

# Memorias XX

## Foro Nacional de Estadística

6

4

%



INSTITUTO NACIONAL DE ESTADÍSTICA  
GEOGRAFÍA E INFORMÁTICA

# Memorias

## XX

### Foro Nacional de Estadística



INSTITUTO NACIONAL DE ESTADÍSTICA  
GEOGRAFÍA E INFORMÁTICA

**DR © 2006, Instituto Nacional de Estadística,  
Geografía e Informática  
Edificio Sede  
Av. Héroe de Nacozari Sur Núm. 2301  
Fracc. Jardines del Parque, CP 20270  
Aguascalientes, Ags.**

**www.inegi.gob.mx  
atencion.usuarios@inegi.gob.mx**

**Memorias del XX Foro Nacional de Estadística**

**Impreso en México  
ISBN 970-13-4692-0**

# Presentación

El XX Foro Nacional de Estadística se llevó a cabo del 19 al 22 de septiembre de 2005 en el Centro de Investigación en Matemáticas A C (CIMAT) y la Universidad de Guanajuato, siendo estas las instituciones que tuvieron a cargo la organizaron el evento.

En estas memorias se presentan algunos de los resúmenes de las contribuciones presentadas en este foro. Todos los resúmenes recibidos fueron incluidos sin un proceso de arbitraje, aunque con una detallada revisión.

La Asociación Mexicana de Estadística agradece al CIMAT y a la Universidad de Guanajuato por su apoyo para la realización de este foro y al Instituto Nacional de Estadística, Geografía e Informática (INEGI) el apoyo para la edición de estas memorias.

## **El Comité Editorial:**

Alberto Contreras Cristán

Jesús Armando Domínguez Molina

Elida Estrada Barragán

Ramsés Humberto Mena Chávez



# Contenido

Presentación	I
<b>Selección Bayesiana de factores diseños Plackett-Burman</b> <i>Ernesto Barrios Zamudio</i>	1
<b>Modelos lineales generalizados en el contexto de diseño robusto</b> <i>Guadalupe Bocanegra Aguilar y Jorge Domínguez Domínguez</i>	7
<b>Acerca de la contrucción de modelos AR(1) utilizando densidades predictivas que emergen de la estadística Bayesiana no paramétrica</b> <i>Alberto Contreras, Ramsés Mena-Chávez y Stephen Walker</i>	17
<b>Alcances y limitaciones de S-PLUS para la generación de árboles de escenarios por simulación</b> <i>Melisa Contreras-González y Gladys Linares-Fleites</i>	25
<b>Optimización conjunta de diseño de parámetro y diseño de tolerancia</b> <i>Jorge Domínguez Domínguez y Susana Pérez Santos</i>	31
<b>Evaluación numérica de funciones de distribución multivariadas</b> <i>Armando Domínguez Molina y Alonso Nuñez Páez</i>	37
<b>La aplicación del análisis probit a un experimento agronómico</b> <i>Arely Espinosa, Emilio Padrón-Corral y Félix Sanchez Pérez</i>	45
<b>Muestreo de poblaciones humanas de difícil detección</b> <i>Martín Félix Medina</i>	53

<b>Una metodología para determinar el periodo de garantía para un producto</b>	<b>59</b>
<i>Humberto Gutiérrez-Pulido, Víctor Aguirre y Andrés Christen</i>	
 <b>Intervalos de confianza en el modelo de regresión logística, en presencia de separación de los datos y colinealidad entre las variables explicatorias</b>	<b>67</b>
<i>Flaviano Godínez y Gustavo Ramírez</i>	
 <b>Regresión múltiple para la interpretación de configuraciones de suelos en la sierra norte de Puebla obtenidas por escalamiento multidimensional</b>	<b>77</b>
<i>Gladys Linares-Fleites, Miguel Angel Valera, Guadalupe Tenorio y Maribel Castillo</i>	
 <b>Monte Carlo simulations for Rasch model tests</b>	<b>83</b>
<i>Patrick Mair y Thomas Ledl</i>	
 <b>Clasificación multivariada usando algoritmos genéticos y la función lineal discriminante de Fisher</b>	<b>95</b>
<i>Aurora Montano y Sergio Juárez</i>	
 <b>La política monetaria en el periodo 1984-2004. Confrontación teórica-práctica</b>	<b>101</b>
<i>Federico Muller, Adrián Guerrero y Mónica Rodríguez</i>	
 <b>Análisis Bayesiano de un modelo de regresión para datos circulares</b>	<b>109</b>
<i>Gabriel Nuñez y Eduardo Gutiérrez-Peña</i>	
 <b>Componentes principales y su relación al análisis de estabilidad con aplicación agronómica</b>	<b>115</b>
<i>Emilio Padrón, Ignacio Mendez, Armando Muñoz y Félix Sánchez</i>	

<b>Una prueba para normalidad basada en la propiedad de la cerradura de convoluciones</b>	<b>123</b>
<i>José Villaseñor y Elizabeth González</i>	
<b>Extensión de la prueba <i>t</i> para observaciones intercambiables</b>	<b>129</b>
<i>José Villaseñor y Eduardo Gutiérrez</i>	
<b>Pruebas de hipótesis para procesos Gaussianos</b>	<b>137</b>
<i>José Villaseñor y Eduardo Gutiérrez</i>	
<b>Comparing tests of multinormality. A Monte Carlo study</b>	<b>145</b>
<i>Alexander von Eye</i>	



# Selección Bayesiana de factores diseños Plackett-Burman

Ernesto Barrios Zamudio<sup>1</sup>

*Instituto Tecnológico Autónomo de México*

## 1. Introducción

De las principales diferencias entre la experimentación en las ciencias agrícolas o biomédicas y la ingeniería son el tiempo que se lleva la experimentación misma y el número de ensayos o réplicas disponibles. En la industria, las posibilidades de experimentación son más limitadas. De ahí, la relevancia de los diseños experimentales estadísticos de 2 niveles. De particular importancia es el problema de la identificación de los factores, entre varios posibles, que influyen o no en la respuesta de interés. Llamamos a éstos factores *activos* e *inertes*, respectivamente.

En el *problema de la identificación de factores*, los diseños experimentales basados en arreglos ortogonales son de gran utilidad pues permiten asignar un factor a cada una de las columnas en la matriz del diseño sin el problema de la confusión de los factores. Por ejemplo, para un diseño factorial completo (FC)  $2^4$  de 16 corridas, podríamos investigar los efectos principales que corresponden hasta a 15 distintos factores.

## 2. Diseños Plackett-Burman

Los diseños *Plackett-Burman* (PB) (Plackett y Burman, 1946) son otro tipo de arreglos ortogonales útiles en la selección de factores. Son diseños de  $n = 4m$  corridas con  $m = 1, 2, \dots$ . Si  $n = 2^k$ , los diseños PB coinciden con los FC. El cuadro 1 muestra un arreglo Plackett-Burman de 12 corridas.

Los arreglos PB se construyen mediante el corrimiento de sus filas. El Cuadro 2 muestra las filas generadoras para arreglos PB más utilizados en la industria, diseños de 12, 20 y 24 corridas. Note que la segunda fila del arreglo en la Tabla 1 se obtiene recorriendo la primer

---

<sup>1</sup>ebarrios@itam.mx

Cuadro 1: Diseño Plackett-Burman de 12 corridas.

corrida	Factores											
	A	B	C	D	E	F	G	H	J	K	L	
1	+	-	+	-	-	-	+	+	+	-	+	
2	+	+	-	+	-	-	-	+	+	+	-	
3	-	+	+	-	+	-	-	-	+	+	+	
4	+	-	+	+	-	+	-	-	-	+	+	
5	+	+	-	+	+	-	+	-	-	-	+	
6	+	+	+	-	+	+	-	+	-	-	-	
7	-	+	+	+	-	+	+	-	+	-	-	
8	-	-	+	+	+	-	+	+	-	+	-	
9	-	-	-	+	+	+	-	+	+	-	+	
10	+	-	-	-	+	+	+	-	+	+	-	
11	-	+	-	-	-	+	+	+	-	+	+	
12	-	-	-	-	-	-	-	-	-	-	-	

Cuadro 2: Filas generadoras para la construcción de diseños PB.

PB <sub>12</sub>	$n = 12$	+ - + - - + + + - +
PB <sub>20</sub>	$n = 20$	+ + - - + + + + - + - + - - - + + -
PB <sub>24</sub>	$n = 24$	+ + + + + - + - + + - - + + - - + - - - -

fila una columna hacia la derecha. El nivel (signo) de la última columna pasa a la primer columna de la siguiente fila. Agotados los corrimientos, los niveles de la última fila son todos negativos.

Si bien, los arreglos PB tienen la propiedad de considerar muchos factores con pocas corridas, en ocasiones, su interpretación no es posible utilizando las herramientas de análisis más comunes: las gráficas en papel normal de los efectos (Daniel, 1959), o el procedimiento de Lenth (1989). Esto se debe a la compleja estructura *alias* de los diseños. El efecto principal de cada factor (columna) está parcialmente confundido con todas las interacciones de segundo orden donde el mismo efecto no participa. De igual manera, columnas no asignadas a factores, están parcialmente confundidas con todas las interacciones de segundo orden. Por ejemplo, si en un PB<sub>12</sub> como el mostrado en la Tabla 1, asignáramos solamente las primeras 5 columnas a los factores  $A, \dots, E$ , el patrón de confusiones es:

$$\begin{aligned} l_A &\rightarrow A + \frac{1}{3}(-BC + \underline{BD} + BE - CD - CE - \underline{DE}) \\ l_K &\rightarrow \frac{1}{3}(-AB - AC + AD - AE - BC - BD - BE + CD + CE - DE) \end{aligned}$$

donde  $l_A$  y  $l_K$  representan los contrastes asociados a las columnas correspondientes.

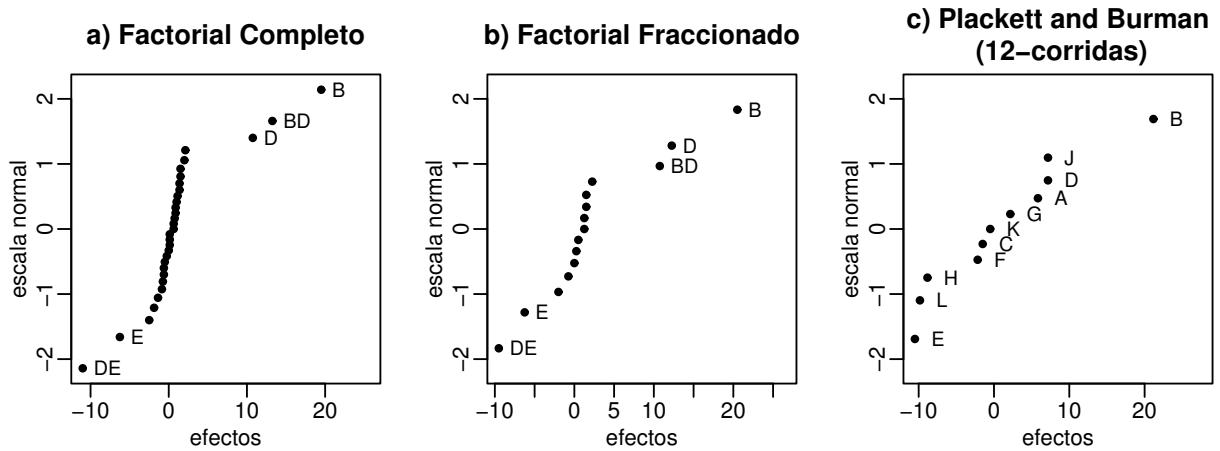


Figura 1: Gráfica normal de efectos de los experimentos ortogonales FC  $2^5$ , FF  $2^{5-1}$  y PB<sub>12</sub>.

### 3. Ejemplo

Para su ilustración, consideremos el ejemplo de un reactor, presentado en Box, Hunter y Hunter (2005, p. 260). La respuesta es la producción (%) del reactor, y como factores experimentales se consideraron: *A*, la tasa de alimentación (lt./min); *B*, la cantidad de catalizador (%); *C*, la tasa de agitación (rpm); *D*, la temperatura ( $^{\circ}C$ ); y *E*, la concentración (%).

La Figura 1.a muestra la gráfica normal de los efectos estimados del diseño factorial completo  $2^5$ . En la gráfica se identifican como posibles efectos activos *B*, *D* y *E*, además de las interacciones de segundo orden *BD* y *DE*. Si en su lugar se hubiese corrido solamente la mitad del diseño, digamos, el factorial fraccionado  $2^{5-1}$ , definido por las corridas 2, 3, 5, 8, 9, 12, 14, 15, 17, 20, 22, 23, 26, 27, 29, 32, la correspondiente gráfica se muestra en el panel b de la Figura 1. Note que los mismos efectos se identifican como activos.

Por otro lado, si se hubiese corrido el diseño PB de 12 ensayos solamente, definido por las corridas 1, 3, 6, 12, 14, 15, 18, 23, 24, 25, 28, 29, difícilmente se identifica uno de los efectos (*B*) como activos, como se muestra en la Figura 1.c. La disminución del número de ensayos ha dificultado la identificación de los efectos activos. Esto es debido a la compleja estructura alias de los experimentos Plackett-Burman. El procedimiento Bayesiano discutido en esta nota permitiría aún la identificación de los factores activos.

## 4. Selección Bayesiana de factores

La selección Bayesiana de factores desarrollada por Box y Meyer (1993), al igual que los procedimientos de Daniel y Lenth, se basa en el supuesto de *pocos factores activos*. El procedimiento se puede resumir de la siguiente manera:

Se supone que la respuesta  $y$  es aproximada razonablemente por el modelo lineal  $y = X\beta + \epsilon$ , donde  $\epsilon \sim n(0, \sigma^2 I_n)$ , y solamente una pequeña proporción ( $\pi$ ) de los factores  $\beta_u$  es activa. Si  $\beta_u$  es activo,  $\beta_u \sim n(0, \gamma^2 \sigma^2)$ , de otra forma,  $\beta_u \sim n(0, \sigma^2)$ .

Estamos interesados en seleccionar los factores activos de una colección de  $k$  posibles  $F_1, \dots, F_k$ . Para esto, suponga que una combinación particular de factores  $f_i$ , ( $0 \leq f_i \leq k$ ), es activa. Sea  $M_i$  el modelo que incluye  $f_i$  efectos principales, efectos de interacciones de segundo orden, etc. Esto es,  $y = X^{(i)}\beta^{(i)} + \epsilon$ , donde  $X^{(i)}$  es la matriz de diseño de los factores  $f_i$ , y  $\beta^{(i)}$  el vector de los coeficientes del correspondiente modelo.

El procedimiento considera las asignaciones *a priori* de la proporción de factores activos,  $p_j = P\{F_j \text{ es activo}\} = \pi$ ,  $j = 1, \dots, k$ , y por lo tanto,  $P_i = P\{M_i \text{ es correcto}\} = \pi^{f_i}(1 - \pi)^{k-f_i}$ ,  $0 \leq i \leq 2^k$ . De igual forma, se proveé valores al factor de inflación  $\gamma$ . Asignaciones *a priori* de  $0.2 \leq \pi \leq 0.3$  y  $2 \leq \gamma \leq 3$  son útiles en la práctica.

Box y Meyer (1993) muestran que la probabilidad *posterior* del modelo  $M_i$  está dada por

$$P(M_i|y) = C \left( \frac{\pi}{1 - \pi} \right)^{f_i} \gamma^{-t_i} \frac{|X'_0 X_0|^{1/2}}{|X'_i X_i|^{1/2}} \left( \frac{S(\hat{\beta}_i) + \hat{\beta}'_i \Gamma_i \hat{\beta}_i}{S(\hat{\beta}_0)} \right)^{-(n-1)/2}$$

donde  $C$  es la constante de integración,  $t_i$  es el número de coeficientes en el modelo,  $\Gamma_i = \gamma^{-2} \text{diag}\{0, 1, \dots, 1\}$  y  $S(\hat{\beta}_i)$  la suma de cuadrados del modelo  $M_i$ . En la expresión anterior, el primer factor depende de los valores *a priori* de los parámetros, el segundo factor favorece los modelos con pocos términos (parsimoniosos), y el tercero es un factor de bondad de ajuste. Finalmente, las correspondientes probabilidades marginales de los factores, son:

$$p_{j|y} = P\{F_j \text{ es activo} | y\} = \sum_{i: F_j \in M_i} P\{M_i | y\}.$$

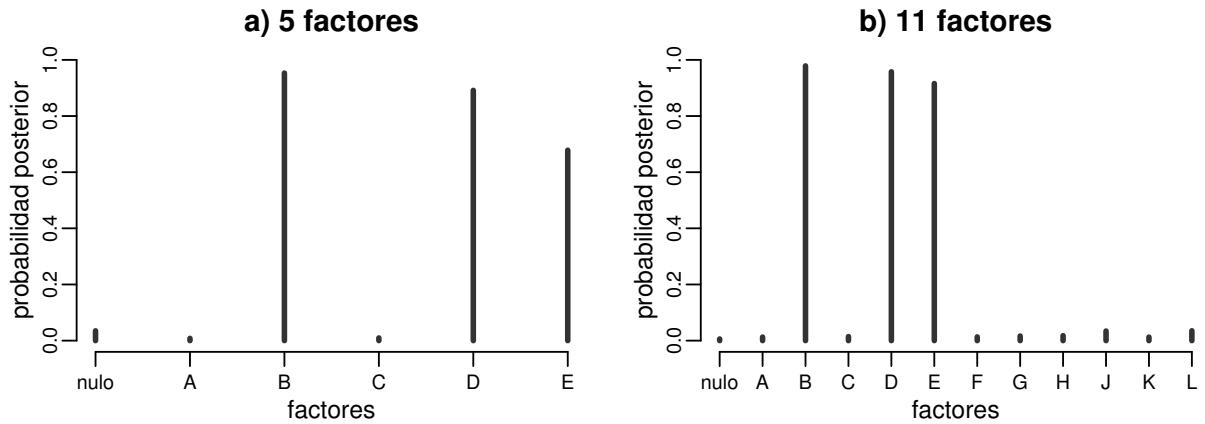


Figura 2: Probabilidades marginales *posteriores* de factores obtenidas del experimento PB<sub>12</sub>.

## 5. Ejemplo

Si aplicamos el procedimiento al ejemplo del reactor. Hay  $k = 5$  factores. Usando  $\pi = 0.25$  y  $\gamma = 2.5$ , mediante el paquete BsMD (Barrios, 2004) se ajustaron 32 modelos considerando interacciones hasta de tercer orden. Las probabilidades marginales de los factores obtenidas se muestran en la Figura 2.a. Note como la actividad de los factores  $B$ ,  $D$  y  $E$  vuelve a hacerse evidente. El análisis puede extender para identificar también la interacciones activas aprovechando que los diseños PB de 12, 20 y 24 corridas son de proyectividad 3 (Box y Tyssedal, 1996). La Figura 2.b muestra las probabilidades marginales calculadas considerando las 11 columnas del diseño PB-12 asociadas a factores. En este caso, se ajustaron 2048 modelos. Nuevamente, los únicos factores identificados como activos son  $B$ ,  $D$  y  $E$ .

## Referencias

Barrios, E. (2004) BsMD Bayesian Screening and MD Follow-up Designs.

URL: <http://cran.at.r-project.org/src/contrib/PACKAGES.html>

Box, G. E. P., Hunter J. S., and Hunter, W. C. (2005). *Statistics for Experimenters* (Second Edition) New York: Wiley.

Box, G.E.P. and Meyer, R. D. (1993). Finding the Active Factors in Fractionated Screening Experiments, *Journal of Quality Technology*, **25**, 94–105.

Box, G.E.P. and Tyssedal, J. (1996). Projective Properties of Certain Orthogonal Arrays, *Biometrika*, **83**, 950–955.

Daniel, C. (1959). Use of Half-Normal Plots in Interpreting Factorial Two-Levels Experiments, *Technometrics* **1**, 311–341.

Lenth, R. V. (1989). Quick and Easy Analysis of Unreplicated Factorials. *Technometrics*. **31**, 469–473.

Plackett, R. L. and Burman, J. P. (1946). The Design of Optimum Multifactorial Experiments, *Biometrika*, **33**, 305–325.

# Modelos lineales generalizados en el contexto de diseño robusto

**Guadalupe Bocanegra Aguilar<sup>1</sup>**

*Universidad Juárez Autónoma de Tabasco*

**Jorge Domínguez Domínguez<sup>2</sup>**

*Centro de Investigación en Matemáticas*

## 1. Introducción

El estudio de variación fuera de línea en la fase de desarrollo de productos y procesos recientemente ha recibido mucha atención. Esto es debido a la influencia de Taguchi (1987). Al contrario de la situación de diseño experimental clásico, qué trata con factores que son fijados durante el experimento y bajo las condiciones de procesamiento (factores de diseño), Taguchi introdujo los factores de ruido en el experimento. Su inclusión permite un estudio sistemático del efecto del proceso de variación en la calidad del producto y proceso.

Las investigaciones de Shoemaker *et al.* (1991) muestran que un arreglo combinado, en el cual los factores de diseño y de ruido son tratados simétricamente, es con frecuencia más flexible y económico que el arreglo cruzado propuesto por Taguchi.

A causa de la equivalencia de los factores de diseño y de ruido en la etapa experimental, la Metodología de Superficie de Respuesta (MSR), conocida de la estadística clásica, se puede aplicar a Diseño Robusto. En el contexto de MSR Vining and Myers (1990) proponen un enfoque de respuesta dual, minimizando la variabilidad mientras que se lleva la media al target. Por otra parte, Lee y Nelder (1998) sugieren aplicar modelos lineales generalizados (GLM) al diseño robusto y Grego (1993) usa GLM para la varianza de la respuesta de experimentos clásicos con réplicas.

En este trabajo se expondrá la utilidad de los Modelos Lineales Generalizados (GLM) para estudiar el diseño del parámetro en el contexto del diseño robusto. Los GLM se aplican

---

<sup>1</sup>032A10004@dacb.ujat.mx

<sup>2</sup>jorge@cimat.mx

para modelar de manera conjunta la media y la varianza, que resulta un objetivo vital en la estrategia del diseño robusto. Trabajar con los GLM permite que las variables de respuesta no necesariamente sigan una distribución normal sino que tengan una distribución de probabilidad más amplia llamada la familia exponencial. Mientras que con la regresión lineal la distribución normal juega un papel central.

La meta de este trabajo es describir diferentes métodos de optimización, considerando de manera específica la estrategia experimental conocida como doble arreglo ortogonal. Se hará una comparación de los métodos usados evaluando sus características potenciales en la práctica de la experimentación industrial.

## 2. Modelos Lineales Generalizados en el diseño robusto

En este apartado se describen el método desarrollado por Engel-Huele (1996) y el de Modelos Lineales Generalizados para obtener el diseño robusto del parámetro. El primero es una extensión de la superficie de respuesta enfocada al diseño robusto para adoptar a los modelos lineales generalizados.

### Modelo en el caso Engel-Huele

Considere  $k$  factores de diseño por  $X_1, \dots, X_k$  y  $q$  factores de ruido por  $Z_1, \dots, Z_q$ . Sea  $x_i$  que denota los niveles de los factores de diseño en la  $i - \text{ésima}$  corrida. El  $i - \text{ésimo}$  renglón de la matriz de diseño se denota por  $g'(x_i)$ ,  $i = 1, \dots, m$ . Además de los efectos lineales, puede también contener los términos cuadráticos y interacciones entre los factores de diseño. Los niveles de los factores de ruido en la  $i - \text{ésima}$  corrida se denota por  $z'_i$ . Suponga que los datos  $Y = (y_1, \dots, y_m)'$  son el resultado del experimento de un arreglo que combina los factores de control con los de ruido y se asume que la varianza es independiente de la media. Por lo tanto, puede a veces ser necesario transformar los datos antes de ajustar el modelo.

El objetivo es obtener un modelo para  $E(y_i)$ , la media del proceso y  $\text{var}(y_i)$ , la varianza del proceso. Ya que durante el experimento el vector de respuesta  $Y$  se observa condicionalmente en los niveles de los factores de ruido, así:  $\mu_i = E(y_i|z_i)$  y  $\sigma_i^2 = \text{var}(y_i|z_i)$ . Ahora, se propone un modelo de regresión heteroscedástico (varianza no constante) para el problema del diseño

robusto:  $y_i = \mu_i + \varepsilon_i$ , con  $\varepsilon_i \sim N(0, \sigma_i^2)$ . La media condicional  $\mu_i$  es lineal en los ajustes de los factores de ruido en la  $i$ -ésima corrida, está dada por:

$$\mu_i = \beta_0 + g'(x_i)\beta + z'_i\delta + g'(x_i)\Lambda z'_i. \quad (1)$$

Aquí  $\beta_0$  es una constante y  $\beta$  &  $\delta$  son dos vectores de parámetros.  $\Lambda$  es una matriz de parámetros que contiene los coeficientes de regresión de la interacción entre los factores de diseño y de ruido. La varianza condicional de la respuesta se modela como:

$$\sigma_i^2 = \exp(u'_i\gamma), \quad (2)$$

con  $u'_i$  que contiene los niveles de los factores de diseño  $p - 1 \leq k$  en la  $i$ -ésima corrida. El primer elemento de  $u_i$  es 1. El vector  $\gamma$  es un vector de parámetros de dimensiones  $p \times 1$ .

