carácter global de la sociedad actual, ha obligado a tener un sistema metrológico mundial, a fin de poder tener mediciones comparables universalmente. Lo anterior ha generado un Sistema Metrológico Internacional, preocupado por la globalización de los procesos de medición.

La construcción del Sistema se inició con la Convención del Metro (CM), tratado diplomático firmado 1875, que otorga autoridad a la Conferencia General de Pesas y Medidas (CGPM), al Comité Internacional de Pesas y Medidas (CIPM), y al Buró Internacional de Pesas y Medidas (BIPM) para actuar a nivel internacional en materia de metrología.

En otro nivel del sistema se encuentran los Centros Nacionales de Metrología (CNM),

que mantienen los patrones de referencia de las diferentes unidades de medición en cada país, y son la autoridad metrológica nacionalmente. En cada país, los laboratorios secundarios (LS), realizan el trabajo amplio de metrología, que consiste en calibrar todo tipo de instrumentos de medición, y difundir el sistema internacional de unidades a su más amplia aplicación. Estos laboratorios orientan en gran medida su trabajo hacia la industria, debido a la demanda creciente que se tiene, fundamentalmente por los requerimientos de las certificaciones en diferentes sistemas de calidad, que los clientes exigen. La última etapa en esta secuencia de organismos metrológicos, son los laboratorios de las industrias, encargados de la organización interna de los sistemas de medición, conformados por la combinación de operarios, instrumentos y procedimientos de medición, que se llevan a cabo en el piso de la fábrica.

En 1978 al reconocerse la falta de consenso alrededor de la expresión de incertidumbres en las mediciones, el CIPM sugirió al BIPM que abordara el problema conjuntamente con los laboratorios nacionales y que hiciera una recomendación. Esta iniciativa generó una amplia discusión sobre la expresión de incertidumbres de medición y culminó con la publicación de un documento, conocido mundialmente como la Guía (ISO 1993) que se ha convertido en un puente entre la estadística y la metrología debido a el papel relevante asignado a la estadística en el estudio de la incertidumbre.

### **3 Incertidumbre de medición**

El desarrollo de la ciencia y la tecnología descansa en las mediciones. En el terreno de la ciencia, los científicos comparan sus resultados, comparando sus mediciones. En la industria se comparan las mediciones de los procesos, con los valores de especificación requeridos por los clientes. Al analizar la diferencia entre un par de mediciones es importante conocer la discrepancia que podemos esperar como consecuencia de el error aleatorio de las mediciones. De esta manera cuando el par de mediciones difiere en una magnitud mayor a la incertidumbre de las mediciones, se puede concluir que la diferencia entre ellas es evidencia suficiente para asevarar que las magnitudes que se miden, son en realidad distintas.

Cuando medimos una magnitud ( $\mu$ ), estamos conscientes de la desviación que hay entre nuestra medición y la magnitud que se mide, por esto reportamos un valor puntual ( $m$ ) como la medición, y además, una medida de la incertidumbre ( $u$ ) de medición.

En el caso de mediciones simples, en donde la medición  $m$  difiere del mensurando  $\mu$  por una cantidad aleatoria  $\varepsilon$ , que expresa todas las fuentes de error, el modelo de medición es

$$m_i = \mu + \varepsilon_i, i = 1, 2, \dots, n, \quad (1)$$

donde suponemos que  $\varepsilon_i$  es una variable aleatoria con media cero. En este modelo  $\varepsilon_i$  representa el error aleatorio en que se incurre al hacer la medición, y en él se encuentran las influencias de las variables ambientales, que el modelo no explica.

Este modelo se hace más complejo a medida que vamos separando las fuentes de error, como por ejemplo, el error debido al operador del instrumento de medición, en tal caso el modelo es

$$m_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, 2, \dots, n, \quad j = 1, 2, \dots, m, \quad (2)$$

donde,  $\alpha_i$  representa el efecto operador en la medición.

Y si además, ubicamos otra fuente de error, como por ejemplo, la variabilidad de las piezas que se miden, el modelo es,

$$m_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, i = 1, 2, \dots, n, \quad j = 1, 2, \dots, m, \quad k = 1, 2, \dots, l, \quad (3)$$

donde  $\beta_j$  es el efecto pieza. De esta forma podemos considerar diferentes efectos aditivos en el proceso de medición.

Este refinamiento de las fuentes de error, se considera cuando se hacen estudios inter-laboratorios, para comparar procesos de medición de diferentes laboratorios, ya sea dentro de un mismo país o entre diferentes naciones. Igualmente es una práctica común cuando se hacen estudios de repetibilidad y reproducibilidad de instrumentos de medición.

El modelo anterior es usual en la práctica de la metrología que se da en los talleres y en las fábricas donde los operadores toman un instrumento de medición como puede ser un voltímetro, una regla, un vernier, etc., y registran directamente la lectura del instrumento.

Sin embargo, en la ciencia y en la tecnología los procesos son cada vez más complejos y se requiere una mayor exactitud, por lo tanto el modelo anterior queda rebasado por no incorporar explícitamente la forma en que influyen las variables ambientales en la medición. Un modelo de medición más global surge al considerar la relación entre la magnitud de interés  $\mu$  (que algunas veces no se mide directamente), y otras magnitudes que influyen de acuerdo a la función

$$\mu = g(\theta_1, \theta_2, \dots, \theta_p; \lambda_1, \lambda_2, \dots, \lambda_r), \quad (4)$$

donde  $\theta_1, \theta_2, \dots, \theta_p$  son magnitudes que se miden, y su incertidumbre se determina por métodos estadísticos, mientras que para las magnitudes  $\lambda_1, \lambda_2, \dots, \lambda_r$ , sus valores e incertidumbre se determinan por métodos no estadísticos, como por ejemplo, juicio experto o estimaciones previas reportadas por otras personas.