(1) es el modelo de respuesta condicional:  $\mu_i = E(y_i|z_i)$ . Este señala que el valor esperado de la respuesta, depende de los factores de ruido. La función de varianza (2) es un modelo para la varianza del error alrededor del modelo de la respuesta condicional para la media. Myers *et al.* (1992) asumen que  $\sigma_i^2$  es constante.

Una vez que se ajustan los modelos condicionales, se tiene que los factores de ruido varían aleatoriamente. Así, las variables aleatorias son sustituidas por los factores de ruido que aparecen en los modelos condicionales. Las superficies de respuesta para la media del proceso  $E(y_i)$  y la varianza del proceso  $var(y_i)$  son obtenidas usando las relaciones siguientes:

$$E(y_i) = E[E(y_i|z_i)], \quad (3)$$

$$var(y_i) = var[E(y_i|z_i)] + E[var(y_i|z_i)]. \quad (4)$$

Las operaciones entre corchetes son calculadas sobre la distribución del vector de los factores de ruido  $Z = (Z_1, \dots, Z_q)'$ . A lo largo de ésta sección se supone que  $E(Z)$  y  $var(Z)$  son conocidas. En el proceso de estimación se obtienen los modelos ajustados para la media y varianza respectivamente son:  $\hat{\mu}_i = \hat{\beta}_0 + g'(x_i)\hat{\beta} + z'_i\hat{\delta} + g'(x_i)Mz_i$ ,  $\sigma_i^2 = \exp(u'_i\hat{\gamma})$ .

Bajo el supuesto de que el vector de los factores de ruido  $Z$  tiene esperanza 0 y matriz de varianza-covarianza  $\Phi = \varphi_{ij}$ , con  $\varphi_{ij} = Cov(Z_i, Z_j)$ . Entonces, la superficie de respuesta

estimada para la media y la varianza del proceso está dada por:

$$\widehat{E(y_i)} = \widehat{\beta}_0 + g'(x_i)\widehat{\beta}, \quad (5)$$

$$\widehat{Var(y_i)} = \left[ \widehat{\delta} + M'g(x_i) \right]' \Phi \left[ \widehat{\delta} + M'g(x_i) \right] + \exp(u_i'\widehat{\gamma}). \quad (6)$$

## Modelos Lineales Generalizados

Los GLM fueron introducidos por Nelder and Wedderburn (1972). Estos están definido por tres componentes:

1. Un modelo para la varianza  $Var(y_{ij}) = \phi_{ij}V(\mu_{ij})$ ;
2. Una función liga  $\eta = g(\mu)$ , y
3. Un modelo de regresión para el predictor lineal  $\eta = x'\beta$ .

En ocasiones una sola transformación de los datos puede fallar y por lo tanto modificar las propiedades necesarias para un análisis. Con los GLM, la identificación de la relación media y varianza y la selección de la escala en la cual los efectos deben ser medidos se deben realizar por separado, con el fin de superar los defectos del enfoque de la transformación de los datos. También los GLM proporcionan una extensión del enfoque de superficie de respuesta.

El objetivo de ésta sección es modelar de manera conjunta la media y la dispersión, es decir, usar un GLM tanto para la media como para la dispersión.

## Modelos para el Diseño Robusto

El objetivo de utilizar el doble arreglo ortogonal es minimizar la varianza mientras que se lleva la media al target. Taguchi utilizó expresiones denominadas señales razón a ruido SNR, tal como:  $10 \log \left( \sum \frac{\bar{y}_i^2}{s_i^2} \right)$ , para alcanzar este objetivo. Esta situación expresada en términos de un modelo lineal generalizado GLM es como sigue:

$$Var(y_{ij}|z_j) = \phi_{ij}V(\mu_{ij}), \quad (7)$$

con  $V(\mu_{ij}) = \mu_{ij}^2$  y  $\mu_{ij} = E(y_{ij}|z_j)$ , donde  $\mu_{ij} = \mu_i$ ,  $\phi_{ij} = \phi_i$ ,  $V(\mu_{ij})$  es la función

de varianza, y  $\phi_{ij}$  es el parámetro de dispersión. Las suposiciones en las ecuaciones ( $\mu_{ij} = \mu_i$ ,  $\phi_{ij} = \phi_i$ ) implican que los factores de ruido no tienen ningún efecto en la media o la dispersión.

Las suposiciones en las ecuaciones (7) y ( $\mu_{ij} = \mu_i$ ,  $\phi_{ij} = \phi_i$ ) implican que  $E(y_{ij}|z_j) = \mu_i$  y  $Var(y_{ij}|z_j) = \phi_i\mu_i^2$ , dando  $var(y_{ij}) = \phi_i\mu_i^2$ , debido a que  $var[E(y_{ij}|z_j)] = 0$ . Así, la varianza relativa,  $\phi_i = \frac{var(y_{ij})}{[E(y_{ij})]^2}$ . La SNR anterior tiene sólo sentido cuando  $E(y_{ij}|z_j) = \mu_i$  y  $Var(y_{ij}|z_j) = \phi_i\mu_i^2$ . Por lo que, Taguchi minimiza la varianza minimizando la varianza condicional  $Var(y_{ij}|z_j)$ , tema discutido por Lee and Nelder (2003).

## 2.1. Ejemplo

En un proceso de soldadura de circuitos integrados se propone un experimento para determinar las condiciones que produzcan el número mínimo de defectos por millón de uniones en un soldador. Los factores de control y sus niveles se muestran en la Tabla 1.

Factores de Control	-1	1	Factores de Ruido	-1	1
A: Temperatura de soldado ( $^{\circ}F$ )	480	510	O: Soldado	5	-5
B: Velocidad de la cinta ( $ft/min$ )	7.2	10	P: Cinta	0.2	-0.2
C: Densidad del flujo	0.9	1.0	Q: Tipo de ensamble	1	2
D: Temperatura de precalentado ( $^{\circ}F$ )	150	200			
E: Amplitud de onda ( $in.$ )	0.5	0.6			

Tabla 1. Factores de control y factores de ruido

En este experimento dos de los tres factores de ruido, Tabla 1, presentan dificultades para fijarlos en sus niveles nominales durante el proceso. Estos factores son la temperatura de soldado y la velocidad de la cinta, se sabe que la temperatura varía  $\pm 5^{\circ}F$  del valor nominal y que la velocidad de la cinta varía  $\pm 0.2 ft/min$ , esta variabilidad es transmitida a la respuesta. El esquema experimental se llevó a cabo en un doble arreglo ortogonal, donde el arreglo interno corresponde a un diseño  $2^{5-2}$  y el externo a un  $2^{3-1}$ . La matriz cruzada con

los valores de la respuesta se muestran en la Tabla 2.

Arreglo Interno					Arreglo Externo (O,P,Q)			
A	B	C	D	E	(-1, -1, -1)	(1, 1, -1)	(1, -1, 1)	(-1, 1, 1)
1	1	1	-1	-1	194	197	193	275
1	1	-1	1	1	136	136	132	136
1	-1	1	-1	1	185	261	264	264
1	-1	-1	1	-1	47	125	127	42
-1	1	1	1	-1	295	216	204	293
-1	1	-1	-1	1	234	159	231	157
-1	-1	1	1	1	328	326	247	322
-1	-1	-1	-1	-1	186	187	105	104

Tabla 2. Matriz cruzada para el proceso de soldadura

## Análisis

Primero realizando un análisis gráfico de la razón señal a ruido para el caso "mientras más pequeña es mejor", es decir,  $SNR = -10 \log \left( \frac{1}{4} \sum_{i=1}^4 y_i^2 \right)$  y la media de los datos contra los factores. Los resultados son:

1. Al analizar la variable  $SNR = -10 \log \left( \frac{1}{4} \sum_{i=1}^4 y_i^2 \right)$ , se tiene que los factores  $A$ ,  $C$ , y  $E$  afectan de manera significativa a la  $SNR$ . En ese sentido estos factores influyen sobre la variación de las condiciones que dan un mínimo número de defectos de soldado. De estos resultados se recomienda utilizar el factor  $A$  en su nivel alto,  $C$  y  $E$  en su nivel bajo.
2. De manera análoga al realizar el análisis para la media, se observa que los factores  $A$ ,  $C$ , y  $E$  también tienen efecto significativo sobre la media.

Por lo tanto, las condiciones robustas del proceso se dan con  $A$  en su nivel alto y  $C$  y  $E$  en su nivel bajo y los factores  $B$  y  $D$  se pueden ubicar en las condiciones más económicas, presumiblemente en su nivel bajo.

## Engel & Huele

Aplicando el procedimiento de estimación de Engel & Huele, se tiene:

Los modelos condicionales finales están dados por

$$\hat{\mu} = 197.13 - 25.78A + 56.88C + 21.36E + 14.22AO$$

$$-14.56BO - 14.55CO + 14.26CP + 15.09AQ$$

$$\hat{\sigma}^2 = \exp(5.92 - 0.78C)$$

Por lo que, se obtienen los modelos para  $\widehat{E(y)}$  y  $\widehat{Var(y)}$  sustituyendo las variables aleatorias por los factores de ruido.

Finalmente se aplica el método de mínimos cuadrados para el diseño de parámetro. Se supone que los factores de ruido  $O$ ,  $P$ , y  $Q$  tienen media 0 y desviación estándar  $\sigma_o$ ,  $\sigma_p$ , y  $\sigma_q$ , respectivamente. Aplicando la fórmula (5) resulta que el modelo estimado de la media del proceso está dada por

$$\widehat{E(y)} = 197.13 - 25.78A + 56.88C + 21.36E$$

El modelo estimado para la varianza de proceso es

$$\begin{aligned}\widehat{Var(y)} &= (14.22A - 14.56B - 14.55C)^2 \sigma_o^2 + (14.26C)^2 \sigma_p^2 \\ &\quad + (15.09A)^2 \sigma_q^2 + \exp(5.92 - 0.78C)\end{aligned}$$

Nuestro planteamiento de optimización es minimizar  $\widehat{Var(y)}$  sujeto a que la  $\widehat{E(y)}$  sea mínima. Así que mediante el proceso de optimización se obtienen los resultados, éstos se describen en la Tabla 3a. Considerando que  $\sigma_o^2$ ,  $\sigma_p^2$ , y  $\sigma_q^2$  es igual a 1, respectivamente.

## GLMs

Ahora, usando modelos lineales generalizados. Nuestro modelo es:

$$Var(y_{ij}) = \phi_{ij}V(\mu_{ij})$$

con  $V(\mu_{ij}) = \mu_{ij}^2$ . Obteniendo lo siguiente:

$$\log \mu_{ij} = 5.29 - 0.07A - 0.004B + 0.28C + 0.09E - 0.08O - 0.05BO$$

$$\log \phi_{ij} = -4.36 - 0.60A - 1.06C - 1.41O - 1.93AO$$

Todos los efectos son significativos con un nivel de significancia de 5 % tanto para el modelo de la media como la dispersión, excepto para el efecto principal  $B$  en el modelo de la media. El objetivo de este estudio es minimizar el número de defectos de soldadura, es decir, minimizar tanto la media como la dispersión. Se quita el efecto de ruido  $O$  en el modelo de la dispersión, esto se logra despejando  $A = -1.41/1.93 = -0.73$ , la cual corresponde a la temperatura de soldado de  $484.05^{\circ}F$ . Luego, hay que quitar el efecto de ruido  $O$  en el modelo de la media, despejando  $B = -0.08/0.05 = -1.6$ , la cual corresponde a la velocidad de la cinta de  $6.36 ft/min$ . De esta manera, se logra un diseño el cual es insensible a los factores de ruido. Primero minimizando la media se tiene que:  $\mu_i^0 = 144.681$  donde  $A = -0.73$ ,  $B = -1.6$ ,  $C = -1$  y  $E = -1$ .

Luego, la varianza es  $Var(y_{ij}) = \exp(-4.36 + 0.60(0.73) + 1.06)(145.11)^2 = 1197.886$ . Los resultados obtenidos se muestran en la Tabla 3b.

Óptimo	Resultados	Óptimo	Resultados
$(A, B, C, D, E)$	$(1, 1, -1, 0, -1)$	$(A, B, C, D, E)$	$(-0.73, -1.6, -1, 0, -1)$
Media estimada	93.11	Media estimada	144.681
Varianza estimada	1447.82	Varianza estimada	1197.886
CME	10117.40		
Tabla 3a. Engel-Huele		Tabla 3b. Modelos Lineales Generalizados	

Comparando los resultados obtenidos de los métodos usados para este ejemplo, se observa que con los modelos lineales generalizados se logra la mínima varianza de este proceso.

### 3. Discusión

El método de Engel-Huele da mucha importancia en ajustar los modelos condicionales. Además describen su metodología en la que los modelos de superficie de respuesta para la media y la varianza del proceso son desarrollados simultáneamente de manera que agregan una contribución a la modelación lineal generalizada a los resultados existentes al modelar la varianza. Los MLG son una extensión del problema de superficie de respuesta y resultan ser eficientes en la estimación del diseño de parámetro.

## Referencias

- Engel, J. and Huele, A. F. (1996). A Generalized Linear Modelling Approach to Robust Design. *Technometrics* **38**, 365-373.
- Grego, J. M. (1993). Generalized Linear Models and Process Variation. *Journal of Quality Technology*, **25**, 288-295.
- Lee, Y. and Nelder, J. A. (1998). Generalized Linear Models for the Analysis of Quality Improvement Experiments. *The Canadian Journal of Statistics*, **26**, 95-105.
- Lee, Y. and Nelder, J. A. (2003). Robust Design via Generalized Linear Models. *Journal of Quality Technology*, **35**, 2-12.
- Shoemaker, A. C., Tsui, K. L., and Wu, C. F. J. (1991). Economical Experimentation Methods for Robust Design. *Technometrics*, **33**, 415-427.
- Taguchi, G. (1987). *Systems of Experimental Design*. UNIPUB/Kraus International Publications, New York.
- Vining, G. G., and Myers, R. H. (1990). Combining Taguchi and Response Surface Philosophies: A Dual Response Approach. *Journal of Quality Technology*, **22**, 38-45.



# Acerca de la contrucción de modelos AR(1) utilizando densidades predictivas que emergen de la estadística Bayesiana no paramétrica

Alberto Contreras-Cristán<sup>1</sup>

*Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas*

Ramsés H. Mena-Chávez<sup>2</sup>

*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas*

Stephen G. Walker<sup>3</sup>

*University of Kent*

## 1. Procesos estacionarios

Sea  $T = \{0, 1, \dots\}$ , un proceso a tiempo discreto  $\{X_t, t \in T\}$  es *estrictamente estacionario* si dados  $t_1, t_2, \dots, t_n \in T$  y  $h$  tal que  $t_{1+h}, \dots, t_{n+h} \in T$ , se tiene que

$$\{X_{t_1}, X_{t_2}, \dots, X_{t_n}\} \stackrel{d}{=} \{X_{t_{1+h}}, X_{t_{2+h}}, \dots, X_{t_{n+h}}\},$$

para cada  $n = 1, 2, \dots$ , donde  $\stackrel{d}{=}$  denota igualdad en distribución.

En el caso de procesos *Markovianos* (homogéneos) la estacionariedad estricta se puede asegurar con la existencia de una medida de probabilidad invariante, es decir,  $Q(\cdot)$  tal que

$$Q(A) = \int_{\mathbb{R}} P_t(x, A) Q(dx), \tag{1}$$

donde  $P_t(x, A) = \mathbb{P}(X_t \in A | X_{t-1} = x)$ , para  $A$  un subconjunto de Borel en  $\mathbb{R}$ .

En Pitt *et al.* (2002), se plantea la construcción de modelos de series de tiempo estrictamente estacionarios como sigue (para esta descripción asumimos el caso de variables aleatorias continuas y la existencia de las densidades que se mencionan):

---

<sup>1</sup>alberto@sigma.iimas.unam.mx

<sup>2</sup>ramses@sigma.iimas.unam.mx

<sup>3</sup>S.G.Walker@kent.ac.uk

Supóngase que se requiere un modelo con densidad marginal  $p(x)$ .

Se propone un modelo paramétrico de la densidad condicional  $p(y | x)$ .

La transición que gobierna el modelo AR(1) se obtiene calculando

$$p(x | x_{t-1}) = \int p(x | y) p(y | x_{t-1}) dy, \quad (2)$$

donde

$$p(x | y) = \frac{p(y | x) p(x)}{\int p(y | x) p(x) dx}.$$

Se puede demostrar que la densidad  $p(\cdot)$  es invariante para la transición  $p(x | x_{t-1})$ , es decir

$$p(x) = \int p(x | x_{t-1}) p(x_{t-1}) dx_{t-1}.$$

La variable aleatoria  $y$  juega un papel de variable aleatoria latente. En un contexto Bayesiano, podemos pensar en  $y$  como el parámetro. De esta manera, la densidad de transición tiene la interpretación

$$p(x | x_{t-1}) = E\{p(x | y) | x_{t-1}\}, \quad (3)$$

donde la esperanza es con respecto a la *posterior Bayesiana*  $p(y | x_{t-1})$ .

Interpretamos a  $p(x_{t-1} | y)$  y a  $p(y)$  como las correspondientes *verosimilitud* y *distribución* inicial. La estructura de dependencia en el modelo, se impone mediante la elección de la familia paramétrica asumida para  $p(y | x_{t-1})$ .

Mena y Walker (2005) proponen flexibilizar la forma de asignar esta dependencia entre  $X_t$  y  $X_{t-1}$  usando modelos no paramétricos para  $p(y | x_{t-1})$ . La forma de llevar a cabo esto es utilizando modelos no paramétricos para  $p(y | x_{t-1})$ , en otras palabras, en el esquema anterior se reemplaza a la variable latente  $y$  por una densidad (distribución) de probabilidad aleatoria  $F$ .

Denotemos a  $\mathcal{B}(\mathbb{R})$  los subconjuntos de Borel en  $\mathbb{R}$ . Podemos pensar en una distribución aleatoria  $F$  como un proceso estocástico  $\{F(A) | A \in \mathcal{B}(\mathbb{R})\}$ , indexado por estos subconjuntos. Este proceso estocástico está caracterizado por su ley o distribución  $\Pi$  que es una

medida de probabilidad sobre un espacio de distribuciones de probabilidad. De esta forma nos referiremos en forma indistinta al proceso (la distribución aleatoria) o a la ley que lo caracteriza ( $\Pi$ ).

Sea  $\Pi$  una medida de probabilidad definida sobre un espacio de distribuciones de probabilidad en  $\mathbb{R}$  que denotamos  $(\mathcal{F}, \mathcal{B}_{\mathcal{F}})$ , pensamos a  $\Pi$  como una distribución inicial sobre  $\mathcal{F}$ . Dada la observación  $X$ , requerimos de una distribución posterior para construir  $\mathbb{P}(X_t \leq x_t | X_{t-1} = x_{t-1})$  como en (3).

Para un conjunto  $A \in \mathcal{B}_{\mathcal{F}}$ , consideremos la distribución conjunta

$$\mathbb{P}\{X \leq x; F \in A\} = \mathbb{E}_{\Pi}\{F(x)1_A(F)\} = \int_A F(x)\Pi(dF).$$

donde  $F \in \mathcal{F}$ .

Notemos que para  $A = \mathcal{F}$  obtenemos  $\mathbb{P}\{X \leq x\} = \mathbb{E}_{\Pi}\{F(x)\}$ . Utilizando el teorema de Bayes la distribución posterior requerida se puede calcular como

$$\mathbb{P}(F \in A | X \leq x) = \frac{\mathbb{E}_{\Pi}\{F(x)1_A(F)\}}{\mathbb{E}_{\Pi}\{F(x)\}}.$$

Asumiendo las condiciones necesarias para su existencia, podemos definir  $\Pi_x(A) = \mathbb{P}\{F \in A | X = x\}$  como la probabilidad posterior de la distribución aleatoria  $F$ , dado que la observación  $X$  vale  $x$ . Con esta posterior podemos actuar como en (3) y calcular

$$p(X_t \leq x_t | X_{t-1} = x_{t-1}) = \int F(x_t)\Pi_{x_{t-1}}(dF). \quad (4)$$

En este trabajo se plantea el problema de obtener modelos de series de tiempo estrictamente estacionarias y con distribuciones marginales fijas, utilizando medidas de probabilidad  $\Pi$  definidas sobre espacios de distribuciones de probabilidad discretas.

## 2. La medida de probabilidad aleatoria beta-Stacy

Walker y Muliere (1997) presentan al proceso beta-Stacy como una generalización del proceso de Dirichlet el cual es una medida de probabilidad aleatoria muy utilizada en el contexto Bayesiano no-paramétrico, véase por ejemplo, Ferguson (1973). Denotemos a  $\mathcal{F}$  como el espacio de funciones de distribución con soporte en el conjunto  $\mathbb{N} = \{0, 1, \dots\}$ .

Sean  $B(\alpha, \beta)$  la función beta y  $\mathcal{C}(\alpha, \beta, \xi)$  la distribución beta-Stacy, cuya función de densidad de probabilidades está dada por

$$f(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1} \frac{(\xi - y)^{\beta-1}}{\xi^{\alpha+\beta-1}} 1_{(0, \xi)}(y).$$

Consideremos la sucesión de variables aleatorias positivas  $\{Y_k \mid k \in \mathbb{N}\}$  dada por

$$\begin{aligned} Y_1 &\sim \mathcal{C}(\alpha_1, \beta_1, 1), \\ Y_2|Y_1 &\sim \mathcal{C}(\alpha_2, \beta_2, 1 - Y_1), \\ &\vdots \\ Y_k|Y_{k-1}, \dots, Y_1 &\sim \mathcal{C}(\alpha_k, \beta_k, 1 - F_{k-1}), \end{aligned}$$

donde  $\{\alpha_k\}$  and  $\{\beta_k\}$  son sucesiones en  $\mathbb{R}_+$  y  $F_k = \sum_{j=1}^k Y_j$ .

Walker y Muliere (1997) probaron que

- Con probabilidad 1, el proceso beta-Stacy a tiempo discreto

$$F(k) = \begin{cases} 0 & \text{if } k = 0, \\ \sum_{j \leq k} Y_j & \text{if } k > 0 \end{cases}$$

toma valores en  $\mathcal{F}$ .

- El tamaño (aleatorio) del brinco de  $F$  en  $k$  está dado por  $Y_k$ .
- Para cada  $m = 1, 2, \dots$  la distribución conjunta del vector  $(Y_1, \dots, Y_m)$  es la distribución de Dirichlet generalizada.

Es posible centrar el proceso beta-Stacy en una función de distribución específica  $Q \in \mathcal{F}$ . Con este fin, definimos  $A_k = \{0, 1, \dots, k\}$ ,  $q(k) = Q(k) - Q(k-1)$  y tomamos

$$\alpha_k = c_k q(k) \text{ and } \beta_k = c_k \{1 - Q(A_k)\} = c_k \left\{ 1 - \sum_{l=0}^k q(l) \right\}, \quad (5)$$

donde  $\{c_k\}$  es una sucesión de números reales positivos. Esta forma de lograr que  $\mathbb{E}_\Pi\{F(x)\} = Q(x)$  fue sugerida por Walker y Muliere (1997).

Dada una muestra  $X_1, \dots, X_n$  de una distribución discreta desconocida  $F$ , Walker y Muliere (1997) probaron que condicionalmente al hecho de que  $F$  proviene de un proceso beta-Stacy a tiempo discreto, es decir  $\Pi = \text{beta-Stacy}(Q, \{c_k\})$ , entonces al aplicar la ecuación (4), la distribución predictiva basada en una observación es

$$\mathbb{P}(X_t = x_t \mid X_{t-1} = x_{t-1}) = h(x_t \mid x_{t-1}) \prod_{\xi < x_t} \{1 - h(\xi \mid x_{t-1})\}, \quad (6)$$

donde

$$\begin{aligned} h(\xi \mid x_{t-1}) &= \frac{\alpha_\xi}{\alpha_\xi + \beta_\xi} 1(\xi > x_{t-1}) + \frac{\alpha_\xi + 1}{\alpha_\xi + \beta_\xi + 1} 1(\xi = x_{t-1}) \\ &+ \frac{\alpha_\xi}{\alpha_\xi + \beta_\xi + 1} 1(\xi < x_{t-1}). \end{aligned}$$

En particular, si usamos la elección de  $\{\alpha_k\}$  y  $\{\beta_k\}$  dada en (5), la densidad de transición (6) tendrá a la distribución  $Q$  como distribución invariante (véase la ecuación 1). En otras palabras la estructura de dependencia dada por  $\Pi$ , nos permite construir un proceso de tipo autoregresivo de orden 1, cuyas variables toman valores discretos, estrictamente estacionario, con distribución marginal (estacionaria)  $Q$  y con función de transición (6). Llamaremos a este modelo el *proceso beta-Stacy AR(1)*.

### 3. Inferencia estadística

Dado un conjunto de observaciones, para ajustar el modelo beta-Stacy se requiere de estimar la sucesión  $\{c_k\}$  y los parámetros que sean desconocidos en la distribución estacionaria  $Q$ .

Para el caso de que  $Q$  tenga soporte finito, digamos en el conjunto  $\{x_0, x_1, \dots, x_l\}$ , tendremos que estimar un número finito  $\{c_0, c_1, \dots, c_l\}$  de valores.

Para ilustrar estas ideas, consideremos el proceso AR(1) con valores discretos ideado por Al-Osh y Alzaid (1991). Este es un proceso  $\{X_t\}$  estacionario con distribución marginal  $Q = \text{Binomial}(N, p)$  y para un entero positivo  $M \leq N$  evoluciona de acuerdo a la ecuación

$$X_t = A_t(X_{t-1}) + Z_t, \quad (7)$$

donde

$\{Z_t\}$  son variables aleatorias, independientes y con distribución  $\text{Binomial}(N - M, \rho)$  para  $\rho = \frac{M}{N}$ .

$A_t$  es un operador aleatorio que queda definido usando la distribución condicional de  $X_1 | X_1 + X_2 = x$  y que satisface

$$A_t(X) | X = x \sim \text{Hipergeométrica}(N, x, M).$$

La función de auto-correlación de este proceso está dada por  $\phi(k) = \rho^k$ ,  $k = 0, 1, \dots$

Para el ejemplo de estimación que sigue, simulamos 500 muestras, cada una de tamaño  $W = 200$  del proceso de Al-Osh y Alzaid con parámetros  $N = 5$ ,  $p = 0.5$ ,  $M = 1$ , de donde la autocorrelación a un lag vale  $\rho = 0.2$ . Con el fin de ajustar el modelo beta-Stacy a cada una de estas muestras, notemos que la función de verosimilitud del proceso beta-Stacy se puede escribir como un producto de  $W - 1$  factores dados como en (??) (6) por un factor que es la distribución del proceso al tiempo inicial, es decir,  $q$ . Asumiendo que al ver una trayectoria de los datos simulados proponemos un modelo  $\text{Binomial}(5, p)$  para  $Q$ , estaremos optimizando una función de verosimilitud con parámetros  $\{c_0, \dots, c_5\}$  y  $p$ .

Con el propósito de presentar un resumen de los resultados de la estimación para cada una de las 500 muestras, denotemos la  $i$ -ésima muestra simulada por  $\{X_1^i, \dots, X_{200}^i\}$ . Supongase que para esta muestra medimos el siguiente error cuadrático (correspondiente a estimar la

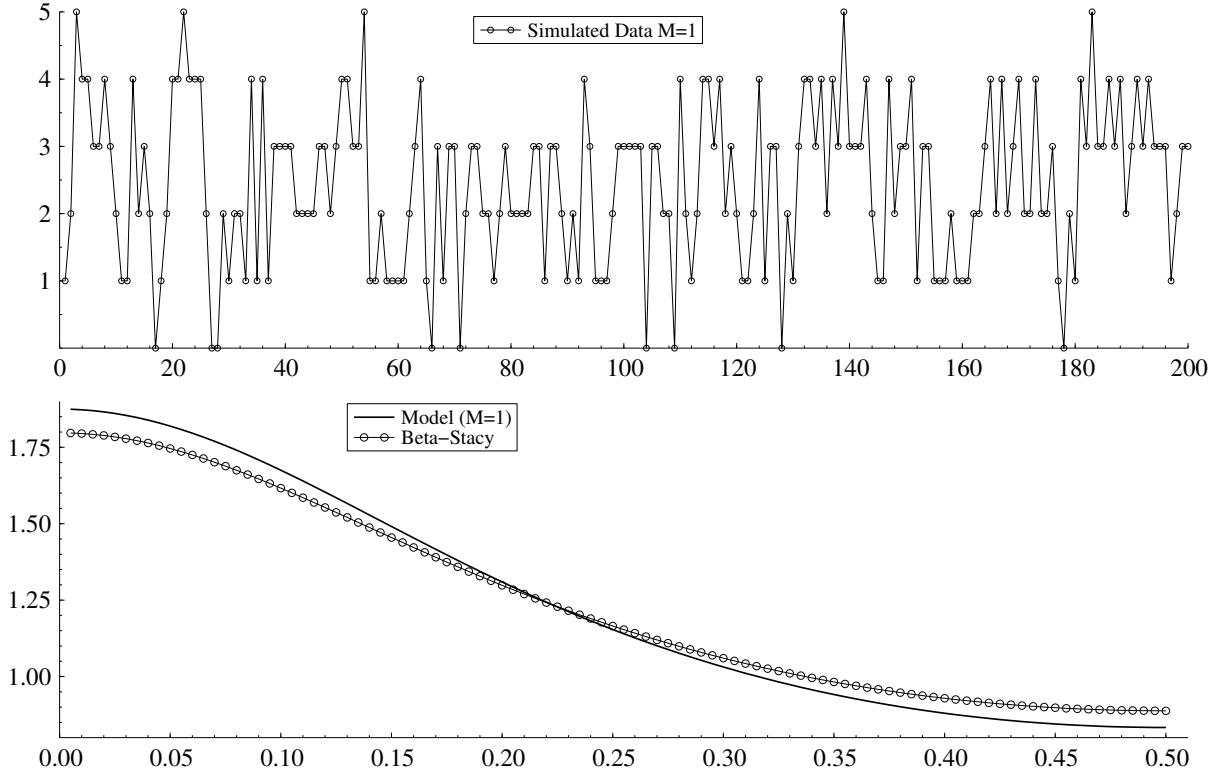


Figura 1: Arriba: Datos simulados del proceso de Al-Osh y Alzaid. Abajo: Estimación de la densidad espectral

densidad espectral del proceso

$$RMSE(i) = \left\{ \frac{1}{W/2 + 1} \sum_{k=0}^{W/2} \left( \hat{S}^i(k/W) - S(k/W) \right)^2 \right\}^{1/2},$$

donde  $W = 200$ ,  $S$  es la verdadera densidad espectral del modelo Al-Osh y Alzaid y  $\hat{S}^i$  es la densidad espectral estimada usando la  $i$ -ésima muestra. Para calcular  $\hat{S}^i$  se encuentran  $\{\hat{c}_0, \dots, \hat{c}_5\}$  y  $\hat{p}$  que maximizan la verosimilitud de la  $i$ -ésima muestra, se utilizan estos estimadores para evaluar las probabilidades de transición (6) y con estas se estima la sucesión de autocorrelación  $\{\hat{\phi}(k)\}$ , por último se calcula la transformada de Fourier de  $\{\hat{\phi}(k)\}$ . La Figura 1 reporta los resultados de la estimación correspondiente a RMSE localizado en el cuantil 50 % de la población de 500 RMSE estimados.

## 4. Comentarios finales

En este trabajo se presenta una alternativa para construir modelos de series de tiempo de tipo autoregresivo de orden 1, estrictamente estacionarias y con valores discretos. Creemos que los métodos no-paramétricos Bayesianos nos permiten flexibilizar las posibles estructuras de dependencia entre las variables del proceso. Para el ejemplo conocido de la serie de tiempo con distribución marginal Binomial sugerido por Al-Osh y Alzaid, nuestro modelo es capaz de captar en buena forma la estructura de correlación.

## Referencias

- Al-Osh, M. A. and Alzaid, A. A. (1991). Binomial autoregressive moving average models. *Communications in Statistics - Stochastic Models*, **7**, 261–282.
- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.
- Mena, R. H. and Walker, S. G. (2005). Stationary autoregressive models via a Bayesian nonparametric approach. *Journal of Time Series Analysis*, **26**-6, 789–805.
- Pitt, M. K., Chatfield, C., and Walker, S. G. (2002). Constructing first order autoregressive models via latent processes. *Scandinavian Journal of Statistics*, **29**, 657–663.
- Walden, A. T., Percival, D. B., and McCoy, E. J. (1998). Spectrum estimation by wavelet thresholding of multitaper estimators. *IEEE Transactions on Signal Processing*, **46**(12), 3153–3165.
- Walker, S. G. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Annals of Statistics*, **25**, 1762–1780.

# Alcances y limitaciones de S-PLUS para la generación de árboles de escenarios por simulación

Meliza Contreras González<sup>1</sup>

*Facultad de Ciencias de la Computación. Benemérita Universidad Autónoma de Puebla*

Gladys Linares Fleites<sup>2</sup>

*Departamento de Investigaciones en Ciencias Agrícolas. Benemérita Universidad Autónoma de Puebla*

## 1. Introducción

En la toma de decisiones son muy utilizados los árboles de escenario, al igual que en el cómputo científico por su desempeño.

Por lo que el motivo de este trabajo es generar árboles de escenario mediante S-PLUS y se organiza como sigue. En la sección 2 se definen los árboles de escenario. En la sección 3 se describen las ventajas de S-PLUS. En la sección 4 se plantean dos metodologías para generar estos árboles cuyos resultados se presentan en la sección 5. Por último se establecen las conclusiones en la sección 6.

## 2. Arboles de escenario en problemas de decisión

Una alternativa para crear un abanico de posibilidades es el generador de escenarios, que es un proceso de construcción de distribuciones discretas para variables de decisión. La estructura resultante de su aplicación es un árbol de escenarios (ver Figura 1.a), donde cada nodo representa el valor de cada variable de decisión en un instante de tiempo, las ramas representan las probabilidades de ocurrencia de éstos valores. Las ramas que inician en el nodo raíz y culminan hasta los nodos hojas se les llama escenarios (ver Figura 1.b), cuya

---

<sup>1</sup>mel\_22281@hotmail.com

<sup>2</sup>gladys.linares@icbuap.buap.mx

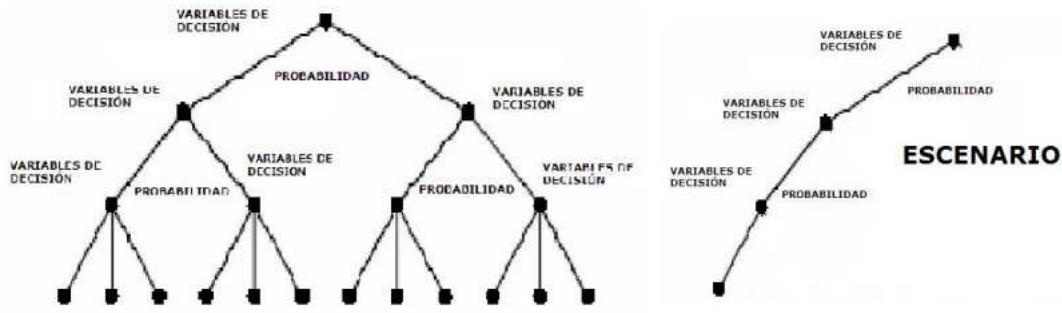


Figura 1: a)Arbol de escenario b)Escenario

probabilidad de ocurrencia se determina por el producto de cada rama a lo largo del camino (Kaut y Wallace, 2003).

### 3. Importancia de S-PLUS

En las últimas décadas las demandas en la resolución de problemas estadísticos han ocasionado el desarrollo de plataformas especializadas en esta área. Dentro de ellas está S-PLUS (Venables y Ripley 1999), que realiza cálculos exhaustivos con gran rapidez y proporciona opciones desde la generación de muestras hasta métodos para realizar análisis multivariado, así como gráficos potentes. Como herramienta adicional, incluye el lenguaje de programación S, cuya finalidad es la expresión de modelos estadísticos ajustables a las necesidades del investigador.

### 4. Generación de escenarios

Así, este trabajo centra su atención en la economía, porque ésta presenta procesos altamente inestables. Para la toma de decisiones financieras es necesario resolver programas estocásticos que presenten un abanico de posibilidades de los pros y contras de invertir en determinadas opciones de cartera (Kaut y Wallace 2003 y Kouwenberg 2001), finalidad de los árboles de escenarios. Entonces el proyecto se conforma en dos etapas: la primera es la generación de vectores aleatorios lognormales, en este caso se utilizó un generador de números aleatorios

Variables	BIMBO	BANCOMER	BANORTE	KOF	PENOLE	TELME	GMODEL	TELECOM
Media	0.02922	0.02836	0.09580	0.22412	0.01234	0.25654	0.03856	0.009340
Desviación Estándar	0.00478	0.00306	0.00375	0.00306	0.00028	0.00114	0.00631	0.001529

Figura 2: Medias y desviaciones estándar de los principales índices de la BMV

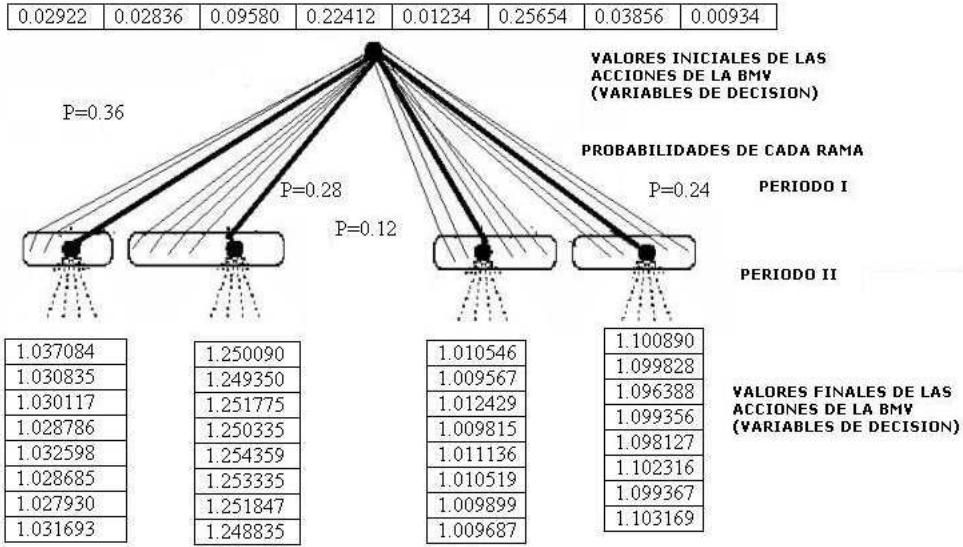


Figura 3: Árbol de escenarios aplicando conglomerados

de S-PLUS que inicialmente tomó como medias y desviaciones estándar, las provenientes de las acciones de la bolsa mexicana de valores (véase Reynoso 2004). La Figura 2 ejemplifica lo anterior. Una vez que se cuenta con la muestra, la segunda etapa es aplicar técnicas de análisis multivariado, debido a que se tiene más de una variable de respuesta de cada observación.

## 5. Resultados

Al utilizar S-PLUS se facilitó la obtención de valores de las variables de decisión, así como de las probabilidades de ocurrencia de cada escenario. A continuación se detalla la aplicación de los métodos de análisis multivariado propuestos: conglomerados y componentes principales conglomerados.

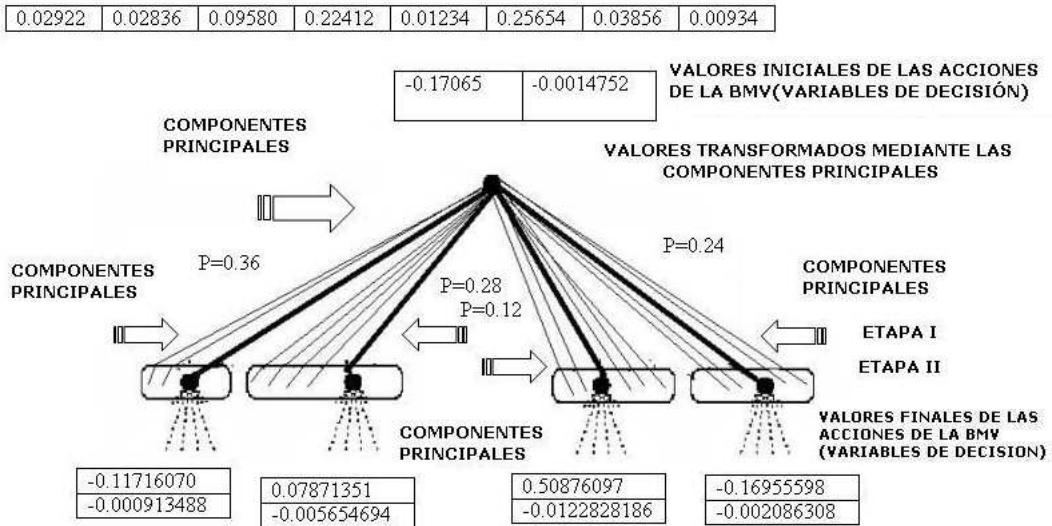


Figura 4: Arbol de escenarios aplicando componentes principales-conglomerados

## 5.1. Conglomerados

En este caso el lenguaje S, proporciona distintas opciones para encontrar los dendogramas asociados a la muestra, se realizaron pruebas teniendo dendogramas mediante los métodos: connect, compact y average. Así para cada periodo del árbol de escenario, a partir de las muestras se aplicaron las técnicas de Conglomerados, obteniendo dendogramas para cada nodo representante del valor de la variable de decisión en el tiempo, obteniendo un árbol de escenario (ver Figura 3).

## 5.2. Componentes principales-conglomerados

En este método, inicialmente a la muestra proporcionada se le aplica la técnica de componentes principales. Entonces, considerando sólo las dos primeras componentes se caracteriza adecuadamente a la población y se evita la manipulación excesiva de datos. Así se siguen los mismos pasos del método de conglomerados, sólo que previamente se construye una nueva muestra a partir de las dos primeras componentes principales obtenidas anteriormente, por lo que resulta un árbol de escenario de menor complejidad (ver Figura 4).

## 6. Conclusiones

Definitivamente con los resultados del proyecto se han encontrado importantes ventajas al aplicar el análisis multivariado en los árboles de escenario. Con la utilización de los conglomerados y los centroides resultantes de los distintos grupos, se determinan los nodos, correspondientes a la etapa que se encuentre simulando, mientras que con el método componentes principales-conglomerados desde el principio se reduce la dispersión de los datos enfocándose en las variables más significativas para el fenómeno, reduciendo la dimensión del árbol, ocasionando que las búsquedas de caminos idóneos no sean exhaustivas. Por otro lado la plataforma S-PLUS, facilitó en tiempo y esfuerzo el camino de la realización del árbol de escenario.

## Referencias

- Gülpinar, N. Rustem, B. Settergren, R. (2004). Optimisation and simulation approaches to scenario tree generation. *Journal of Economics Dynamics and Control*, **28**, 1291-1315.
- Hoyland, K. y Wallace, S. W. (2001). Generating scenario trees for multi stage decision problems. *Management Science*, **47**, 295-307.
- Kaut, M. y Wallace, S. W. (2003). Evaluation of scenario-generation methods for stochastic programming. *Stochastic Programming E-Print Series*, 1-14.
- Kouwenberg, R. (2001). Scenario generation and stochastic programming models for asset liability management. *European Journal of Operation Research*, **134**, 279-292.
- Reynoso, A. (2004). Opening up a Securities Market: Mexico's New Push for Liberalization 2003-2004, Mexican Stock Exchange. En SCID Latin America Conference, pp 1-61.
- Venables, W.N. y Ripley, B.D. (1999). *Modern Applied Statistics with S-PLUS*. Nueva York: Springer - Verlag.



# Optimización conjunta de diseño de parámetro y diseño de tolerancia

Jorge Domínguez Domínguez<sup>1</sup>

*Centro de Investigación en Matemáticas*

Susana Pérez Santos<sup>2</sup>

*Universidad Juárez Autónoma de Tabasco*

## 1. Introducción

La Mejora continua es sin duda alguna un objetivo importante que es deseable alcanzar en diversas actividades y proyectos en diferentes áreas del conocimiento, principalmente si se refiere a procesos industriales. El desarrollo y aplicación de métodos para la optimización permiten determinar las condiciones en un proceso, así establecer las condiciones de la mejora continua.

El procedimiento de calidad fuera de línea se utiliza para alcanzar la efectividad en la estrategia de mejorar la calidad en un producto, y que a la vez redunde en el impacto económico. El método propuesto por Taguchi (1996) combina las técnicas de diseño experimental con las consideraciones de pérdida de calidad. El método para diseñar el producto consiste en tres etapas. Ellas son el diseño de sistema, el diseño de parámetro y el diseño de tolerancias. El primero involucra el desarrollo de un prototipo que cumpla con los requisitos establecidos por el cliente.

El producto y el procedimiento operacional están influenciados por el **diseño de parámetro** esto es por los factores que son controlados por el experimentador y los factores de ruido. Así el diseño de parámetro trata de encontrar los niveles de los factores que minimice los efectos de los factores de ruido. Esto es, las condiciones del diseño de parámetro para un producto o proceso determinan que la varianza de la variable de respuesta  $Y$ (característica del producto) debe ser mínima y que la media esta cerca de un valor objetivo  $M$ . Durante el diseño de parámetro, se supone que materiales y componentes de bajo nivel permite

---

<sup>1</sup>jorge@cimat.mx

<sup>2</sup>032A10005@dacb.ujat.mx

una tolerancia para los factores de control mientras minimiza la sensitividad del ruido y la variación de la calidad. El **diseño de tolerancia** se aplica si la reducción en la varianza de la calidad generada por el diseño de parámetro es insuficiente. En el diseño de tolerancia surge la necesidad de negociar entre reducir la varianza de la calidad o incrementar el costo de la manufactura.

En este trabajo, se presentan el planteamiento para la optimización de un proceso en relación al diseño robusto. En esa dirección se describe la estrategia para optimizar de manera compuesta el diseño de parámetro y diseño de tolerancias, en este caso la variable de respuesta siguen una distribución de probabilidad normal, y se describe de manera breve la estrategia global de optimización.

## 2. Diseño de producto

El método de calidad fuera de línea tiene como finalidad mejorar la calidad de un producto o su procedimiento operacional, y éste tiene una interpretación en el costo, expresada en la pérdida financiera debido a la desviación de  $M$ . Es decir:  $P(Y(x)) = k(Y(x) - M)^2$ , el valor esperado de esta función se indica por:

$$C(x) = E(P(Y(x))) = k(\sigma^2 + (\mu - M)^2), \quad (1)$$

donde  $k$  es el costo de calidad asociado a una unidad producida,  $\mu$  y  $\sigma^2$  son la media y la varianza de  $Y$  respectivamente. El interés es incorporar la expresión anterior a la situación en la que se presenta la necesidad del diseño de tolerancias.

La figura 1 ilustra que el desarrollo de tolerancias toma un camino que inicia y termina con las expectativas y requerimientos de los clientes. El diseño y las tolerancias de manufactura se derivan de las tolerancias de los clientes. El diseño y capacidad del proceso de manufactura ( $C_p$ ) deben ser optimizados para minimizar pérdida del cliente.

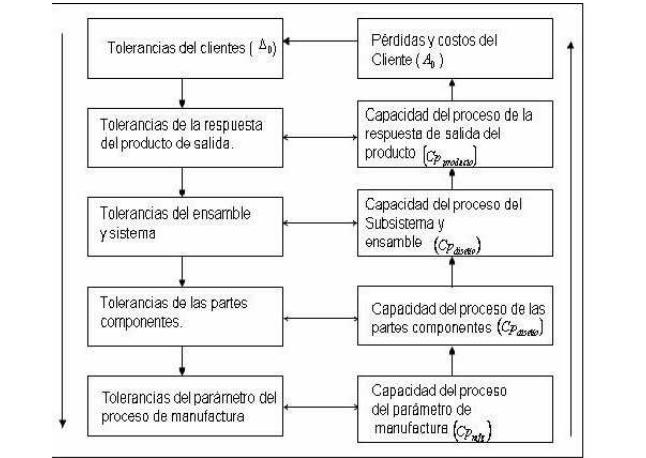


Figura 1. Camino para el desarrollo de tolerancias

Figura 1. Camino para el desarrollo de tolerancia

Las tolerancias deben ser vistas en el contexto de cómo se restringe la transformación eficiente de energía, durante la función del diseño. Como un proceso de manufactura puede producir rendimiento sobre el valor objetivo  $M$ , es también una función de la eficiencia y la estabilidad de la energía. Esto se vuelve importante cuando se estudia la variabilidad de salida asociada con un proceso de manufactura.