Este modelo fue propuesto en la Guía (ISO 1993), en donde el estado del conocimiento tanto de la magnitud de interés  $\mu$ , como de las de influencia  $\theta_s$  y  $\lambda_s$ , se expresa a través de distribuciones de probabilidad, siendo las más comunes, la distribución normal, la uniforme y la triangular.

La distribución uniforme modela el estado del conocimiento de una magnitud, cuando sólo sabemos que ésta se encuentra en un intervalo  $(a, b)$ , y no tenemos evidencia para asignarle a determinada región del intervalo mayor probabilidad de contener a la magnitud. El principio de máxima entropía justifica el modelo uniforme en este caso (Kessel, 2000). Este principio también justifica las otras dos distribuciones cuando tenemos otros niveles de información.

La recomendación dada en la Guía (ISO 1993) para la determinación de la incertidumbre de medición, ha sido de interés para varios estadísticos, porque parece combinar medidas de desempeño frecuentista, con índices de distribuciones subjetivas, en una forma que no es ni frecuentista ni Bayesiana, mas bien parece que establece una solución de compromiso entre los dos enfoques.

Gleser (1998) hace una revisión de la Guía (ISO 1993) y muestra cómo las recomendaciones pueden tomarse como soluciones aproximadas a ciertos problemas de inferencia frecuentista y Bayesiana. Una revisión adicional de la Guía (ISO 1993), es la que hace Bich (1997), quien introduce una generalización a casos multivariados, en donde tenemos  $m$  mediciones de interés que resultan de mediciones que hacemos de  $n$  magnitudes de influencia. Esta es una tarea común en metrología, ya que es usual, calibrar varios instrumentos de medición, utilizando un mismo conjunto de instrumentos o patrones de referencia.

## 4 Areas de oportunidad

El campo metrológico, como cualquier otro campo de aplicación de la estadística, ofrece la posibilidad de introducir una gran variedad de herramientas. Incluso, la diversidad de herramientas es una necesidad por la amplitud del espectro de áreas de investigación y desarrollo de la metrología. Así por ejemplo, en el área de tiempo y frecuencia, el análisis de

series de tiempo es fundamental, mientras que en el área de materiales, la determinación de mediciones ligadas a gases, materiales cerámicos y metálicos, etc., se apoya fundamentalmente en los modelos lineales, y el diseño y análisis de experimentos. En la determinación de intervalos de calibración de instrumentos de medición, la confiabilidad es una herramienta de gran utilidad, ya que el crecimiento de la incertidumbre del instrumento al paso del tiempo, puede interpretarse como una manifestación del proceso de degradación del instrumento. En el desarrollo de programas de aseguramiento de mediciones son necesarias las cartas de control, para una gran variedad de procesos. La presencia de observaciones aberrantes no es extraña en los datos de metrología, por lo cual deben aplicarse técnicas robustas para el análisis de los datos. La determinación de la incertidumbre de medición en casos en que se reducen algunos supuestos del modelo de medición, puede efectuarse, empleando técnicas de bootstrap. La variedad de herramientas que se utilizan en metrología, podemos encontrarla en algunas revistas que publican regularmente artículos en donde interactúan la estadística y la metrología. Entre estas revistas tenemos principalmente, Metrología, Technometrics y el Journal of Quality Technology.

Una fuente adicional de artículos y notas sobre aplicaciones estadísticas en metrología es la serie de textos Advanced Mathematical Tools in Metrology, que contienen los trabajos discutidos en los talleres internacionales sobre herramientas matemáticas avanzadas en metrología que se han realizado en Europa desde 1995, y que han sido publicados por la editorial World Scientific.

## 5 Conclusiones

Por el desarrollo de la ciencia y la tecnología, en un lapso muy breve la estadística y la metrología han entrado en una relación sinérgica. Hay mucho aún por desarrollar en esta relación. En México hasta hace algunos años empezó a ser necesaria la colaboración de estadísticos y metrólogos. Se ha iniciado una relación que es preciso continuar y fortalecer.

## Agradecimientos

Agradezco a el personal del Centro Nacional de Metrología por el apoyo recibido en diferentes formas, a través de material impreso, así como de discusiones. También agradezco a el CONACyT el apoyo recibido a través del proyecto de investigación I32824-E.

## Referencias

- W. Bich. (1997). The ISO Guide to the Expression of Uncertainty in Measurement: A Bridge Between Statistics and Metrology, en *Advanced Mathematical Tools in Metrology III*, P. Ciarlini, M. G. Cox, D. Richter Eds. World Scientific, 1-11.
- Birtz G. C., Emerling D. W., Hare L. B., Hoerl R. W., Hanes S. J., and Shade J. E., (2000). *Improving Performance Through Statistical Thinking*, ASQ Quality Press, Milwaukee, WI. USA
- Gleser, L. J. (1998). Assessing Uncertainty in Measurement, *Statistical Science*, **13**, No. 3, 227-290.
- Kessel, W. (2000). Notas del curso *BIPM-GUM, Concept of Measurement Uncertainty, Bayesian Background?*, impartido en el Taller de Estadística y Metrología, realizado en el CIMAT, del 22 al 24 de enero de 2000.

Esta publicación consta de 410 ejemplares y se terminó de imprimir en el mes de julio del 2001 en los talleres gráficos del **Instituto Nacional de Estadística, Geografía e Informática**  
Av. Héroe de Nacozari Núm. 2301 Sur, Acceso 11, PB  
Fracc. Jardines del Parque, CP 20270  
Aguascalientes, Ags.  
**México**