Las tolerancias están definidas fundamentalmente como *límites de fronteras*. En el mundo de ingeniería de calidad de Taguchi, las tolerancias están económicamente establecidas operando ventanas de variabilidad funcional para optimizar conjuntos de puntos de factores de control y limitar pérdidas de los clientes. Típicamente los límites están basados en el rendimiento que posee la calidad  $3\sigma$ .

$y$  es la respuesta de calidad, la cual se expresa en función de los factores  $x$  con el modelo de regresión :  $y(x) = \beta_0 + x^t\beta + x^tBx + \varepsilon$ , con  $\beta_0$  constante,  $\beta = (\beta_1, \dots, \beta_k)$  un vector de parámetros  $B = (\beta_{11}, \dots, \beta_{1k}, \beta_{k1}, \dots, \beta_{kk})$  matriz de parámetros de segundo orden, y  $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ . Además considere que los factores  $x_i$  siguen una distribución normal  $N(x_i, \sigma_i^2)$ . En esta dirección, la finalidad del diseño de parámetros es encontrar los valores  $x_i$  que dado  $\sigma_i^2$  minimicen (1). Se indicó con anterioridad que es común definir la tolerancia  $t_i$  de  $x_i$  como:  $t_i = 3\sigma_i$ . De acuerdo a esta planteamiento la función  $C(x)$  de la expresión (1) se puede

expresar como una función de  $x_{i0}$  y  $t_i$ . Entonces (1) se puede reescribir como  $C(x_0, t)$ . Así el diseño de parámetro se replantea como el proceso de optimizar  $C(x_0, t)$  para  $x_o$  con  $t$  fija.

## 2.1. Diseño compuesto del diseño de parámetro y tolerancia

Como se ha indicado el diseño de tolerancia trae consigo costos de calidad asociados, esto da lugar a plantear una nueva función objetivo que considere el costo total esta se expresa por:

$$H(x_0, t) = C(x_0, t) + C(t) = k(\sigma^2 + (\mu - M)^2) + \sum_i^n C(t_i), \quad (2)$$

El procedimiento para reducir la varianza desempeña un papel importante en la práctica, sin embargo lograr esta meta usualmente causa un incremento en el costo de manufactura porque requiere: procedimientos operacionalmente más precisos, mejores medios de operación y técnicos mejor entrenados. En ese sentido disminuir la varianza de un proceso implica tolerancias más estrechas.

El desarrollo de en serie de Taylor permite estudiar el efecto de las componentes de tolerancia en la variabilidad de  $y$ . Considere la función:

$$y = g(x_1, \dots, x_n) = g(x_{10}, \dots, x_{n0}) + \sum_{i=1}^n d_i[x_i - x_{i0}] + \sum_{i=1}^n d_{ii}[x_i - x_{i0}]^2 \quad (3)$$

donde  $m < n$ ,  $d_i = \partial g / \partial x_i$  y  $d_{ii} = \frac{1}{2} \partial^2 g / \partial x_i^2$  evaluadas en  $x_i = x_{i0}$ . La relación entre  $\sigma$ ,  $\mu$  y  $x_{i0}$  y  $t_i$  se obtiene a partir de la expresión (3). La varianza  $\sigma_y^2$  se obtiene mediante el término de primer orden de la serie de Taylor y se sustituye  $t_i = 3\sigma_i$ :

$$\sigma_y^2 = var(y) \approx \sum_{i=1}^n d_i^2 \sigma_i^2 = \frac{1}{9} \sum_{i=1}^n d_i^2 t_i^2 \quad (4)$$

Así  $\mu - M$  en (2) se obtiene utilizando término de segundo orden de la serie de Taylor y se sustituye  $t_i = 3\sigma_i$ :

$$\mu - M \approx g(x_{10}, \dots, x_{n0}) - M + \sum_{i=1}^n d_{ii} \sigma_i^2 = g(x_{10}, \dots, x_{n0}) - M + \frac{1}{9} \sum_{i=1}^n d_i^2 t_i^2 \quad (5)$$

Con las expresiones (4) y (5) se obtiene la función objetivo en términos de  $x_0$  y  $t$ :

$$H(x_0, t) = k \left( \frac{1}{9} \sum_{i=1}^n d_i^2 t_i^2 + (g(x_{10}, \dots, x_{n0}) - M + \frac{1}{9} \sum_{i=1}^n d_{ii}^2 t_i^2)^2 \right) + \sum_i^n C(t_i), \quad (6)$$

En la estrategia experimental la función  $g(x_{10}, \dots, x_{n0})$  representa al modelo  $\hat{y}(x_0) = b_0 + \sum b_i x_{0i} + \sum \sum b_{ij} x_{0i} x_{0j}$  en términos de las variables originales. Derivando esta expresión y sustituyendo en la ecuación (6), se obtiene:

$$H(x_0, t) = k \left( \frac{1}{9} \sum_{i=1}^n d_i^2 t_i^2 + (\hat{y}(x_0) - M + \frac{1}{9} \sum_{i=1}^n b_{ii} t_i^2)^2 \right) + \sum_i^n C(t_i), \quad (7)$$

donde  $d_i = b_i + 2b_{ii}x_{i0} + \sum_{j=i+1}^n b_{ij}x_{j0} + \sum_{j=i+1}^{i-1} b_{ij}x_{j0}$ .

Finalmente el modelo de optimización es:

$$\text{Minimizar } H(x_0, t)$$

$$\begin{aligned} & \text{Sujeto a } t_i^I \leq t_i \leq t_i^S \\ & \sum_{i=1}^n d_i^2 t_i^2 \leq T_{\max} \\ & x \in R : \text{Región experimental.} \end{aligned} \quad (8)$$

donde  $T_{\max}$  la tolerancia máxima permitida, y  $t_i^I, t_i^S$  los límites inferior y superior respectivamente para la tolerancia.

## Estrategia para integrar los costos

Considere que se ha realizado un experimento, en cada tratamiento a los factores se les asigna su tolerancia como se muestra en la tabla de abajo, se proponen diferentes tolerancias  $t_{ij}$  para cada factor ( $i = 1, \dots, n$  y  $j$  niveles de tolerancia) por consiguiente en cada tratamiento se genera un costo. De esta manera se tienen dos respuestas, la correspondiente a la variable de interés ( $y$ ) y el costo. Luego en el proceso de optimización se propone una  $t_{ij}^*$  y se sustituye en el modelo (7), se establece la función costo  $C(t)$ , esta se obtiene ajustando los valores resultantes de costo, por ejemplo un modelo para ajustar una de estas funciones puede ser el recíproco esto es:  $C(t) = c_0 + c_1 t^{-1}$ . Finalmente se optimiza el modelo (8).

Factor	Tolerancia	Costo asociado
$x_1$	$t_{1j}$	$C_{1j}$
.	.	.
.	.	.
.	.	.
$x_n$	$t_{nj}$	$C_{nj}$

## Referencias

- Bisgaard, S. and Amkenman, B. (1995). Analytic Parameter Desing. *Quality Engineering*, **8**(1), 75-91.
- Myers, R. H. and Montgomery, D.C. (2002). *Response Surface Methodology: Process and Product Optimization Using Desinged Experiments*. John Wiley and Sons, New York, NY.
- Park, H.S. (1996). *Robust Design and Analysis for Quality Engineering*. Chapman and Hall, London, UK.
- Romano, D., Varetto, M. and Vicario G. (2004). Multiresponse Robust Design: A General Framework Based on Combined Array. *Journal of Quality Technology* **36**(1), 27-37.

# Evaluación numérica de funciones de distribución multivariadas

J. Armando Domínguez Molina<sup>1</sup>

*Facultad de Matemáticas, Universidad de Guanajuato*

Alonso Núñez Páez<sup>2</sup>

*Escuela de Ciencias Físico-Matemáticas de la Universidad Autónoma de Sinaloa.*

## 1. Introducción

Sea  $f(\mathbf{x}; \boldsymbol{\theta})$  una función de densidad del vector aleatorio  $\mathbf{x} \in \mathbb{R}^s$ ,  $s \geq 1$ . Sea  $A \in \mathbb{R}^s$  un conjunto Borel medible dado. El problema que nos concierne<sup>3</sup> es la evaluación o aproximación numérica de la integral de probabilidad

$$I(\boldsymbol{\theta}; A) = \Pr(\mathbf{X} \in A) = \int_A f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}, \quad (1)$$

con énfasis al caso en que  $A$  es un hiper-rectángulo de  $\mathbb{R}^s$ . Este es un problema importante, ya que el valor numérico de  $I(\boldsymbol{\theta}; A)$  es útil en la obtención pruebas de hipótesis, clasificación y análisis discriminante, entre otros.

Recientemente se han generado importantes avances en la teoría y aplicación de las distribuciones elípticas sesgadas (skew-elliptical). Genton (2004) contiene una reseña de una gran cantidad de avances, recolectando resultados teóricos y aplicaciones previas dispersas a través de la literatura. Azzalini (2005) contiene una visión general introductoria de la teoría de distribuciones relacionada con las distribución normal-sesgada y las distribuciones elípticas sesgadas. La evaluación numérica de (1) es de particular importancia en la teoría de distribuciones sesgada, debido a que para la evaluación de la función de densidad de los miembros de las familias elípticas sesgadas se requiere del cálculo de funciones de distribución de variables aleatorias relacionadas con la familia.

---

<sup>1</sup>[dominguez@quijote.ugto.mx](mailto:dominguez@quijote.ugto.mx)

<sup>2</sup>[alonso@uas.uasnet.mx](mailto:alonso@uas.uasnet.mx)

<sup>3</sup>Trabajo realizado con apoyo del proyecto 04-16-K117-028 de CONCyTEG

También puede requerirse de la evaluación numérica de integrales múltiples dónde el integrando no necesariamente es una función de densidad. Con el fin de dar una exposición general también consideraremos la teoría de integración múltiple para integrandos que no necesariamente son funciones de densidad. De esta manera podremos calcular integrales múltiples en otras áreas de la ciencia en las que también se requieren de manera natural. Ejemplos de lo anterior se dan en finanzas: Paskov y Traub (1995) y física: Keister (1996) y Papageorgiou y Traub (1997).

## 2. Aproximación a Integrales múltiples

Para aproximar la integral de una función  $f$  con valores en los reales definida sobre el hipercubo unitario  $U^s = [0, 1]^s$ , dada por

$$I_s(f) = \int_{U^s} f(\mathbf{x}) d\mathbf{x}. \quad (2)$$

La manera común de resolver este problema consiste en la elección de un conjunto de puntos  $P_n = \{x_0, x_1, \dots, x_{n-1}\}$  con los cuales se toma el promedio de  $f$  sobre  $P_n$ ,

$$Q_s(P_n)f = \frac{1}{n} \sum_{k=0}^{n-1} f(\mathbf{x}_k), \quad (3)$$

como una aproximación de  $I_s(f)$ . Para  $s$  grande, los métodos clásicos basados en el producto cartesiano de una regla unidimensional (*e.g.*, regla del trapecio, regla de Simpson, cuadratura gausiana, entre otras) no son prácticos debido a *la maldición de la dimensionalidad*: la complejidad computacional crece exponencialmente con la dimensión. Las integrales en dimensiones altas se aproximan comúnmente por métodos Monte Carlo o quasi-Monte Carlo (QMC), también conocidos como métodos de la Teoría de Números. En los métodos Monte Carlo se toma  $P_n = \{\mathbf{u}_k : \mathbf{u}_k \sim \text{Uniforme}(0, 1)^s\}$ , en contraparte en los métodos quasi-Monte Carlo  $P_n$  es un conjunto de puntos elegidos de manera determinista de dos grandes clases: (i) Látices de integración y (ii) redes digitales y sucesiones de baja discrepancia. En particular, para las reglas de Korobov o métodos de buenas látices de puntos (*glp* por sus siglas en inglés) introducidas por Korobov en 1959 e independientemente por Hlawka en 1962, se toma  $P_n = \{\{k\mathbf{z}/n\} : k = 0, 1, \dots, n - 1\}$ , donde  $\mathbf{z} \in \mathbb{Z}^s$  es el vector generatriz sin

factores comunes con  $n$  elegido cuidadosamente y  $\{x\}$  denota la parte fraccionaria, es decir,  $\{x\} = x \bmod 1$ . Sloan (1985) y Sloan y Kachoyan (1987) generalizaron las reglas de Korobov introduciendo más de un vector generatriz. La precisión de las reglas de látices dependen del vector generatriz  $y$  de la clase de funciones. La clase de funciones de Korobov  $E_{\alpha,s}(c)$  es el conjunto de funciones en  $L_1(U^s)$ , que sus coeficientes de Fourier satisfacen

$$\hat{f}(\mathbf{h}) \leq \frac{c}{(\bar{h}_1 \cdots \bar{h}_s)^\alpha},$$

donde  $\mathbf{h} = (h_1, \dots, h_s)$  con  $h_j$  entero y

$$\hat{f}(\mathbf{h}) = \int_{U^s} \exp(-2\pi i \mathbf{h} \cdot \mathbf{x}) f(\mathbf{x}) d\mathbf{x},$$

con  $\mathbf{h} \cdot \mathbf{x} = h_1 x_1 + \cdots + h_s x_s$  y  $\bar{h}_j = \max(1, |h_j|)$ . La medida clásica de las reglas de Korobov es el *error de cuadratura del peor caso*  $P_\alpha$  en la clase de funciones de Korobov  $E_{\alpha,s}(1)$  dado por:

$$P_\alpha := \sup \{|I_s(f) - Q_s(P_n)f| : f \in E_{\alpha,s}(1)\} = \sum_{\mathbf{h} \neq \mathbf{0}, \mathbf{h} \cdot \mathbf{z} \equiv 0 \pmod{n}} \frac{1}{(\bar{h}_1 \cdots \bar{h}_s)^\alpha}.$$

Para una función  $f \in E_{\alpha,s}(1)$ , una cota de error es

$$\left| I_s(f) - \frac{1}{n} \sum_{k=0}^{n-1} f\left(\left\{\frac{k\mathbf{z}}{n}\right\}\right) \right| \leq c P_\alpha.$$

Se sabe que existe un vector generatriz  $\mathbf{z}$  tal que  $P_\alpha = O(n^{-\alpha} (\log n)^\beta)$  para alguna  $\beta$  de orden  $s$ . Una manera de elegir el vector generatriz  $\mathbf{z}$  es tomar el que minimiza  $P_\alpha$ . En Keast (1973), Fang y Wang (1994) y Hua y Wang (1981) se proporcionan tablas de valores que minimizan  $P_\alpha$  y otras cantidades.

En general, no es posible obtener una estimación útil para las reglas Korobov al repetir los cálculos con más puntos de cuadratura, ya que los errores tienden a fluctuar erráticamente. La estrategia para estimar el error propuesta por Cranley y Patterson (1976), es usar una regla de Korobov fija, para utilizarla en la forma desplazada

$$Q_s(P_n, \mathbf{c}) f = \frac{1}{n} \sum_{k=0}^{n-1} f\left(\left\{\frac{k}{n}\mathbf{z} + \mathbf{c}\right\}\right), \quad (4)$$

donde  $\mathbf{c}$  es un vector aleatorio de desplazamiento. Al calcular (4) para diferentes elecciones aleatorias de  $\mathbf{c}$ , es posible obtener una estimación de la integral e intervalos de confianza para la magnitud del error. Los errores de cuadratura en  $Q_s(P_n)$  y  $Q_s(P_n, \mathbf{c})$  dependen de los mismos coeficientes de Fourier, por lo que cualquier miembro de la familia de reglas de Korobov desplazadas no es mejor que cualquier otro para un integrando general  $f$ . En la terminología de Cranley y Patterson (1976) es una familia estocástica de reglas de cuadratura. Supongamos que el vector aleatorio  $\mathbf{c}$  se elige de una distribución uniforme multivariada sobre  $C^s$  así que cada componente de  $\mathbf{c}$  es independiente uniformemente distribuida sobre  $[0, 1]$ . Esto nos lleva a que la familia estocástica de reglas de Korobov desplazadas dada en (4) es un estimador insesgado de  $I_s(f)$ ; es decir, el centro de gravedad de la distribución de  $Q(P_n, \mathbf{c})$  es  $I_s(f)$ .

### 3. Evaluación numérica de funciones de distribución

Un problema que surge en muchas aplicaciones estadísticas es la evaluación de la función de distribución normal multivariada

$$F(\mathbf{a}, \mathbf{b}) = \frac{1}{|\Sigma|^{1/2} (2\pi)^s} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_s}^{b_s} e^{-\frac{1}{2}\mathbf{x}^t \Sigma^{-1} \mathbf{x}} d\mathbf{x}, \quad (5)$$

donde  $\mathbf{x} = (x_1, x_2, \dots, x_m)^t$  y  $\Sigma$  es la  $s \times s$  matriz covarianzas simétrica, definida positiva. Mediante una secuencia de tres transformaciones propuesta por Genz (1992) se lleva la integral (5) a una integral sobre el hipercubo unitario de dimensión  $s - 1$  sobre el integrando transformado denotado por  $f$ .

Frecuentemente junto con el cálculo (5) se requiere de la evaluación de integrales de la forma

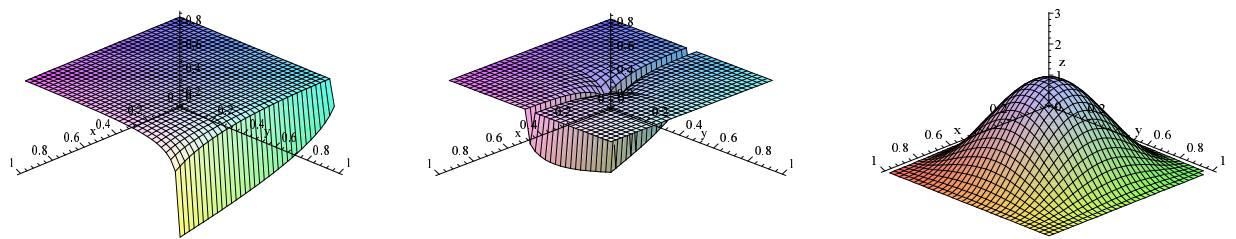
$$E(g) = \frac{1}{\sqrt{|\Sigma| (2\pi)^s}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_s}^{b_s} e^{-\frac{1}{2}\mathbf{x}^t \Sigma^{-1} \mathbf{x}} g(\mathbf{x}) d\mathbf{x},$$

donde  $g(\mathbf{x})$  puede ser una o más aplicaciones específicas de funciones de probabilidad.

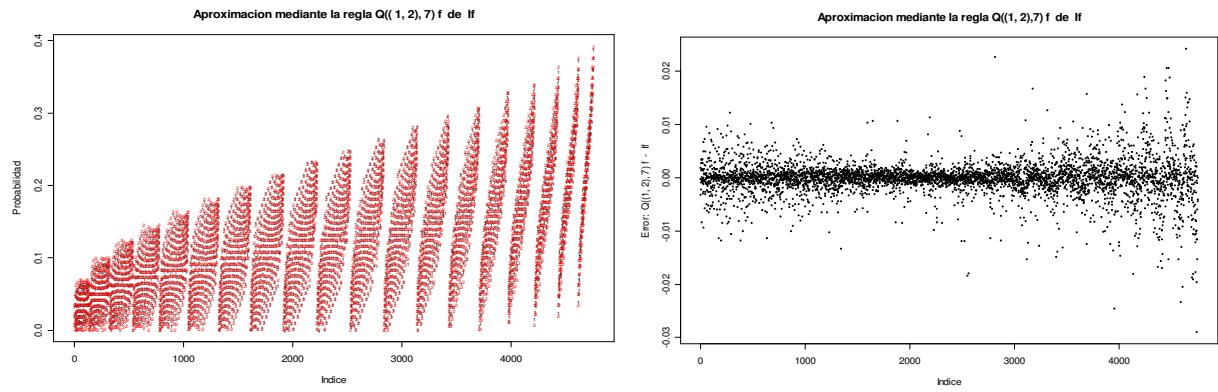
### 4. Ejemplos numéricicos

En esta sección proporcionamos dos ejemplos para ilustrar las ideas discutidas en las dos secciones previas.

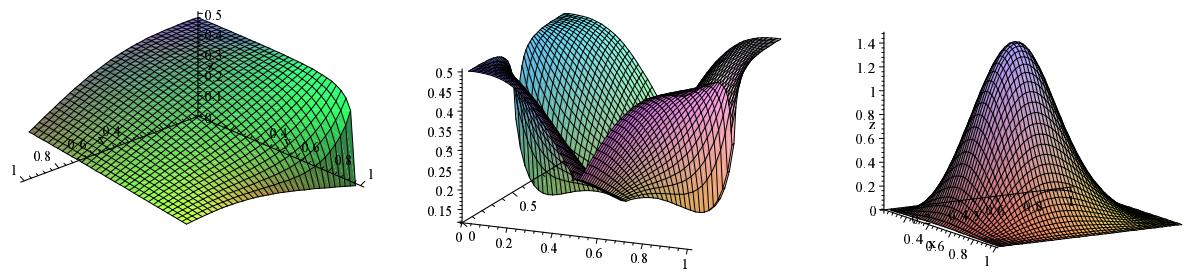
Ejemplo 1. En este ejemplo evaluamos (5) con  $s = 3$ ,  $\mathbf{a} = (-\infty, -\infty, -\infty)$ ,  $\Sigma = \left( \left(1, \frac{3}{5}, \frac{1}{3}\right)^T, \left(\frac{3}{5}, 1, \frac{11}{15}\right)^T, \left(\frac{1}{3}, \frac{11}{15}, 1\right)^T \right)$  y  $\mathbf{b} = (1, 4, 2)$ . El valor de  $F(\mathbf{a}, \mathbf{b})$  es en este caso  $F(\mathbf{a}, \mathbf{b}) = 0.82798$ . El valor aproximado para  $F(\mathbf{a}, \mathbf{b})$  obtenido al aplicar la regla de Korobov  $Q_2(P_7)f$  con  $P_7 = \left\{ \mathbf{x}_k : \mathbf{x}_k = \left\{ \frac{k(1,2)}{7} \right\}, k = 0, \dots, 6 \right\}$  a la función  $f$  obtenida utilizando las transformaciones de Genz fue  $F(\mathbf{a}, \mathbf{b}) = 0.83119$ , aplicando la transformación de periodización de baker dada por  $h(x) = 1 - 2|x - 0.5|$  a  $f$  se obtuvo  $F(\mathbf{a}, \mathbf{b}) = 0.8399$ , y finalmente al aplicar la transformación dada por  $h(x) = x^3(10 - 15x + 6x^2)$  se obtuvo  $F(\mathbf{a}, \mathbf{b}) = 0.8544$ . En la siguiente figura se muestran las gráficas del integrando  $f$  y sus dos periodizaciones



Ejemplo 2. En este ejemplo evaluamos (5) para  $s = 3$ ,  $\mathbf{a} = (-\infty, -\infty, -\infty)$ ,  $\mathbf{b} = (0, 0, 0)$  y  $\Sigma = \left( (1, \rho_1, \rho_2)^T, (\rho_1, 1, \rho_3)^T, (\rho_2, \rho_3, 1)^T \right)$  con  $\rho_i = -1 + 0.1 \times l$ ,  $1 \leq l \leq 20$ ,  $1 \leq i \leq 3$  en los casos donde  $\Sigma$  es una matriz de correlación. En las siguientes figuras se muestran las gráficas de las aproximaciones a  $F(\mathbf{a}, \mathbf{b})$  y los errores al aplicar la regla  $Q_2(P_7)$  a  $f$ , contra los valores exactos de  $F(\mathbf{a}, \mathbf{b})$ .



En las siguientes figuras se muestran las gráficas del integrando  $f$ , y sus periodizaciones donde la regla  $Q_2(P_n)$  obtuvo la peor aproximación a  $F(\mathbf{a}, \mathbf{b})$  con parámetros  $\rho_1 = 0.9$ ,  $\rho_2 = 0.7$  y  $\rho_3 = 0.9$



## Referencias

- Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*. **32**, 159–188.
- Cranley, R. y Patterson, T.N.L. (1976). Randomization of number theoretic methods for multiple integration. *SIAM Journal on Numerical Analysis* **13**(6), 904–914.
- Genton, M. (2004). *Skew-elliptical Distributions and Their Applications: A Journey Beyond Normality*, (Edited volume). Chapman and Hall/CRC.

Genz, A. (1992). Numerical Computation of Multivariate Normal Probabilities, *Journal of Computational and Graphical Statistics*, **1**, 141–149.

Hua, L.K. y Wang, Y. (1981). *Applications of Number Theory to Numerical Analysis*. Springer-Verlag, Berlin, New York, 1981.

Keast, P. (1973). Optimal Parameters for Multidimensional Integration. *SIAM Journal of Numerical Analysis*. **10**, 831-838.

Papageorgiou, A. y Traub, J. F. (1997). Faster valuation of multi-dimensional integrals. *Computers in Physics*. Nov./Dec., 574-578

Paskov, S. H. y Traub, J. F. (1995). Faster valuation of financial derivatives. *Journal of Portfolio Management*, **22**, 113-120.

Sloan, I. H. (1985). Lattice methods for multiple integration. *Journal of Computational and Applied Mathematics*, **12-13**, 131-43.

Sloan, I. H. y Kachoyan, P. J. (1987). Lattice methods for multiple integration: theory, error analysis and examples. *SIAM Journal on Numerical Analysis*, **24**, 116-28.

Fang, K.T. y Wang, Y. (1994). *Number-theoretic Methods in Statistics*. Chapman and Hall, London.

Keister, B. D. (1996). Multidimensional quadrature algorithms. *Computers in Physics*. **10**, 119-122.

Sloan, I. H. y Joe, S. (1994). *Lattice methods for multiple integration*. Oxford: Clarendon Press.



# La aplicación del análisis probit a un experimento agronómico

**Arely Elizabeth Espinosa Jiménez<sup>1</sup>**

*Facultad de Ciencias Físico Matemáticas, Universidad Autónoma de Coahuila*

**Emilio Padrón Corral**

*Centro de Investigación en Matemáticas Aplicadas*

**Félix de Jesús Sánchez Pérez<sup>2</sup>**

*Centro de Investigación en Matemáticas Aplicadas*

## 1. Introducción

El análisis probit es una técnica que permite analizar datos categóricos y dependencia estocástica en los mismos a través de variables cuantitativas. La distribución Normal de probabilidades se utiliza como parte de la transformación que define probabilidades en función de los datos categóricos. Así, primeramente se transforman los datos categóricos a variables cuantitativas y posteriormente estas variables se utilizan como argumento en la función de distribución Normal, esto define una probabilidad.

Se realizó un experimento para detener la plaga del barrenillo o picudo de chile *Anthomus eugenii Cano* existente en el municipio de Ramos Arizpe, Coahuila. En el experimento se estudiaron nueve insecticidas de diferente grupo toxicológico. Mediante el probit se determinará cual de las nueve dosis es la más eficiente para la eliminación o reducción de la plaga la cual puede causar pérdidas muy significativas en la región agrícola.

## 2. Análisis probit

El método consiste en la aplicación de correlaciones estadísticas para estimar las consecuencias desfavorables sobre la población u otros elementos vulnerables a los fenómenos físicos peligrosos que resultan de los accidentes, asociando la probabilidad de un daño.

---

<sup>1</sup>aespinosa@mate.uadec.mx

<sup>2</sup>fel1925@yahoo.com

El valor de la variable probit se determina por la expresión:

$$Y = \alpha + \beta x \text{ y } x = \log(d)$$

donde  $x$  es la variable física representativa de la dosis.

Teniéndose  $j$  dosis (p. ej. tóxicos, insecticidas, etc.) y siendo cada dosis aplicada a  $n_i$  individuos ( $i = 1, \dots, j$ ). Al finalizar el experimento se observa, si en cada individuo existió manifestación o no a la dosis. El objetivo del experimento es estimar la dosis necesaria para lograr que una proporción de la muestra presente respuesta al estímulo realizado.

Sean  $D_1, D_2, \dots, D_j$  las dosis aplicadas y sean  $r_1, r_2, \dots, r_j$  los individuos que muestran la respuesta de cada dosis. La distribución probabilística de las tolerancias a las dosis del estímulo generalmente es asimétrica, ya que se ve afectada por los individuos que tienen una tolerancia muy grande al estímulo. Por esta razón, y con objeto de volver simétrica la distribución se trabaja con los logaritmos de la dosis.

Siendo  $P_i$  un área en la distribución probabilística de tolerancia. Si se supone, que la distribución del logaritmo de la tolerancia es normal se tiene:

$$P_i = \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2} dt$$

En particular, si  $P_i = \frac{1}{2}$  la expresión anterior se convierte en :

$$P_i = \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2} dt$$

## 2.1. Método de máxima verosimilitud

Encontrando los estimadores de  $\beta_0$  y  $\beta_1$ , por mínimos cuadrados, siendo la primera interacción.

Calculando después los valores de:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_i, \hat{\mu} = \frac{5 - \hat{\beta}_0}{\hat{\beta}_1}, \hat{\sigma} = \frac{1}{\hat{\beta}_1}, \hat{z}_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}} \text{ para } i = 1, 2, \dots, k$$

Los valores calculados por tablas:

$$P_i = \varphi(\hat{z}), \quad \psi(\hat{z}_i) \text{ para } i = 1, 2, \dots, k$$

$$w_i = \frac{\left( \psi(\hat{z}_i) \right)^2}{P_i Q_i}$$

Se obtienen los nuevos probit:

$$V_i = \hat{Y}_i + \frac{\hat{P}_i - P_i}{\psi(\hat{z}_i)}$$

El método termina hasta encontrar el mejor ajuste del probit.

La dosis letal del 50% se determina para encontrar umbrales de toxicidad a un nivel del 50%:

$$DL_{50} = e^{\hat{\mu}}$$

Se tiene un proceso de interacción para localizar la mejor estimación de  $\beta_0$  y  $\beta_1$ , encontrando la dosis letal al 50%.

En el caso de tener mortalidad en el testigo los datos fueron corregidos por medio de la ecuación de Abbott.

$$MC = \frac{X - Y}{1 - Y} * 100$$

donde MC=% Mortalidad corregida, X=% Mortalidad en el tratamiento, Y=% Mortalidad en el control.

### 3. Resultados y discusión

El picudo de chile es una plaga aparentemente de origen mexicano que ataca los botones florales y frutillos de todas las variedades de chile, en las cuales oviposita y provoca su caída prematura. Esto ha provocado que la forma de atacarlo sea por medio de a químicos pero esto tiene sus desventajas ya que encarecimiento del cultivo, genera resistencia en las plagas y los riegos ambientales potenciales.

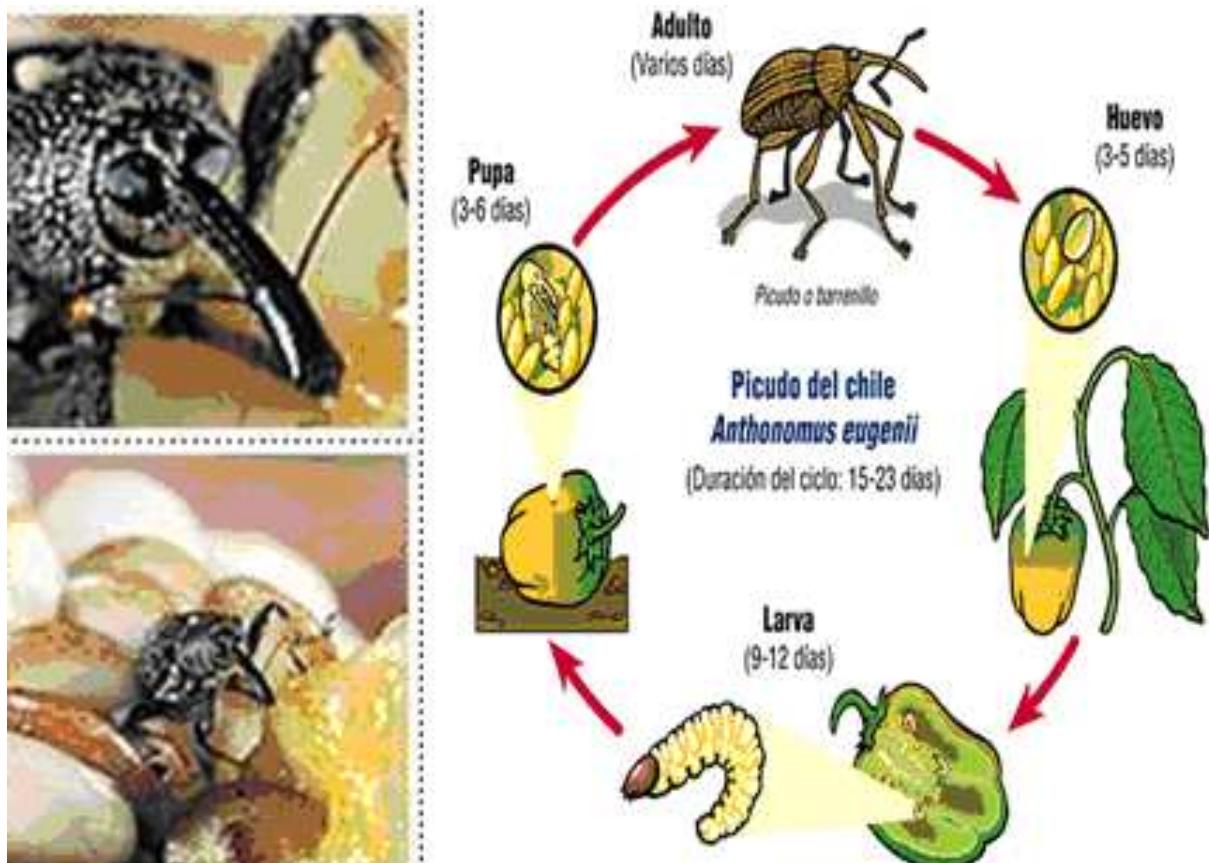


Figura 1: picudo del chile

Por ser el adulto la fase expuesta a los tóxicos y considerando su tamaño pequeño se optó por utilizar la técnica de película residual para contaminar a los adultos, y después realizar el respectivo análisis a cada insecticida.

Se obtuvieron las líneas de respuesta dosis-mortalidad para cada uno de los productos por separado para conocer la concentración o dosis letal que matan al cincuenta porciento de la población ( $DL_{50}$ ). Los productos con  $DL_{50}$  más bajas para controlar la plaga del picudo fueron metamilo, malation y paration, siendo los productos con mayor efectividad. Los productos menos efectivos fueron  $DL_{50}$  con un alto valor fueron azinfos, permetrina. A continuación ln denota datos en escala de logaritmo natural y log10 denota datos en escala de logaritmo base 10.

Al analizar los valores de la pendiente estimada se observó que las líneas de los insecticidas en escala log10 azinfos, malation, metamidofos y paration, y los de escala ln paration, metamidofos, malation y metamilo tienen los valores máximos tendiendo a una vertical.

Los insecticidas en escala log10 deltametrina y endesolfan, en cuanto a los de base ln son azifos, endosulfan y permetrina son los que tienden más a la horizontal.

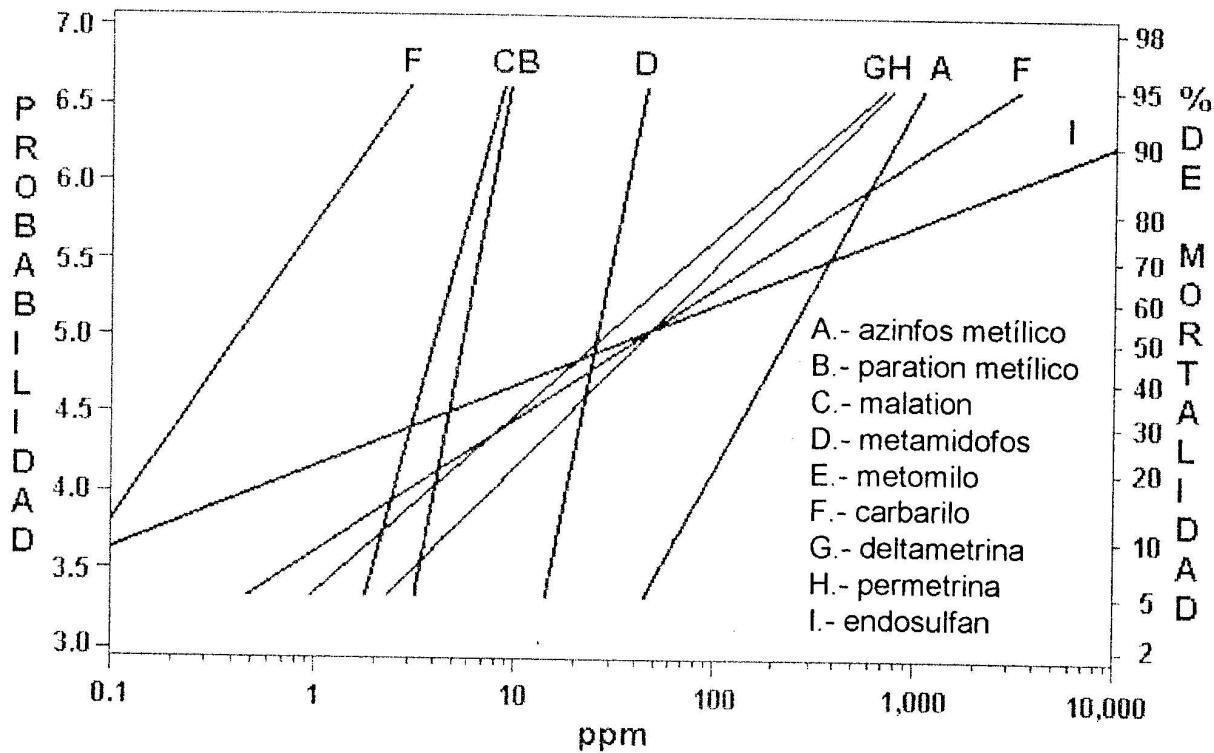


Figura 2: Líneas de respuesta dosis-mortalidad.

Insecticidas	Ec. Probit	Base	Percentil	Error Estándar	Fiduciales al 95 %	
					Inferior	Superior
Azinfos	-1.793+0.007x	In	235.203	18.790	202.080	283.762
	-5.5+2.329x	Log10	229.716	18.502	202.048	283.462
Carbarilo	-0.476+0.009x	In	52.173	12.893	20.041	102.137
	-1.381+0.857x	Log10	40.723	12.844	18.147	143.158
Deltametrina	-0.475+0.009x	In	38.678	10.0829	17.005	65.386
	-0.715+0.542x	Log10	20.844	10.331	6.436	107.046
Endosolfan	-0.430+0.007x	In	58.000	16.060	20.102	149.516
	-0.941+0.594x	Log10	38.344	17.515	13.817	198.937
Malation	-1.746+0.422x	In	4.131	0.321	3.282	4.706
	-2.917+4.816x	Log10	4.0337	0.2643	3.3967	4.530
Metamidofofos	-2.188+0.089x	In	24.545	1.557	20.386	27.292
	-8.575+6.187x	Log10	24.320	1.2585	21.255	26.658
Metamilo	-0.977+1.920x	In	0.508	0.058	0.389	0.639
	-8.575+6.187x	Log10	0.413	0.057	0.310	0.558
Paration	-2.450+0.436x	In	5.609	0.274	5.005	6.148
	-4.918+6.673x	Log10	5.457	0.229	4.978	5.922
Permetina	-0.408+0.007x	In	55.157	17.164	-0.022	87.136
	-2.144+1.289x	Log10	45.964	10.562	21.647	69.862

Cuadro 1. Concentraciones letales a un 50 % en logaritmo natural (ln), logaritmo en base 10 (log10), su respectiva ecuación probit de predicción y límites Fiduciales sobre adultos de *Anthomus eugenni Cano.*

## 4. Conclusiones

Para la obtención de los resultados fue necesario aplicar el método de máxima verosimilitud. Así este proceso daría el mejor valor en los fiduciales, en la ecuación probit así como de las demás variables.

Mediante el análisis probit se puede lograr encontrar la dosis de mortalidad  $DL_{50}$  para poder eliminar las plagas.

Los insecticidas más utilizados han generado inmunidad en los picudo de chile. Para los agricultores de la región los insecticidas más vulnerables son los más económicos .

## Referencias

- Fineey, D.J. (1971). *Probit Analysis*. Cambrige University Press 3.ed. Pág.1-5,22-80.
- Pérez Zubiri, J. R. (2000). *Susceptibilidad de Anthonomus eugenii Cano (Coleoptera: Curculionidae) a Mezclas de Dos Inergistas con Nueve Insecticidas de Diferente Grupo Toxicológico*. Tesis de Maestría de la Universidad Autónoma Agraria Antonio Narro.
- Sánchez Pérez, F. J. (1991). *Introducción al Análisis Probit*. Revista Interfase V.II no.1. Pág.9-20



# Muestreo de poblaciones humanas de difícil detección

Martín H. Félix Medina<sup>1</sup>

*Escuela de Ciencias Físico-Matemáticas de la Universidad Autónoma de Sinaloa*

## 1. Introducción

El propósito de este trabajo<sup>2</sup> es presentar una revisión actualizada de los métodos que se han propuesto para muestrear poblaciones de difícil detección. Cabe aclarar que sólo nos enfocaremos a métodos de muestreo que permiten realizar inferencias estadísticas válidas. En este trabajo, una población humana de difícil detección significará un conjunto de personas con las siguientes características: (1) representan una pequeña parte o proporción de una población más general; (2) no son geográficamente localizables; (3) posiblemente no son fácilmente distinguibles de las personas que componen la población más general; (4) posiblemente tienen alguna conducta que no es socialmente aceptable por las personas que componen la población más general, y (5) tienen lazos sociales entre ellas. Como ejemplos de este tipo de población tenemos los drogadictos, niños de la calle, indigentes, trabajadoras sexuales, homosexuales, trabajadores ilegales y enfermos de SIDA.

Las técnicas convencionales del muestreo probabilístico no son adecuadas para muestrear y estimar parámetros de poblaciones humanas de difícil detección debido a la carencia de marcos muestrales apropiados, y a la resistencia de las personas a participar en el muestreo.

Debido a lo inadecuado de los métodos convencionales de muestreo para seleccionar muestras de poblaciones de difícil detección, se han propuesto un buen número de métodos para muestrear este tipo de población. Muchos de esos métodos permiten únicamente la obtención de muestras a conveniencia del investigador, y por tanto, no posibilitan la realización de inferencias válidas acerca de los parámetros de la población de interés. Sin embargo, también se han propuesto métodos, que aunque tienen deficiencias, conducen a la obtención de muestras que permiten la realización de inferencias razonablemente buenas. Dentro de estos últimos

---

<sup>1</sup>[mhfelix@uas.uasnet.mx](mailto:mhfelix@uas.uasnet.mx)

<sup>2</sup>Trabajo realizado con apoyos parciales de los proyectos UASIN-EXB-01-01 y PIFI-2003-25-28 de la SEP y del proyecto PAFI-UAS-2002-I-MHFM-06 de la UAS

métodos, destacan el Muestreo de Redes y el Muestreo por Seguimiento de Nominaciones. Éste último con diferentes variantes, algunas de las cuales describiremos en este trabajo.

## **2. Muestreo de redes**

El Muestreo de Redes, denominado en Inglés como Network Sampling o Multiplicity Sampling, fue propuesto por Birnbaum y Sirken (1965). El desarrollo teórico de este método, así como algunas aplicaciones interesantes del mismo se pueden ver en Sirken (1970, 1972a, 1972b) y Sirken y Levy (1974). La idea detrás del Muestreo de Redes es la de dividir la población en conglomerados de personas, por ejemplo, personas que comparten una vivienda, pacientes de un hospital o trabajadores de una empresa. Luego, seleccionar una muestra aleatoria de conglomerados y pedirle a una o a varias de las personas que se localizan en cada uno de los conglomerados seleccionados que proporcionen información acerca de las personas de ese conglomerado así como también acerca de las personas fuera de ese conglomerado, pero que están asociadas, de acuerdo con algún criterio de asociación, con las personas de ese conglomerado. Esto permite incrementar la probabilidad de obtener información acerca de personas que tienen la característica de interés. Por ejemplo, si las personas que viven en una misma vivienda se definen como un conglomerado, y la relación de asociación es padres-hijos, entonces de cada vivienda seleccionada se obtiene información acerca de las personas que habitan en esa vivienda, así como también acerca de sus hijos o padres que habitan en otras viviendas.

## **3. Muestreo por seguimiento de nominaciones**

El Muestreo por Seguimiento de Nominaciones, denominado en Inglés como Link-Tracing Sampling, Snowball Sampling o Chain Referral Sampling, fue propuesto por Coleman (1958) y formulado matemáticamente por Goodman (1961). Este método es uno de los que más frecuentemente se usan en el muestreo de poblaciones de difícil detección. Ha sido usado para muestrear, por ejemplo, adictos a la cocaína (Bielesman, *et al.*, 1993), adictos a la heroína

(Frank and Snijders, 1994), drogadictos (Heckathorn, 1997) e indigentes (David y Snijders, 2002). La idea detrás de este método es la selección de una muestra inicial de elementos de la población de interés y pedirles que nominen a otros miembros de la población. Se les puede pedir a los elementos nominados que nominen a otros elementos, y continuar de esta manera con el procedimiento de nominación hasta que se cumpla alguna regla de terminación del muestreo previamente especificada. Inferencias acerca de los parámetros de la población de interés se realizan mediante el uso de modelos probabilísticos que describen los mecanismos de selección de la muestra inicial y de nominaciones.

De entre las diferentes variantes de este método, destacan las siguientes:

- *Variante de Goodman.* En este método, conocido como Muestreo por bola de nieve (Goodman, 1961), se supone que la muestra inicial es una muestra Bernoulli de la población (todos los individuos tienen la misma probabilidad de ser incluidos en la muestra y las inclusiones son independientes), que cada persona en la muestra nomina  $k$  individuos, y que el número de etapas se decide previamente al muestreo.
- *Variante de Klovdahl.* En este método, conocido como caminata aleatoria (Klovdahl, 1989), se selecciona aleatoriamente un individuo de la población. Se le pide que liste a otros miembros de la población. De esa lista se selecciona aleatoriamente un individuo, y se repite el procedimiento anterior con esa persona. El procedimiento termina hasta que se obtiene un tamaño muestral previamente especificado. Inferencias se basan en propiedades de cadenas de Markov y métodos estadísticos secuenciales.
- *Variante de Frank y Snijders.* En este método, propuesto en Frank y Snijders (1994), se supone una muestra inicial Bernoulli, y a cada individuo seleccionado se le pide que nomine a otros miembros de la población. (Las nominaciones de personas se suponen que son independientes e igualmente probables.) Estos autores proponen estimadores máximo verosímiles del tamaño poblacional.
- *Variante de Heckathorn.* En este método, conocido como Muestreo conducido por las personas muestreadas (Heckathorn, 1997 y 2002, y Salganik y Heckathorn, 2004), se selecciona una muestra inicial, no necesariamente aleatoria. A los miembros seleccionados se les remunera por participar en el estudio y se les pide que recluten a otros miembros de la población. Por cada miembro reclutado, el reclutador es remunerado. Los nuevos

miembros reclutados son remunerados, y se les pide que recluten a otros miembros, por lo cual también serán remunerados. El proceso de reclutamiento continúa hasta que se cumple alguna regla de terminación del muestreo. Estimadores de proporciones subpoblacionales se obtienen mediante propiedades límites de cadenas de Markov.

- *Variante de Félix Medina y Thompson.* En este método, propuesto por Félix Medina y Thompson (2004), se construye un marco muestral de sitios en donde los individuos de la población se pueden encontrar con alta probabilidad. (No se supone que el marco cubre a toda la población.) Se toma una muestra aleatoria simple de sitios y se identifican los miembros que pertenecen a los sitios seleccionados. En cada sitio muestreado se les pide a los miembros que nominen a otros elementos de la población. Inferencias acerca del tamaño muestral se basan en estimadores máximo verosímiles. Félix Medina y Monjardin (2004a) han propuesto estimadores del tamaño poblacional obtenidos bajo el enfoque Bayesiano. Asimismo, Félix Medina y Monjardin (2004b) han propuesto una variante en la que la muestra inicial de sitios se obtiene secuencialmente.

## 4. Conclusiones

El muestreo de poblaciones humanas de difícil detección es un problema de gran complicación ya que por las características de este tipo de población las técnicas convencionales del muestreo probabilístico no son apropiadas. Por tal razón, se han desarrollado un buen número de métodos para muestrear este tipo de población. Algunos de esos métodos no tienen sustento estadístico y conducen a la selección de muestras a conveniencia del investigador y a partir de las cuales no es posible realizar inferencias estadísticas válidas. Otros métodos, como los que presentamos en este trabajo, permiten la selección de muestras basadas en la teoría Estadística, y por tanto, es posible, a partir de ellas, realizar inferencias estadísticas aceptables. Sin embargo, aun estos métodos tienen deficiencias, por lo que ante un problema particular, se debe realizar un análisis de las ventajas y desventajas de cada uno de ellos antes de decidirse por uno en especial.

## Referencias

- Bieleman, B., Díaz, A., Merlo, G., and Kaplan, C. D. (1993). *Lines across Europe. Nature and extent of cocaine use in Barcelona, Rotterdam and Turin*. Amsterdam: Swets and Zeitlinger.
- Birnbaum. Z.W. and Sirken, M.G. (1965). Design of sample surveys to estimate the prevalence of rare deseases: three unbiased estimates. *National Center for Health Statistics, Series 2*, No. 11. U.S. Washington, D.C.: Government Printing Office.
- Coleman, J.S. (1958). Relational analysis: the study of social organizations with survey methods. *Human Organization*, **17**, 28-36.
- Félix-Medina, M.H., and Thompson, S.K. (2004). Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. *Journal of Official Statistics*, **20**, 19-38.
- Félix-Medina, M.H., and Monjardin, P.E. (2004a). Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population: a Bayesian assisted approach. En revisión en *Survey Methodology*.
- Félix-Medina, M.H. and Monjardin, P. (2004b). Link-tracing sampling with an initial sample of sites sequentially selected: estimation of the population size. *Proceedings of the Statistical Canada Symposium 2004: Innovative Methods for Surveying Difficult-to-reach Populations*. Por aparecer.
- Frank, O. and Snijders, T.A.B. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, **10**, 53-67.
- Goodman, L. (1961). Snowball sampling. *Annals of Mathematical Statistics*, **32**, 148-170.
- Heckathorn, D. (1997). Respondent-Driven Sampling: a new approach to the study of hidden populations. *Social Problems*, **44**, 174-199.
- Heckathorn, D. (2002). Respondent-Driven Sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, **49**, 11-34.

Salganik, M. and Heckathorn, D. (2004). Sampling and estimation in hidden populations using Respondent driven sampling. *Sociological Methodology*. **34**, 193-240.

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, **65**, 257-266.

Sirken, M.G. (1972a). Variance components of multiplicity estimators. *Biometrics*, **28**, 869-873.

Sirken, M.G. (1972b). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, **67**, 224-227.

Sirken, M.G. and Levy, P.S. (1974). Multiplicity estimation of proportions based on ratios of random variables. *Journal of the American Statistical Association*, **69**, 68-73.

# Una metodología para determinar el periodo de garantía para un producto

Humberto Gutiérrez-Pulido<sup>1</sup>

*Universidad de Guadalajara*

Víctor Aguirre Torres

*Instituto Tecnológico Autónomo de México*

J. Andrés Christen

*Centro de Investigación en Matemáticas*

## 1. Introducción

Uno de los objetivos básicos de muchos de los estudios de confiabilidad es decidir el tiempo de garantía para un producto, sin embargo esto prácticamente no se aborda en la literatura clásica de confiabilidad, ver por ejemplo Ansell and Phillips (1994), Lawless (1982) y Meeker y Escobar (1998). En estas obras se limitan a señalar, las más de las veces en forma implícita, que la garantía se fija con base en cuantiles bajos de la distribución del tiempo de falla. La garantía juega un papel cada día más importante, ya que esto habla de la calidad y confiabilidad del producto, y muchas de las veces es una especie de carta de presentación que ayuda a comunicar características del producto y a reforzar las estrategias de mercadotecnia. De esta manera si el periodo de garantía es pequeño esto podrá tener un efecto negativo en la elección del consumidor (Emons, 1989) y además los competidores podrán utilizarlo como una herramienta comparativa. En la literatura estadística, varios trabajos han sido orientados a aprovechar las bases de datos de las empresas en cuanto a reclamaciones de garantía para estimar el número esperado de reclamaciones, así como los costos asociados, ver por ejemplo Kalbfleisch *et al.* (1991) y Kim y Rao (2000). En este tipo de trabajos no se cuestiona la forma en la que se ha seleccionado la garantía, mas bien se estudian las consecuencias de esa elección. También se ha propuesto el uso de estos datos junto con la información de las unidades que no han fallado para estimar la distribución del tiempo de vida, ver por ejemplo Lawless (1998).

---

<sup>1</sup>humpulido@yahoo.com

Si la fijación de la garantía no se basa en la verdadera calidad y confiabilidad del producto, se puede incurrir en serios problemas: altos costos de garantía e insatisfacción del cliente. En este sentido el objetivo de este trabajo es proponer una metodología Bayesiana para decidir la garantía de un producto. Se parte de una estimación del tiempo de falla del producto y de una función de utilidad que incorpora diferentes consideraciones de costos, mercadotecnia y calidad. El lector interesado en profundizar en lo que aquí se expone lo remitimos a Gutiérrez-Pulido *et al.* (2004) y Gutiérrez-Pulido *et al.* (2006).

## 2. Función de utilidad

Sea  $L$  el número de productos a vender por la empresa en el período de referencia ( $L$  es una variable aleatoria), y sea  $t_i$  el tiempo de falla de la manufactura  $i$ -ésima  $i = 1, \dots, L$ . Determinar el tiempo de garantía  $t_w$ , de un producto es una decisión que debe ser soportar en un esquema coherente de toma de decisiones de tipo cuantitativo. Por ello proponemos una función  $u(t_i, t_w)$  que mide la utilidad monetaria cuando el producto  $i$  falla al tiempo  $t_i$  y la garantía que se le da al consumidor es  $t_w$ . Para definir  $u(t_i, t_w)$  proponemos que se tomen en cuenta tres aspectos fundamentales que contemplan las diferentes consecuencias de tomar la decisión  $t_w$ :

- ◊ Los beneficios económicos asociados a una cierta garantía,  $b(t_w)$ . Beneficios en mercadotecnia, imagen y probabilidad de venta.
- ◊ El costo directo en la que incurre el fabricante,  $r(t_i, t_w)$ , cuando el producto falla en  $t_i$  dentro del periodo de garantía  $t_w$ .
- ◊ El costo de insatisfacción del cliente debido a que el producto  $i$  falla al tiempo  $t_i$  dentro del periodo de garantía,  $I(t_i, t_w)$ .

De acuerdo a estos tres puntos, y suponiendo que todos están en las mismas unidades, la función de utilidad propuesta está dada por

$$u(t_i, t_w) = b(t_w) - r(t_i, t_w) - I(t_i, t_w). \quad (1)$$

De aquí que la utilidad total  $U(t_w)$  esté dada por

$$U(t_w) = \sum_{i=1}^L u(t_i, t_w). \quad (2)$$

A continuación detallamos cómo definir cada uno de los componentes de (1).

**Función beneficio  $b(t_w)$ .** Consideramos que  $b(t_w)$  debe ser una función creciente y acotada superiormente. Sería poco realista suponer que  $b(t_w)$  creciera sin ninguna cota superior, como en Singpurwalla y Wilson (1998), ya que dar un periodo de garantía mucho mayor al de los competidores es probable que ya no traiga un beneficio real e incluso puede causar duda o suspicacia de parte del cliente. Por lo anterior proponemos usar la función siguiente,

$$b(t_w) = A_2[1 - e^{-A_1 t_w}], \quad (3)$$

para constantes positivas  $A_1$  y  $A_2$ . Ésta es una familia flexible de funciones que son positivas, crecientes y acotadas por  $A_2$ , y cuya rapidez de crecimiento es proporcional a  $A_1$ . Los parámetros  $A_1$  y  $A_2$  deberán ser derivados de consideraciones proporcionadas por el fabricante. Específicamente consideramos que es factible que el fabricante proporcione la siguiente información en relación al producto y su garantía:

- ◊  $v = p_s - c$ , utilidad directa, con  $p_s$  el precio de venta y  $c$  el costo de producción.
- ◊  $c_r$  es el costo para la empresa para reparar o reemplazar el producto.
- ◊  $t_e$  la garantía actual o estándar del mercado.
- ◊  $\pi(t_e, p_s)$  la participación en el mercado con  $t_e$  y  $p_s$  dados.
- ◊  $M$  el tamaño del mercado potencial para el producto.
- ◊  $t_a$  es una garantía que es más atractiva para el cliente pero que más allá de ella el fabricante no espera un aumento significativo en cuanto a su participación en el mercado al precio actual. Note que  $t_a > t_e$ .

◊  $I_a$  es el aumento esperado por el fabricante en penetración en el mercado si se ofrece la garantía  $t_a$  con el precio  $p_s$ .

◊  $\pi(t_a, p_s) = \pi(t_e, p_s)[1 + Ia]$ , porción del mercado con  $t_a$  y  $p_s$  dados.

◊  $C_a = M\pi(t_e, p_s)[1 + Ia]$  es el número esperado de unidades a vender con  $t_a$  y  $p_s$  dados.

De acuerdo a lo anterior  $E(L) = C_a$ , y los beneficios totales para el fabricante con  $t_e$  y  $t_a$  están dados por  $M\pi(t_e, p_s)b(t_e) = M\pi(t_e, p_s)v$  y  $C_a b(t_a) = M\pi(t_e, p_s)[1 + Ia]v$ , respectivamente. Es fácil ver que

$$\frac{b(t_e)}{b(t_a)} = \frac{1 - e^{-A_1 t_e}}{1 - e^{-A_1 t_a}} = \frac{1}{1 + I_a}, \quad (4)$$

por lo tanto sea  $g(x) = (1 - e^{-xt_e})/(1 - e^{-xt_a})$ , entonces  $A_1$  es la solución de  $g(A_1) = \frac{1}{1+I_a}$ , que se obtiene en forma numérica. La existencia y unicidad de la solución se garantiza si  $\frac{t_e}{t_a} < \frac{1}{1+I_a}$ . En cuanto a  $A_2$  dado el significado de (3) y que está acotada por  $A_2$  entonces es razonable suponer que  $A_2 \approx b(t_a)$ , y por lo tanto  $A_2 = v$ .

**Función de costo de garantía  $r(t, t_w)$ .** Existen tres planes típicos de garantía : reemplazo, reparación e inversamente proporcional al uso o prorrateo (Menezes y Currim, 1992). El costo de garantía para estos esquemas puede ser expresados por:

$$r(t, t_w) = A_4(1 - \frac{A_3 t}{t_w}), \quad \text{para } t < t_w. \quad (5)$$

Si se tiene una garantía de reemplazo o reparación, entonces  $A_3 = 0$  y la constante  $A_4$  debe ser igual  $c_r$ , es decir  $A_4 = c_r$ . Bajo una garantía del tipo prorrateo  $(1 - \frac{A_3 t}{t_w})$  es la proporción de  $p_s$  o  $c_r$  que el usuario recibe si el producto falla al tiempo  $t$ , con  $t < t_w$ , por lo tanto  $A_3$  debe definirse bajo esa consideración y  $A_4$  será igual a  $p_s$  ó  $c_r$ .

**Función costo de insatisfacción  $I(t, t_w)$ .** El cliente no espera que el producto funcione para siempre, pero sí tiene ciertas expectativas que son reforzadas por una garantía larga. Por

lo que si el producto falla relativamente pronto, la insatisfacción del consumidor podría ser significativa. Aunque la garantía minimize o atenúe tal insatisfacción, cualquier reclamo de una garantía genera costos para el consumidor que no son cubiertos por la garantía (tiempos, desplazamientos, la interrupción en el uso del producto, frustración por no cumplimiento de expectativa, etc.). Por ello es necesario tomar en cuenta esta insatisfacción en la función de utilidad. De tal manera que se penalicen garantías largas si es que no están respaldadas por la confiabilidad del producto. Proponemos que este costo indirecto se cuantifique mediante:

$$I(t, t_w) = A_5 \left(1 - \frac{t}{t_w}\right), \text{ para } t < t_w. \quad (6)$$

La especificación de  $A_5$  se puede hacer a partir de considerar el costo del máximo nivel de insatisfacción que se daría si el producto falla en forma muy temprana. Como es difícil cuantificar esto, proponemos que se asigne como una proporción  $q$ , del precio de venta del producto  $p_s$ , por lo tanto  $A_5 = qp_s$ .

**Utilidad esperada.** De acuerdo a (1) y (2) la utilidad esperada está dada por

$$E[U(t_w)] = E[Lb(t_w)] - E\left[\sum_{i=1}^L r(t_i, t_w)\right] - E\left[\sum_{i=1}^L I(t_i, t_w)\right].$$

Como  $b(t_w)$  no depende de  $t$ , entonces  $E[Lb(t_w)] = E[L]b(t_w)$ , y de acuerdo a la información proporcionada por el productor se está suponiendo que  $L$  tiene distribución binomial ( $M, \pi(t_e, p_s)[1 + I_a]$ ), por lo que  $E[L] = C_a$ . Para obtener  $E[\sum r(t_i, t_w)]$ , se puede suponer que al menos en el periodo de la decisión  $L$  y  $t_i$  son independientes, y en consecuencia

$$E\left[\sum_{i=1}^L r(t_i, t_w)\right] = E(L)E[r(t, t_w)] = C_a A_4 \int_0^{t_w} \left(1 - \frac{A_3 t}{t_w}\right) f(t|\mathbf{X}) dt,$$

donde  $f(t|\mathbf{X})$  es la distribución posterior predictiva del tiempo de falla. Similarmente,

$$E \left[ \sum_{i=1}^L I(t_i, t_w) \right] = C_a A_5 \int_0^{t_w} \left( 1 - \frac{t}{t_w} \right) f(t|X) dt,$$

Por lo tanto

$$\begin{aligned} E[U(t_w)] &= C_a v [1 - e^{-A_1 t_w}] - C_a \int_o^{t_w} \left[ c_r \left( 1 - \frac{A_3 t}{t_w} \right) + q p_s \left( 1 - \frac{t}{t_w} \right) \right] f(t|X) dt \\ &\propto v [1 - e^{-A_1 t_w}] - \int_o^{t_w} \left[ c_r \left( 1 - \frac{A_3 t}{t_w} \right) + q p_s \left( 1 - \frac{t}{t_w} \right) \right] f(t|X) dt. \end{aligned} \quad (7)$$

Por lo que el valor de  $C_a$  no influye en la elección de  $t_w$ , y con ello realmente no es necesario conocer  $M$ . La decisión óptima para la garantía  $t_w^*$ , está dada por el  $t_w$  que maximiza (7).

### 3. Implementación de la Metodología

Para aplicar el procedimiento anterior, hay algunos pasos críticos:

- ◊ Seleccionar el modelo  $f(t|\theta)$  que describe adecuadamente el tiempo de falla del producto. Para ello, dados los datos  $X$ , en Gutiérrez-Pulido *et al.* (2003), se describe un procedimiento Bayesiano de selección de modelos que considera los modelos más usuales en confiabilidad.
- ◊ Para obtener  $f(t|X)$  es necesario especificar la distribución a priori para sus parámetros  $\theta$ . En Gutiérrez-Pulido *et al.* (2005), se describe un procedimiento que a partir de poca información inicial sobre el tiempo de falla se especifican los parámetros de los modelos más usuales en confiabilidad.
- ◊ En los más de los casos  $f(t|X)$  y (7) no tienen forma analítica. En estos casos será necesario obtener por simulación a  $f(t|X)$ . Un método relativamente fácil de implementar para los modelos de confiabilidad es el Sampling-importance-resampling (Robert y Casella, 1999). Sea  $t^{(1)}, t^{(2)}, \dots, t^{(N)}$ , una muestra de tiempos de falla obtenidos a partir de  $f(t|X)$ , entonces es claro que

$$E[U(t_w)] \approx \frac{1}{N} \sum_{k=1}^N u(t^{(k)}, t_w). \quad (8)$$

◊ El óptimo  $t_w^*$  para (8) se puede encontrar fácilmente en forma numérica.

El lector interesado puede encontrar la aplicación del procedimiento anterior a un problema de selección del tiempo de garantía para balatas de frenos para automóvil, en Gutiérrez-Pulido *et al.* (2006).

## Referencias

- Ansell, J.I. and Phillips M.J.(1994). *Practical methods for reliability data analysis*. Clarendon Press, Oxford.
- Emons, W. (1989). On the limitation of warranty duration. *Journal of industrial economics*, 37, 3, 287-301.
- Gutiérrez-Pulido, H., Aguirre-Torres, V. and Christen, J.A. (2003). Bayesian model evaluation in reliability. *Technical report DE-C03.17*, Statistics department, ITAM, México.
- Gutiérrez-Pulido, H., Aguirre-Torres, V. and Christen, J.A. (2004). A Bayesian approach for the determination of warranty length. *Reporte técnico DE-C04.6*, Departamento de Estadística, ITAM, México.
- Gutiérrez-Pulido, H., Aguirre-Torres, V. and Christen, J.A. (2005). A practical method for obtaining prior distributions in reliability. *IEEE transactions on reliability*, 54, 2, 262-269.
- Gutiérrez-Pulido, H., Aguirre-Torres, V. and Christen, J.A. (2006). A Bayesian approach for the determination of warranty length. Aceptado para su publicación en *Journal of quality technology*.
- Kalbfleisch, J.D., Lawless, J.F., and Robinson, J.A. (1991). Methods for the analysis and prediction of warranty claims. *Technometrics*, 33, 3, 273-285.

Kim, H.G. and Rao, B.M. (2000). Expected warranty cost of two-attribute free-replacement warranties based on a bivariate exponential distribution. *Computers and industrial engineering*, 38, 425-434.

Lawless, J.F. (1982). *Statistical models and methods for lifetime data*. Wiley:N.York.

Lawless, J.F. (1998). Statistical analysis of product warranty data. *International statistical review*, 66, 41-60.

Meeker, W.Q. and Escobar, E. (1998). *Statistical methods for reliability data*. Wiley:N.York.

Menezes, M.A.J. and Currim, I.S. (1992). An approach for determination of warranty length *International journal of research in marketing*, 9, 177-195.

Robert, C.P. and Casella, G. (1999). *Monte Carlo statistical methods*, Springer:N.York.

Singpurwalla, N.D. and Wilson, S.P. (1998). Failure models indexed by two scales. *Advances in applied probability.*, 30, 1058-1072.

# Intervalos de confianza en el modelo de regresión logística, en presencia de separación de los datos y colinealidad entre las variables explicatorias

Flaviano Godínez Jaimes<sup>1</sup>

*Unidad Académica de Matemáticas de la Universidad Autónoma de Guerrero*

Gustavo Ramírez Valverde

*Especialidad de Estadística. ISEI. Colegio de Postgraduados*

## 1. Introducción

Sea  $\{(Y_1, x_{11}, \dots, x_{1p}), \dots, (Y_n, x_{n1}, \dots, x_{np})\}$  una muestra aleatoria en regresión binaria donde las  $Y_i$ 's son variables aleatorias independientes con distribución Bernoulli con probabilidad de éxito desconocida  $\pi_i = P(Y = 1)$  y  $x_{i1}, \dots, x_{ip}$  son valores fijos de las variables explicatorias  $X_1, \dots, X_p$ . Sean  $\mathbf{X}$  la matriz diseño de  $n \times (p + 1)$  cuyos renglones son  $x_i^T = (1, x_{i1}, \dots, x_{ip})$ . Los datos siguen el modelo de regresión logística si

$$\pi_i = P(Y = 1) = e^{x_i^T} / (1 + e^{x_i^T})$$

donde  $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$  es un vector de parámetros desconocidos.

El estimador de máxima verosimilitud (EMV)  $\hat{\beta}$ , del modelo de regresión logística no existe cuando hay separación o cuasi separación en los datos y éste existe y es único cuando hay traslape en los datos (Albert y Anderson, 1984; y Santner y Duffy, 1986). Hay separación en los datos si existe un hiperplano que separa los éxitos ( $Y = 1$ ) de las fracasos ( $Y = 0$ ), hay cuasi separación si además, al menos una observación esta en el hiperplano y hay traslape si no existe tal hiperplano.

La matriz de varianzas y covarianzas estimada del EMV es la inversa de la matriz de información (MI)  $X^T \hat{V} X$ , donde  $\hat{V} = \text{diag}\{\hat{\pi}_1(1 - \hat{\pi}_1), \dots, \hat{\pi}_n(1 - \hat{\pi}_n)\}$ . La MI es importante

---

<sup>1</sup>fgodinezj@colpos.mx

para obtener el EMV, construir intervalos de confianza y probar hipótesis. La MI puede ser singular debido a la existencia de dependencias lineales cercanas entre las variables explicativas (x-colinealidad) y/o la existencia de separación en los datos. Lesaffre y Marx (1993) reconocieron el efecto de estos factores en la singularidad de la MI y llamaron mv-colinealidad al problema de colinealidad en la MI. La x-colinealidad y la mv-colinealidad son medidos con el número de condición escalado de  $\mathbf{X}$  y MI respectivamente. Cuando los números de condición escalados son mayores a 30 entonces las inferencias obtenidas a partir del EMV pueden no ser confiables y se dice que la colinealidad es severa.

Rousseeuw y Christmann (2003) y Heinze y Schemper (2002) propusieron estimadores que existen en presencia de separación en los datos pero que son negativamente afectados por la existencia de x-colinealidad.

En este trabajo se presentan intervalos de confianza bootstrap paramétricos basados en un estimador de tipo ridge que existe aun cuando hay separación en los datos y/o x-colinealidad. En la sección dos se describen el estimador y los intervalos de confianza estudiados. En la sección tres se presenta un estudio de simulación para valorar el cubrimiento y amplitud de los intervalos de confianza bootstrap paramétricos.

## 2. Estimador e intervalos de confianza estudiados

El estimador ridge logístico (ERL) está dado por

$$\hat{\beta}_R(k) = \left[ X^T \hat{V} X + kI \right]^{-1} X^T \hat{V} X \hat{\beta}$$

donde  $\hat{\beta}$  es el EMV del modelo de regresión logística e I es la matriz identidad de orden  $p+1$ . En presencia de x-colinealidad en el modelo de regresión lineal, Liu (2003) propuso calcular  $k$  con

$$k_L = (\lambda_1 - 100\lambda_p)$$

donde  $\lambda_1$  y  $\lambda_p$  son el mayor y el menor de los valores propios de  $X^T X$ . Con esta propuesta  $k$  no depende de  $\beta$ . le Cessie y van Houwenlingen (1992) obtuvieron un estimador ridge penalizando la log-verosimilitud con el cuadrado de la norma de  $\beta$  y estimando el vector de

parámetros de forma iterativa con el método de Newton-Raphson

$$\beta_R^{(s+1)} = \beta_R^{(s)} + \left\{ X^T \hat{V} X + 2kI \right\}^{-1} \left\{ X^T (Y - \pi) - 2k\beta_R^{(s)} \right\}$$

denominaremos estimador ridge-Liu iterativo logístico (RLI) al estimador obtenido,  $\hat{\beta}_{RLI}$ , cuando  $k = k_L$ . El estimador RLI existe aun en casos en que hay separación en los datos y/o x-colinealidad y en pruebas preliminares se observó que tiene menor error cuadrático medio que el EMV y que los estimadores estudiados por Heinze y Schemper y Rousseeuw y Christmann.

La matriz de varianzas y covarianzas estimada del estimador RLI es

$$\left( X^T \hat{V} X + 2k_L I \right)^{-1} \left( X^T \hat{V} X \right) \left( X^T \hat{V} X + 2k_L I \right)^{-1},$$

pero esta estimación de  $V(\hat{\beta}_{RLI})$  no puede usarse para construir intervalos de confianza para  $\beta$  porque no se conoce la distribución de este estimador. Aunque se sabe que en estos casos se pueden usar técnicas de remuestreo, esto no ha sido usado pues aun cuando en la muestra original no haya separación en los datos, es posible que en la muestra bootstrap si exista este problema. Esta es la razón por lo que es necesario un estimador como  $\hat{\beta}_{RLI}$  pues existe en los casos extremos donde hay separación en los datos y/o colinealidad.

## 2.1. Intervalos de confianza de máxima verosimilitud en regresión logística

Es conocido que asintóticamente  $\hat{\beta} \sim N\left(\beta, \left(X^T \hat{V} X\right)^{-1}\right)$ , por tanto un intervalo de confianza de máxima verosimilitud del  $100(1 - \alpha)\%$  para  $\beta$  se define por  $\hat{\beta} \pm z_{1-\alpha/2} s_{\hat{\beta}}$ .

## 2.2. Intervalos de confianza bootstrap paramétricos percentiles en regresión logística

La distribución de probabilidad bootstrap paramétrica de  $\hat{\beta}_{RLI}$  se obtiene en la forma siguiente:

- a) Dada la muestra original  $z = \{z_i, \dots, z_n\}$  donde  $z_i = (Y_i, x_{1i}, \dots, x_{pi})$  se obtiene  $\hat{\beta}_{RLI}$ .
- b) Se generan nuevos valores de las variables explicatorias, independientes y en el rango de los valores muestrales:  $X_i^N \sim U [\min(X_i), \max(X_i)], i = 1, \dots, p$ .
- c) El vector respuesta  $Y^* = (y_1^*, y_2^*, \dots, y_n^*)^T$  se genera con el modelo de regresión logística en función de  $X_1^N, \dots, X_p^N$  y  $\hat{\beta}_{RLI}$ .
- d) Se obtiene  $\hat{\beta}_{RLI}^*(i)$  usando  $z^* = \{z_1^*, \dots, z_n^*\}$ , donde  $z_i^* = (y_i^*, x_{1i}^N, \dots, x_{pi}^N)$ .
- e) Se repiten B veces b), c) y d) y se asigna probabilidad  $1/B$  a cada  $\hat{\beta}_{RLI}^*(i)$ .
- f) La función de distribución acumulada bootstrap paramétrica de  $\hat{\beta}_{RLI}$  se define por

$$G(s) = \# \left( \hat{\beta}_{RLI}^*(i) < s \right) / B.$$

Sea  $\hat{\beta}_{RLI}^*[\alpha] = \hat{G}^{-1}(\alpha)$  el percentil  $\alpha$ -ésimo de la distribución acumulada bootstrap paramétrica de  $\hat{\beta}_{RLI}$  y  $z_\alpha$  al percentil  $\alpha$ -ésimo de la distribución Normal estándar. Los intervalos de confianza bootstrap paramétricos percentiles del  $100(1 - \alpha)\%$  para  $\beta$  (Manly, 1991) se definen por:

1. De Efron:  $[\hat{\beta}_{RLI}^*[\alpha/2], \hat{\beta}_{RLI}^*[1 - \alpha/2]]$ .

2. De Hall:  $[ 2\hat{\beta}_{RLI} - \hat{\beta}_{RLI}^* [1 - \alpha/2], 2\hat{\beta}_{RLI} - \hat{\beta}_{RLI}^* [\alpha/2] ]$ .
  3. Corregido por sesgo:  $[ \hat{\beta}_{RLI}^* [\Phi(2z_0 + z_{\alpha/2})], \hat{\beta}_{RLI}^* [\Phi(2z_0 + z_{1-\alpha/2})] ]$ , donde  $z_0 = z_{1-p}$  es el percentil  $1-p$  de la distribución Normal estándar y  $p$  es la proporción de veces que  $\hat{\beta}_{RLI}^*$  excede a  $\hat{\beta}_{RLI}$ .
  4. Corregido por sesgo acelerado:
- $$\left[ \hat{\beta}_{RLI}^* \left[ \Phi \left( z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})} \right) \right], \hat{\beta}_{RLI}^* \left[ \Phi \left( z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})} \right) \right] \right],$$
- donde  $z_0$  se define igual que en 3) y  $a \approx \sum_{i=1}^n (\bar{\hat{\beta}}^J - \hat{\beta}_{RLI}(i))^3 / 6 \left\{ \sum_{i=1}^n (\bar{\hat{\beta}}^J - \hat{\beta}_{RLI}(i))^2 \right\}^{1.5}$ .

### 3. Simulación y resultados

Se hizo una simulación para estudiar el desempeño del estimador  $\hat{\beta}_{RLI}$  donde y de los intervalos de confianza bootstrap paramétricos estudiados usando dos variables y los siguientes factores:

1. Dos grados de correlación muestral (CM): alta ( $r=0.95$ ) y severa( $r=0.99$ ).
2. Dos tamaños muestras (TM) de 20 y 40 observaciones.
3. Dos orientaciones para  $\beta$  , estas corresponden a los vectores propios asociados al valor propio mayor y menor de  $X^T X$ .
4. Dos tamaños de  $\beta$  (TVP): 1 y 9, éstos se obtienen multiplicando por 1 y 3 el vector propio correspondiente.

5. El porcentaje de traslape (PT) se define como  $100 * (\text{número de observaciones traslapadas} / \text{número de observaciones})$ . Este se midió en los datos generados y se clasificó en cinco categorías: PT0, PTI, PTII, PTIII y PTIV en las cuales los porcentajes de traslape están en los intervalos 0, (0, 10], (10, 20], (20, 30], (30, 40].

Las matrices diseño,  $X = [1 X_1 X_2]$ , se generaron de orden  $3 \times 20$  y  $3 \times 40$ . Con  $X_1 \sim U[0, 1]$  y  $X_2$  se obtuvo con  $X_2 = X_1 + cu$ , donde  $u \sim U[0, 1]$ , y  $c$  toma valores que permitieron obtener correlaciones muestrales aproximadas de 0.95 y 0.99. La variable respuesta  $Y$  fue generada con  $Y_i = 1$  y  $\pi_i > w$  y  $Y_i = 0$  en otro caso; donde  $w \sim U[0, 1]$ .

Los intervalos de confianza se comparan respecto a:

1. Cubrimiento,  $C = \frac{1}{R} \sum_{r=1}^R I_{[\tilde{\beta}_{INF,r}, \tilde{\beta}_{SUP,r}]}(\beta)$ .
2. Longitud,  $L = \frac{1}{R} \sum_{r=1}^R (\tilde{\beta}_{SUP,r} - \tilde{\beta}_{INF,r})$ .

En cada una de las combinaciones de CM x TM x VP x TVP se hicieron 10000 repeticiones en las que se calculó el cubrimiento y longitud de los intervalos de confianza del 95 % para  $\beta_1$  y  $\beta_2$ . La determinación del número de observaciones traslapadas se hace con una versión en SAS del procedimiento NOOVERLAP, el cual determina de manera exacta el número mínimo de observaciones que hay que eliminar para que en el resto haya separación completa (Christmann y Rousseeuw, 2001).

Cuadro 1. Cubrimiento (C) y longitud (L) de los intervalos de confianza de máxima verosimilitud (MV) y bootstrap paramétricos percentiles (ICBPP) del 95 % de tamaño de vector propio 1 y vector propio 1.

TM	CM	PTO		PTI		PTII		PTII		PTIV	
		L	C	L	C	L	C	L	C	L	C
40	0.99	MV	B1	†	†	53.14	<b>0.94</b>	38.20	<b>0.96</b>	33.01	<b>0.98</b>
			B2	†	†	52.04	<b>0.94</b>	37.64	<b>0.95</b>	32.57	<b>0.98</b>
		E	B1	†	†	2.58	<b>1.00</b>	2.68	<b>1.00</b>	2.70	<b>1.00</b>
			B2	†	†	2.55	<b>1.00</b>	2.63	<b>1.00</b>	2.63	<b>1.00</b>
		H	B1	†	†	2.58	<b>0.97</b>	2.68	<b>0.99</b>	2.70	<b>1.00</b>
			B2	†	†	2.55	<b>0.96</b>	2.63	<b>0.99</b>	2.63	<b>1.00</b>
		CS	B1	†	†	2.54	<b>0.97</b>	2.65	<b>0.99</b>	2.69	<b>1.00</b>
			B2	†	†	2.51	<b>0.97</b>	2.60	<b>0.99</b>	2.62	<b>1.00</b>
		CSA	B1	†	†	2.59	<b>0.97</b>	2.68	<b>0.99</b>	2.69	<b>1.00</b>
			B2	†	†	2.57	<b>0.97</b>	2.63	<b>0.99</b>	2.62	<b>1.00</b>
		NR				<b>1129</b>		<b>6296</b>		<b>2504</b>	<b>70</b>
20	0.99	MV	B1	*	*	155.79	<b>0.99</b>	93.38	<b>0.99</b>	79.54	<b>1.00</b>
			B2	*	*	148.87	<b>0.98</b>	88.86	<b>0.99</b>	75.68	<b>1.00</b>
		E	B1	7.14	<b>1.00</b>	6.61	<b>1.00</b>	6.12	<b>1.00</b>	5.82	<b>1.00</b>
			B2	7.03	<b>1.00</b>	6.49	<b>1.00</b>	5.99	<b>1.00</b>	5.69	<b>1.00</b>
		H	B1	7.14	<b>1.00</b>	6.61	<b>1.00</b>	6.12	<b>1.00</b>	5.82	<b>1.00</b>
			B2	7.03	<b>1.00</b>	6.49	<b>1.00</b>	5.99	<b>1.00</b>	5.69	<b>1.00</b>
		CS	B1	211.98	<b>1.00</b>	29.15	<b>1.00</b>	7.94	<b>1.00</b>	5.91	<b>1.00</b>
			B2	270.70	<b>1.00</b>	41.43	<b>1.00</b>	9.11	<b>1.00</b>	5.96	<b>1.00</b>
		CSA	B1	10386.55	<b>1.00</b>	561.72	<b>1.00</b>	16.19	<b>1.00</b>	5.92	<b>1.00</b>
			B2	1919.57	<b>1.00</b>	530.13	<b>1.00</b>	25.29	<b>1.00</b>	5.97	<b>1.00</b>
		NR		<b>295</b>		<b>3687</b>		<b>4807</b>		<b>1191</b>	<b>20</b>

NR: Número de replicas; E: ICBPP de Efron; H: ICBPP de Hall; CS: ICBPP corregido por sesgo; CSA: ICBPP por sesgo acelerado CM; Correlación muestral.

\* No existe estimación de MV.

† La simulación no generá estos casos.

### **3.1. Resultados**

El estimador RLI resuelve los problemas de x-colinealidad, mv-colinealidad y separación en los datos. Este estimador existe aún en casos donde hay separación en los datos y elimina el problema de mal condicionamiento de la matriz de información producido por x-colinealidad severa y/o poco traslape.

Los intervalos de confianza de Efron tienen el mejor desempeño en términos de cubrimiento y longitud, seguidos muy de cerca de los intervalos de confianza de Hall. Los intervalos de confianza corregidos por sesgo y corregidos por sesgo acelerados pueden, con 40 observaciones, tener casi el mismo desempeño que los de Efron y de Hall pero con 20 observaciones y en PT0 y PTI tienen longitudes que son mayores, incluso que los intervalos de confianza de máxima verosimilitud.

Los intervalos de confianza de máxima verosimilitud tienen cubrimiento mayor a 0.90 en el 73 % de todos los escenarios estudiados, los de Efron en el 78 %, los de Hall en 67 %, los corregidos por sesgo en 63 % y los corregidos por sesgo acelerado en 61 %.

Los intervalos de confianza de Efron y de Hall tienen la misma longitud y además son los que tienen la menor longitud, después están los corregidos por sesgo, los corregidos por sesgo acelerado y por último los de máxima verosimilitud. Sin embargo, hay escenarios donde la longitud de los intervalos corregidos por sesgo acelerado es mayor que los de máxima verosimilitud.

## **4. Conclusiones**

Los intervalos de confianza bootstrap paramétricos percentiles de Efron y de Hall basados en el estimador ridge-Liu logístico son mejores en cubrimiento y longitud que los de máxima verosimilitud en cualquier porcentaje de traslape y especialmente en los casos donde hay separación en los datos.

## Referencias

- Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 71:1-10.
- Christmann, A., and Rousseeuw, P.J. (2001). Measuring overlap in logistic regression. *Computational Statistics and Data Analysis*. 37:65-75.
- Heinze, G., and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21: 2409-2419.
- le Cessie, S. and van Houwelingen, J.C. (1992). Ridge estimators in logistic regression. *Applied Statistics*. 41(1):191-201.
- Lesaffre, E. and Marx, B.D. (1993). Collinearity in generalized linear regression. *Communications in Statistics-Theory and Methods*. 22(7):1933-1952.
- Liu, K. (2003). Using Liu-type estimator to combat collinearity. *Communications in Statistics-Theory and Methods*, Vol 32, No 5, pp 1009-1020.
- Manly, B. F. (1991). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall. London,UK.
- Rousseeuw, P. J. and Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis* 43, 315-332.
- Santner, T. J. and Duffy, D.E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 73:755-758.



# **Regresión múltiple para la interpretación de configuraciones de suelos en la sierra norte de Puebla obtenidas por escalamiento multidimensional**

**Gladys Linares Fleites**

**Miguel Ángel Valera Pérez**

**María Guadalupe Tenorio Arvide<sup>1</sup>**

*Departamento de Investigaciones en Ciencias Agrícolas. Instituto de Ciencias. Benemérita Universidad Autónoma de Puebla*

**Maribel Castillo Morales<sup>2</sup>**

*Facultad de Ingeniería Química. Benemérita Universidad Autónoma de Puebla*

## **1. Introducción**

Las técnicas de visualización de datos multidimensionales, entre las que se encuentran las de Escalamiento Multidimensional (EM), permiten la proyección en un espacio de baja dimensión de un conjunto de datos de alta dimensión preservando las relaciones topológicas originales.

Se aplicó una de las técnicas de EM a un estudio de suelos de carga variable en la sierra norte del estado de Puebla (Tenorio, 2003), donde se habían seleccionado diferentes unidades de suelo que ya mostraban evidencias de degradación ambiental por causas del uso del mismo. Las muestras de suelo, tomadas de forma representativa en la zona de estudio, fueron preparadas y caracterizadas en el laboratorio mediante la determinación de parámetros físicos, químicos, mineralógicos y biológicos.

---

<sup>1</sup>cs001985@siu.buap.mx

<sup>2</sup>lebiram2702@hotmail.com

Previo a la aplicación de EM se realizó un análisis exploratorio de datos con las determinaciones físicas y químicas que incluían 14 variables de 21 horizontes o perfiles de suelo, hallándose a través de un Análisis de Componentes Principales que con sólo ocho de ellas era posible explicar el comportamiento de la degradación de los perfiles de suelo de la zona. (Valera *et al.*, 2001). Posteriormente, aplicando la técnica de conglomerado jerárquico se logró una clasificación exitosa de los niveles de degradación de estos suelos. (Linares *et al.*, 2004). La representación en un espacio de dos dimensiones de las disimilaridades entre los horizontes de suelos, dados como coordenadas de ocho dimensiones, corroboró los resultados anteriores (Linares *et al.*, 2005).

El objetivo del presente trabajo es indagar la relación entre la disposición espacial de los estímulos y algunas características de éstos que han sido medidas independientemente a través de Regresión Múltiple.

## 2. Resultados del escalamiento multidimensional

Estas técnicas abordan el problema de construir distancias métricas transformando de manera adecuada las disimilaridades entre objetos. (Linares, 2001 y Miret, *et al.*, 2002). La entrada básica de un análisis de Escalamiento Multidimensional son los valores de similaridad o disimilaridad entre todos los pares de  $n$  objetos. En el problema que estamos considerando los objetos son  $n = 21$  horizontes de suelos de carga variable en la Sierra Norte del Estado de Puebla y las disimilaridades se obtuvieron a través de la distancia Euclíadiana entre  $p = 8$  variables. Estas variables son: pH, porcentaje de carbono orgánico (%C), porcentaje de saturación de bases (V%), porcentaje de arcilla, porcentaje de Fe extraído con TAMM (OXAFE), porcentaje de Al extraído con TAMM (OXAAL), porcentaje de Fe extraído con DBC (DBCFE) y porcentaje de Al extraído con DBC (DBCAL).

Los resultados se obtuvieron utilizando el módulo Multidimensional Scaling del software STATISTICA (1998). Se inició el cálculo con la matriz de distancias Euclidianas obtenida a partir de los datos observados. La solución inicial se estimó con el método de Guttman-Lingoes y, a continuación, se encontró la configuración óptima usando un algoritmo iterativo basado en el método de optimización de máximo descenso. El coeficiente de esfuerzo (Stress) de Kruskal obtenido fue de 0.0560076, lo que nos permite calificar como “buena” la precisión

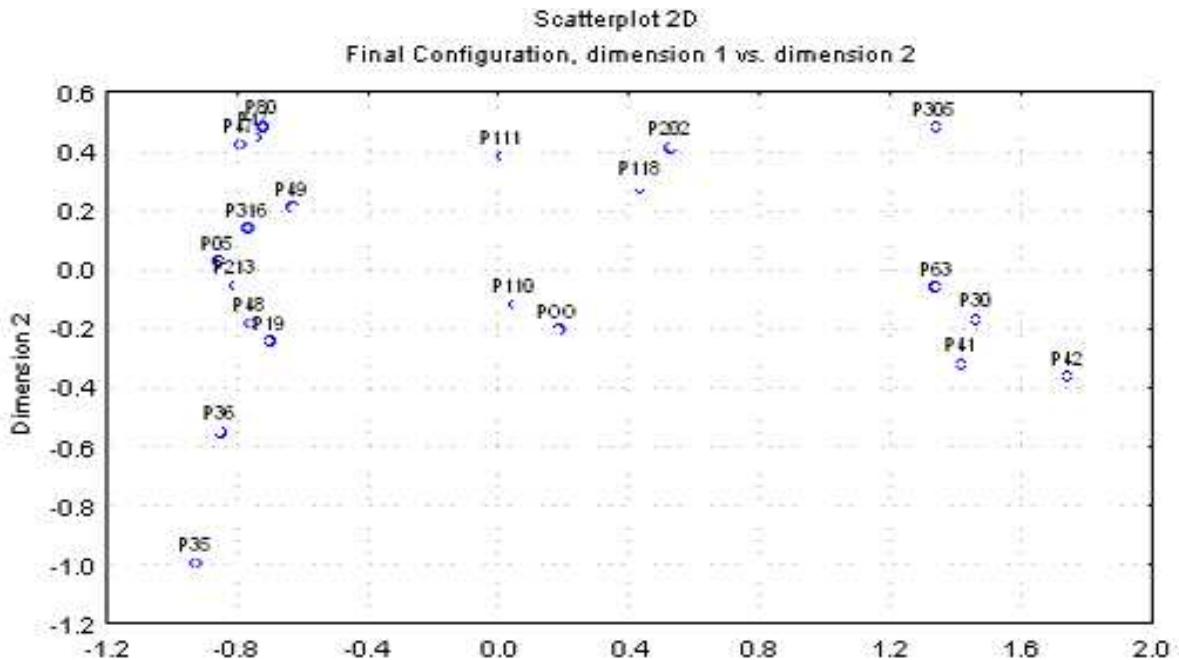


Figura 1: Configuración de puntos

de la configuración de puntos finales. La Figura 1 muestra el gráfico de la configuración de puntos. La dimensión del eje horizontal del gráfico expresa la degradación por erosión mientras que la dimensión del eje vertical expresa la degradación biológica de los suelos.

Los suelos con más alta degradación ambiental por erosión (perfils P305, P42, P41, P30 y P63) son puntos situados en el extremo derecho del eje horizontal del gráfico, y todos excepto el primero están en el cuarto cuadrante, lo que expresa que son suelos con fuerte degradación biológica.

Los suelos con un menor grado de degradación por erosión están situados en el lado extremo izquierdo del gráfico. De éstos, los mejor conservados están situados en el segundo cuadrante del gráfico. La dimensión del eje horizontal del gráfico expresa la degradación por erosión mientras que la dimensión del eje vertical expresa la degradación biológica de los suelos.

Dependiente	R mult	R <sup>2</sup>	F(2,18)	p- empi	$\beta$ dim 1	$\beta$ dim 2
pH	0.47	0.22	2.51	<0.108	-0.4349	-0.1714
%C	0.40	0.16	1.73	<0.206	-0.0085	0.4011
%V	0.57	0.33	4.408	<0.027	0.3186	<b>-0.4767</b>
<b>Arcilla</b>	<b>0.99</b>	<b>0.99</b>	<b>44.92</b>	<b>&lt;0.000</b>	<b>-0.9983</b>	-0.0368
OXA-FE	0.67	0.45	7.33	<0.005	-0.3292	0.1749
OXA-AL	0.61	0.37	5.25	<0.016	-0.4629	0.3927
<b>DBC-FE</b>	<b>0.90</b>	<b>0.81</b>	<b>39.33</b>	<b>&lt;0.000</b>	<b>-0.8685</b>	0.2437
DBC-AL	0.85	0.73	23.84	<0.000	<b>-0.8278</b>	0.2013

Figura 2: Resultados del análisis de regresión múltiple con las 8 variables utilizadas en EM

### 3. Resultados de la regresión múltiple

La interpretación de la configuración puede hacerse averiguando si existe alguna relación entre la disposición espacial de los estímulos y algunas características de éstos que han sido medidas independientemente. (Real, 2001). Las características se utilizan como “datos externos” que pueden ponerse en relación con las coordenadas de los estímulos en las distintas dimensiones mediante Regresión Múltiple. Este procedimiento puede resultar de ayuda cuando no resulta sencillo interpretar las posiciones de los estímulos en el espacio directamente, o bien cuando existen varias posibilidades que compiten en la interpretación.

La figura 2 muestra los resultados obtenidos al aplicar regresión múltiple a las diferentes características consideradas. Aquellas características relacionadas con la solución proporcionada por ED mostraron un coeficiente de regresión múltiple elevado y significativo, y un peso importante y significativo en alguna o algunas de las dimensiones, lo que indica una elevada relación entre esa característica y la dimensión o las dimensiones correspondientes. Las variables arcilla y Al y Fe extraídos explican en gran medida el comportamiento químico de estos suelos, expresando el grado de degradación por erosión, mientras que las variables %V y relación carbono - nitrógeno expresan la degradación biológica de los suelos.

## 4. Conclusiones

La interpretación de la configuración de los puntos obtenida por Escalamiento Multidimensional, puede esclarecerse a través de la regresión múltiple, ya que vemos confirmada, por los coeficientes obtenidos, nuestra suposición inicial de que la dimensión 1 estaba relacionada con la degradación de los suelos por erosión, mientras que la dimensión 2 refleja la degradación biológica de estos suelos.

## Referencias

- Linares, G. (2001). Escalamiento Multidimensional: conceptos y enfoques. *Revista Investigación Operacional*. **28**, 173 - 183.
- Linares, G., Valera, M.A. y Tenorio, M.G. (2004). Análisis de Conglomerados Jerárquicos en la Clasificación de Suelos Ácidos de la Sierra Norte de Puebla. *XIX Foro Nacional de Estadística*. Monterrey, México.
- Linares, G., Tenorio, M.G. y Valera, M.A. (2005). Visualización del diagnóstico de la degradación de suelos de carga variable en la Sierra Norte del estado de Puebla con Escalamiento Multidimensional. *X Congreso Nacional y IV Internacional de Ciencias Ambientales*. Chetumal, México.
- Miret, E., Linares, G. y Mederos, M.V. (2002). Estudio comparativo de procedimientos de Escalamiento Multidimensional a través de Experimentos de Simulación. *Revista Investigación Operacional*. **23**, 73 - 82.
- Tenorío, M. G. (2003). Evaluación de los recursos naturales suelo y agua del municipio de Xochitlán de Vicente Suárez, Puebla. *Tesis de Maestría. Postgrado de Ciencias Ambientales*. Instituto de Ciencias. Universidad Autónoma de Puebla. Puebla, México.
- Real, J.E. (2001). Escalamiento Multidimensional. Madrid: Editorial La Muralla

STATISTICA. versión 5.1 (1998). Copyright, StatSoft&Inc.

Valera, M. A. Tenorio, M. G. y Linares, G.(2001) Aplicación de indicadores químicos de degradación para suelos ácidos de la Sierra Negra de Puebla. *COLOQUIOS Cuba – México sobre manejo sostenible de los suelos* . pp 57-64. Puebla : BUAP.

# Monte carlo simulations for Rasch model tests

**Patrick Mair<sup>1</sup>**

*Vienna University of Economics*

**Thomas Ledl**

*University of Vienna*

## 1. Introduction

Item response theory (IRT) deals with the study of test and item scores based on assumptions concerning the mathematical relationship between abilities (or other hypothesized traits) and has its origins in the 1950's. Different to the early days, where IRT research was focused on psychological tests only, nowadays, it can be considered as the outstanding psychometric method for contemporary tests and it is not limited only onto psychological issues.

An important role in IRT plays its simplest model, namely, the Rasch model (Rasch, 1960). The main implication of the Rasch model is that the abilities of persons on a certain trait as well as the item difficulties can be compared on an interval scale. Moreover, the comparison of two persons can be performed irrespective of the itemset presented and the remaining persons in the sample. Analogously, two items can be compared irrespective of the persons in the sample and the remaining items in the test. From a philosophy of science point of view this property is called ‘specific objectivity’ (Rasch, 1960 and 1977).

From a technical point of view, the Rasch model is a logistic test model for dichotomous items. The 0/1-response of an individual is described by a logistic function which depends on the person ability and the item difficulties:

$$p(X_{vi} = 1 | \theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}, \quad (1)$$

where  $X_{vi}$  is the binary response of person  $v$  to item  $i$  (i.e.  $X_{vi} \in \{0, 1\}$ ),  $\theta_v$  the trait level of person  $v$  (person parameter), and  $\beta_i$  the difficulty of item  $i$  (item parameter). The parameters are usually estimated either by a conditional maximum likelihood approach (CML; see e.g.

---

<sup>1</sup>[patrick.mair@wu-wien.ac.at](mailto:patrick.mair@wu-wien.ac.at)

Andersen, 1972 and Fischer, 1974) or a marginal maximum likelihood method (MML; see e.g. Glas, 1989). The outstanding feature of the CML approach is that  $\beta$  and  $\theta$  can be estimated independently from each other and, thus, they are separable (see e.g. Andrich, 1988 and Fisher, 1992).

Concerning the formal properties of the Rasch model the following issues are to mention: The measured trait must be unidimensional. It is assumed that response probabilities are a function of a single latent characteristic  $\theta$ . For fixed  $\beta_i$ 's the curves resulting from (1) are called item characteristic curves (ICC) and it is obvious that they cannot cross for different  $\beta_i$ 's. As a consequence, we assume parallel ICC. Furthermore, local stochastic independence is assumed. This implies that item responses are correlated only on account of the trait level  $\theta$ . If  $\theta$  is partialized out, this correlation vanishes. The last property is that the raw score  $R_v$  of each person must be sufficient with respect to the item responses. In other words,  $R_v$  must contain all the information of person  $v$  and hence it is not needed to look at the various 0/1-patterns. Formal derivations and discussions regarding these aspects can be found in Fischer (1995).

## 2. Rasch goodness-of-fit statistics

General discussions pertaining to Rasch model tests are given in Andrich (1988). Specific violations of the properties described above influence these fit statistics. As pointed out by Gustafsson (1980) as well as by Glas and Verhelst (1995), it is ambiguous how global Rasch fit statistics react to such specific violations. For instance, even though from a formal point of view a certain test statistic is suited to detect unidimensionality violations, in practice also non-parallel ICC can affect its result. However, it is not clear how certain test statistics respond to specific violations. As a result, the various impacts have to be an issue of simulation studies. In a recent simulation study by Suárez-Falcón and Glas (2003) the following test statistics were investigated:

- Likelihood ratio test (Andersen, 1973): Persons are splitted into  $G$  groups due to their raw scores and the corresponding  $LR$ -statistic is

$$LR = 2 \left[ \sum_{g=1}^{G-1} \log L_c(\hat{\boldsymbol{\beta}}^{(g)}; \mathbf{X}^{(g)}) - \log L_c(\hat{\boldsymbol{\beta}}; \mathbf{X}) \right] \quad (2)$$

which is  $\chi^2$  distributed with  $df = (G - 1)(G - 2)$ .

- $Q_1$ - and  $Q_2$ -statistic by van den Wollenberg (1982): These statistics are based on differences in the oberserved and Rasch-expected positive response frequencies on item  $i$  ( $i = 1, \dots, K$ ) for each group  $g$ , i.e.  $d_{1gi}^* = m_{1gi} - E(M_{1gi}|\hat{\boldsymbol{\beta}})$  and on item  $i$  and  $j$ , respectively, i.e.  $d_{2gij}^* = m_{2gij} - E(M_{2gij}|\hat{\boldsymbol{\beta}})$ . These terms divided by their standard deviations result in  $z_{1gi}$  and  $z_{2gij}$ .

$$Q_1 = \frac{K-1}{K} \sum_{i=1}^K \sum_{k=1}^{K-1} z_{1gi}^2 \quad (3)$$

$$Q_2 = \frac{K-3}{K-1} \sum_{i=1}^K \sum_{j=i+1}^K \sum_{l=1}^{K-1} z_{2gij}^2 \quad (4)$$

$Q_1 \sim \chi^2$  with  $df = (K - 1)(K - 2)$  and  $Q_2 \sim \chi^2$  with  $df = K(K - 1)(K - 3)/2$ .

- $R_1$ - and  $R_2$ -statistic by Glas (1988): Here,  $Q_1$  and  $Q_2$  are extended by taking into account also the covariance structure of the terms  $d_{1gi} = d_{1gi}^*/\sqrt{n}$  and  $d_{2ij} = d_{2ij}^*/\sqrt{n}$ , without regarding  $g$ . The resulting quadratic forms are

$$R_1 = \sum_{g=1}^{K-1} \mathbf{d}_{1g}^T \mathbf{W}_{1g}^{-1} \mathbf{d}_{1g} \quad (5)$$

$$R_2 = \mathbf{d}_2^T \mathbf{W}_2^{-1} \mathbf{d}_2 - \mathbf{d}_1^T \mathbf{W}_1^{-1} \mathbf{d}_1 \quad (6)$$

$R_1 \sim \chi^2$  with  $df = (k - 1)(k - 1)$  and  $R_2 \sim \chi^2$  with  $df = k(k - 1)/2$

The performance of these test statistic were already simulated by Suárez-Falcón and Glas (2003). In this paper, these simulation studies are enhanced with the classical  $S$ -test from Fischer and Scheiblechner (1970) and the Wald approach  $W$  by Glas and Verhelst (1995) which can be viewed as an approximation of  $S$ : The only difference is that  $S$  uses the standard errors from the information function  $\mathbf{I}(\theta_v)$  of the person parameters (denoted by  $\widehat{\sigma}_{inf}$ ) and  $W$  those from the information matrix in the CML estimation.

$$S = \sum_{i=1}^K \frac{(\widehat{\beta}_{1i} - \widehat{\beta}_{2i})^2}{\widehat{\sigma}_{inf,1i}^2 - \widehat{\sigma}_{inf,2i}^2} \quad (7)$$

$$W = \sum_{i=1}^K \frac{(\widehat{\beta}_{1i} - \widehat{\beta}_{2i})^2}{\widehat{\sigma}_{1i}^2 - \widehat{\sigma}_{2i}^2} \quad (8)$$

Here, the persons are splitted only in 2 groups. Fischer and H. Scheiblechner (1970) claim that the components to be summed up are not completely independent from each other. Nevertheless, a  $\chi^2$  approximation with  $df = K - 1$  is proposed. Corresponding simulation studies can be found in Mair (2006) whereas further test statistics and explanations are elaborated in Glas and Verhelst (1995).

### 3. Monte Carlo simulations

The Monte Carlo simulation issues refer to the following conditions and scenarios, respectively: After having studied the behavior of the Type I error rate, specific Rasch violations are presented. Within these violations, certain scenarios concerning the number of items and persons in the samples are simulated and the power of certain test statistics against these violations is discussed. The basic routine for simulation is the eRm package of Hatzinger and Mair (2006) written in R. Note that due to computational issues for large number of items,  $R_2$  was not computable.

### 3.1. Type I Error Rates

Here, Rasch homogeneous 0/1-matrices are produced and the number of  $H_0$ -rejections is counted. Of course, this number should be small with respect to a well performing test. The results are given in Table 1.

Cuadro 1: Type I error rates

K	N	$LR$	$Q_1$	$Q_2$	$R_1$	$R_2$	$S$	$W$
15	100	.04	.06	.48	.06	.08	.21	.11
	250	.04	.04	.17	.04	.07	.25	.10
	500	.05	.04	.07	.04	.06	.19	.10
	1000	.05	.05	.07	.05	.06	.19	.14
	4000	.04	.04	.05	.05	.06	.20	.13
50	100	.03	.05	-	.05	-	.21	.14
	250	.06	.04	.51	.04	-	.22	.16
	500	.06	.07	.21	.07	-	.22	.13
	1000	.06	.06	.14	.06	-	.19	.09
	4000	.04	.05	.13	.05	-	.18	.09
75	100	.01	.05	-	.05	-	.23	.15
	250	.05	.05	.67	.05	-	.20	.16
	500	.05	.06	.36	.06	-	.22	.11
	1000	.05	.04	.17	.04	-	.19	.10
	4000	.04	.05	.22	.05	-	.20	.08

None of the tests is strongly affected by the number of persons  $n$  nor by the number of items  $K$ . The values within a column are fairly constant over the rows. It can be stated that  $LR$ ,  $Q_1$ , and  $R_1$  hold the  $\alpha$ -level whereas  $S$  and  $W$  are slightly above.

### 3.2. Non-parallel ICC

Non-parallel ICC are simulated by imposing an additional item discrimination parameter  $\alpha_i$  into 1. This model corresponds to the 2-PL Birnbaum (1968) and  $\alpha_i$  is drawn from a log-normal distribution with  $\mu = 0$  and  $\sigma = 0.25$  which corresponds to a medium violation (Suárez-Falcón and Glas, 2003).

Cuadro 2: Parallel ICC violation

K	N	LR	$Q_1$	$Q_2$	$R_1$	$R_2$	S	W
15	100	.14	.15	.46	.15	.11	.47	.14
	250	.40	.41	.14	.41	.19	.61	.41
	500	.71	.71	.09	.72	.32	.74	.52
	1000	.93	.93	.07	.93	.61	.83	.85
	4000	1.00	1.00	.07	1.00	.99	.97	1.00
50	100	.29	.40	-	.40	-	.62	.34
	250	.88	.87	.63	.87	-	.80	.52
	500	1.00	1.00	.28	1.00	-	.85	.86
	1000	1.00	1.00	.18	1.00	-	.95	.99
	4000	1.00	1.00	.16	1.00	-	.99	1.00
75	100	.41	.59	-	.60	-	.65	.29
	250	.97	.97	.72	.97	-	.78	.62
	500	1.00	1.00	.46	1.00	-	.91	.98
	1000	1.00	1.00	.27	1.00	-	.99	1.00
	4000	1.00	1.00	.19	1.00	-	1.00	1.00

The results in Table 2 suggest that  $LR$ ,  $Q_1$ ,  $R_1$ ,  $R_2$ ,  $S$ , and  $W$  perform well with respect to a large number of items and thus they are able to detect the deviation from parallel ICC.

### 3.3. Non-unidimensional data

This kind of violation is simulated by using a multidimensional Rasch model (Glas, 1989 and 1992). A correlation parameter  $r_{\theta_1, \theta_2}$  is introduced in order to steer the strength of violation. Due to Suárez-Falcón and Glas (2003), a medium violation corresponds to  $r_{\theta_1, \theta_2} = .50$ . Table 3 suggests that  $Q_2$  and  $R_2$  have the highest statistical power. For the other test statistics it can be stated that the sample size has to be fairly large in order to achieve satisfying results.

### 3.4. Locally dependent data

Finally, the violation of local independence is simulated by using a model proposed by Jannarone (1986) which imposes a dependency parameter  $\delta_{ij}$  for pairwise dependent items  $i$

and  $j$ . A value of  $\delta_{ij} = .50$  can be regarded as medium violation. This leads to the results in Table 4. It is not surprising that only  $R_2$  and  $Q_2$  have an acceptable power since only these statistics formally impose the pairwise solving frequencies as can be seen in 6 and 4. The remaining tests are not able to detect derivations from local item independence.

## 4. Discussion

The question of model fit in Rasch modeling is crucial in the sense that usually the Rasch model is taken for granted whereas the researcher must ‘produce’ appropriate data. The term ‘appropriate data’ means that in the first instance he must find a set of items which is Rasch homogeneous. In practice, the Rasch assumption is rather restrictive and the sources of deviation can be: Firstly, items which do not refer to only one latent trait; secondly, items which produce non-parallel curves; thirdly, items whose answer patterns are not locally independent. Hence, in order to apply global fit statistics it is inevitable to know, whether they are able to detect specific violations. Thus, corresponding power analyses were performed and the results of the last subsections are summarized in Table 5.

Cuadro 3: Violation of unidimensionality

K	N	<i>LR</i>	<i>Q</i> <sub>1</sub>	<i>Q</i> <sub>2</sub>	<i>R</i> <sub>1</sub>	<i>R</i> <sub>2</sub>	<i>S</i>	<i>W</i>
15	100	.07	.07	-	.08	.37	.17	.08
	250	.13	.13	.50	.13	.83	.25	.11
	500	.23	.23	.86	.23	1.00	.28	.19
	1000	.52	.52	.99	.52	1.00	.44	.45
	4000	.98	.98	1.00	.98	1.00	.79	.85
50	100	.06	.09	-	.09	-	.12	.11
	250	.17	.16	.95	.16	-	.19	.21
	500	.26	.26	1.00	.26	-	.20	.29
	1000	.54	.54	1.00	.54	-	.48	.61
	4000	.99	.99	1.00	.99	-	.77	.80
75	100	.07	.12	-	.12	-	.18	.15
	250	.25	.22	.98	.23	-	.21	.29
	500	.41	.42	1.00	.42	-	.38	.43
	1000	.73	.73	1.00	.72	-	.68	.76
	4000	1.00	1.00	1.00	1.00	-	.87	.88

In sum it can be stated that for detecting local dependence, special test statistics as  $R_2$  and  $Q_2$  have to be consulted. For the remaining violations Andersen's classical  $LR$ -statistic performs rather well. The advantage of this approach is that it can be readily applied to more general IRT models such as rating scale models (RSM; Andrich, 1978), partial credit models (PCM; Masters, 1982) and their linear extensions, i.e. LRSM (Fischer and Parzer, 1991) and LPCM (Fischer and Ponocny, 1994). A corresponding implementation of such models is given in the eRm package.

Cuadro 4: Violation of local independence

K	N	LR	$Q_1$	$Q_2$	$R_1$	$R_2$	S	W
15	100	.02	.04	-	.06	.10	.09	.14
	250	.04	.05	.30	.05	.42	.11	.10
	500	.05	.07	.40	.07	.83	.09	.11
	1000	.09	.10	.80	.10	1.00	.15	.13
	4000	.23	.25	1.00	.26	1.00	.21	.23
50	100	.01	.09	-	.09	-	.10	.08
	250	.02	.04	.66	.04	-	.10	.12
	500	.03	.02	.58	.02	-	.12	.11
	1000	.03	.02	.58	.02	-	.12	.16
	4000	.15	.14	1.00	.15	-	.15	.24
75	100	.04	.12	-	.12	-	.09	.09
	250	.01	.04	.64	.04	-	.09	.13
	500	.06	.07	.67	.07	-	.11	.13
	1000	.10	.10	.79	.10	-	.20	.12
	4000	.20	.22	1.00	.22	-	.19	.15

Cuadro 5: Test performance overview

Violation	Good Performance	Medium/Weak Performance
None (Rasch data)	$LR, Q_1, R_1, R_2$	$S, W, Q_2$
Parallel ICC	$LR, Q_1, R_1, S, W$	$Q_2, R_2$
Unidimensionality	$LR, Q_1, Q_2, R_1, R_2$	$S, W$
Local Independence	$Q_2, R_2$	$LR, Q_1, R_1, S, W$

## References

- Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, **34**, 42-54.
- Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, **38**, 123-140.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, **43**, 561-573.
- Andrich, D. (1988). *Rasch Models for Measurement*. Sage. Newbury Park, CA.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores* (edited by F.M. Lord and M.R. Novick), 395-479. Addison-Wesley, Reading, MA.
- Fischer, G.H. (1974). Einführung in die Theorie psychologischer Tests (Introduction to the theory of psychological tests). Huber, Bern.
- Fischer, G.H. (1995). Derivations of the Rasch Model. *Rasch Models: Foundations, Recent Developments, and Applications* (edited by G.H. Fischer and I.W. Molenaar), 15-38. Springer, New York.
- Fischer, G.H. and Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of treatment effects. *Psychometrika*, **56**, 637-651.
- Fischer, G.H. and Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, **59**, 177-192.
- Fischer, G.H. and Scheiblechner, H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch (Algorithms and programs Rasch's probabilistic test model). *Psychologische Beiträge*, **12**, 23-51.

Fisher Jr., W.P. (1992). Objectivity in measurement: A philosophical history of Rasch's separability theorem. *Objective Measurement: Foundations, Recent Developments, and Applications* (edited by M. Wilson), 29-60. Ablex, Norwood, NJ.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, **53**, 525-546.

Glas, C.A.W. (1989). Contributions to estimating and testing Rasch model. *Doctoral Thesis*, University of Twente, Enschede.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. *Objective Measurement: Foundations, Recent Developments, and Applications* (edited by M. Wilson), 236-258. Ablex, Norwood, NJ.

Glas, C.A.W. and Verhelst, N.D. (1995). Testing the Rasch model. *Rasch Models: Foundations, Recent Developments, and Applications* (edited by G.H. Fischer and I.W. Molenaar), 69-96. Springer, New York.

Gustafsson, J.E. (1980). Testing and obtaining the fit of data to the Rasch model. *Journal of Mathematical and Statistical Psychology*, **33**, 205-223.

Hatzinger, R. and Mair, P. (2006). eRm - Extended Rasch modelling: An R package for the application of item response theory models. In preparation

Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika*, **451**, 357-373.

Mair, P. (2006). Simulation Studies for Goodness-of-Fit Statistics in Item Response Theory. *Diploma Thesis*, University of Vienna, Vienna.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149-174.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Rasch, G. (1977). On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, **14**, 58-94.

Suárez-Falcón, J.C. and Glas, C.A.W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, **56**, 127-143.

van den Wollenberg, A.L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, **47**, 123-140.



# **Clasificación multivariada usando algoritmos genéticos y la función lineal discriminante de Fisher**

**Julia Aurora Montano Rivas**  
**Sergio Francisco Juárez Cerrillo**<sup>1</sup>

*Universidad Veracruzana*

## **1. Introducción**

Montano y Cantú (2005) desarrollaron un procedimiento basado en algoritmos genéticos, el cual, según evidencia empírica, mejora la capacidad de discriminación de la función lineal discriminante de Fisher (FLDF). La idea fundamental de la propuesta es transformar a la FLDF en otra función lineal, mediante operadores genéticos, con el objetivo de mejorar la discriminación de la FLDF en los datos. Motivados por éste hecho, en éste trabajo presentamos los resultados de un experimento de simulación Monte Carlo diseñado para evaluar empíricamente el desempeño de la propuesta en la solución del problema de clasificación. El artículo tiene la siguiente estructura. En la Sección 2 presentamos brevemente la FLDF. En la Sección 3 describimos, también brevemente, la propuesta de Montano y Cantú (2005). Las características del experimento de simulación las damos en la Sección 4. En la Sección 5 presentamos los resultados obtenidos. Observamos que la FLDF fue superior a las funciones lineales producidas por el algoritmo genético para clasificar. En la Sección 6 concluimos el trabajo con una discusión de las posibles razones por las cuales sucede ésto. También identificamos posibles líneas para investigación futura.

---

<sup>1</sup>[sejuarez@uv.mx](mailto:sejuarez@uv.mx)

## 2. La función lineal discriminante de Fisher

Sean  $\mathbf{x}_{11}, \mathbf{x}_{21}, \dots, \mathbf{x}_{n_1 1}$  y  $\mathbf{x}_{12}, \mathbf{x}_{22}, \dots, \mathbf{x}_{n_2 2}$  muestras aleatorias de las distribuciones  $N_p(\mu_1, \Sigma)$  y  $N_p(\mu_2, \Sigma)$ , respectivamente, y sea  $\mathbf{x}_0$  una observación de la cual sólo se sabe que proviene de alguna de éstas distribuciones. El problema de clasificación consiste en determinar de cuál de estas distribuciones proviene  $\mathbf{x}_0$ . El análisis discriminante de Fisher resuelve este problema clasificando a  $\mathbf{x}_0$  como proveniente de la  $N_p(\mu_1, \Sigma)$  si  $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_c^{-1} \mathbf{x}_0 \geq m$  y clasificando a  $\mathbf{x}_0$  como proveniente de la  $N_p(\mu_2, \Sigma)$  si  $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_c^{-1} \mathbf{x}_0 < m$ , donde

$$m = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_c^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) / 2, \quad \bar{\mathbf{x}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_{i1}, \quad \bar{\mathbf{x}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{x}_{i2},$$

$$\mathbf{S}_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\mathbf{x}_{i1} - \bar{\mathbf{x}}_1) (\mathbf{x}_{i1} - \bar{\mathbf{x}}_1)^T, \quad \mathbf{S}_2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (\mathbf{x}_{i2} - \bar{\mathbf{x}}_2) (\mathbf{x}_{i2} - \bar{\mathbf{x}}_2)^T,$$

y

$$\mathbf{S}_c = \frac{(n_1 - 1) \mathbf{S}_1 + (n_2 - 1) \mathbf{S}_2}{n_1 + n_2 - 2}.$$

La FLDF se define por  $l(\mathbf{x}) = (\mu_1 - \mu_2)^T \Sigma^{-1} \mathbf{x}$  y es óptima en el sentido de que alcanza la *tasa óptima de error* de clasificación, TOE, la cual resulta ser

$$\text{TOE} = p \int_{R_1} f_1(\mathbf{x}) d\mathbf{x} + (1 - p) \int_{R_2} f_2(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{-\Delta/2} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx, \quad (1)$$

donde  $f_1$  es la densidad de la  $N(\mu_1, \Sigma)$  y  $f_2$  es la densidad de la  $N(\mu_2, \Sigma)$ ,  $p$  es la probabilidad apriori de que una observación provenga de la  $N(\mu_1, \Sigma)$ , de modo que  $1 - p$  es la probabilidad de que provenga de la  $N(\mu_2, \Sigma)$ ,  $\Delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$ ; y

$$R_1 = \{ \mathbf{x} \in \mathbb{R}^p : (\mu_1 - \mu_2)^T \Sigma^{-1} \mathbf{x} - (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 + \mu_2) / 2 \geq 0 \},$$

y  $R_2 = \mathbb{R}^p \setminus R_1$ . En nuestra exposición hemos seguido Johnson y Wichern (1998).

### 3. Algoritmos Genéticos

La propuesta de Montano y Cantú (2005) consiste en lo siguiente:

1. Se calcula la FLDF muestral en base a los datos de entrenamiento.
2. Se construye con remuestreo una población inicial de  $N$  seudomuestras a partir de los datos de entrenamiento. Se calcula la FLDF de cada seudomuestra junto con sus respectivos errores de mala clasificación. En este trabajo usamos  $N = 100$ .
3. Se eligen a las  $k$  FLDF muestrales con menor error de discriminación. Estas funciones se mutan usando los operadores de cruce aritmético, mutación con distribución normal y mutación con distribución uniforme. En este trabajo usamos  $k = 20$ . Con esto se produce un nuevo conjunto de funciones lineales discriminantes.
4. Las funciones lineales resultantes se evalúan con los datos de entrenamiento y de prueba para obtener sus errores de mala discriminación y mala clasificación. El procedimiento termina si las funciones lineales tienen errores de clasificación menores o iguales que la TOE en (1) o si ya se han realizado  $M$  iteraciones sin éxito. Nosotros usamos  $M = 10$  y elegimos a las funciones lineales de mutación normal (FLMN) y uniforme (FLMU) con menores errores de clasificación. De otro modo el procedimiento se itera a partir del paso 3.

Los detalles de la propuesta, el cruce aritmético, y las mutaciones normal y uniforme se pueden ver en Montano y Cantú (2005). Un programa en S-plus que implementa la propuesta está disponible con el primer autor.

### 4. Experimento de Simulación

Consideremos muestras aleatorias de tamaño  $n_1$  de la  $N(\mu_1, \Sigma)$  y de tamaño  $n_2$  de la  $N(\mu_2, \Sigma)$ . Nosotros elegimos  $n_1 = n_2 = n$  con  $n$  en  $\{20, 50, 100\}$ . Al vector  $\mu_1$  lo fijamos

en  $(0, 0)^T$  y para  $\mu_2$  consideramos  $(1, 1)^T$  y  $(-1, 1)^T$ . Para la matriz de varianzas y covarianzas usamos

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

con  $\rho$  en  $\{0, 0.25, 0.50, 0.75, 0.95\}$ . Para cada combinación de  $n$ ,  $\mu_2$ , y  $\rho$  generamos muestras aleatorias de entrenamiento de tamaño  $n$  de cada una de las poblaciones  $N(\mu_1, \Sigma)$  y  $N(\mu_2, \Sigma)$ . Estimamos la FLDF y las funciones lineales discriminantes del algoritmo genético. Para realizar la clasificación generamos 500 observaciones de cada una de las poblaciones  $N(\mu_1, \Sigma)$  y  $N(\mu_2, \Sigma)$ . Para comparar a las FLDF muestrales con las funciones producidas supongamos que tenemos dos clasificadores  $A$  y  $B$ . Sea  $n_A$  el número de errores hechos por  $A$  y no por  $B$  y viceversa para  $n_B$ . Bajo la hipótesis nula de que ambos clasificadores tienen la misma tasa de error, el estadístico de prueba de McNemar  $z = (|n_A - n_B| - 1)/\sqrt{n_A + n_B}$  tiene aproximadamente una distribución  $N(0, 1)$  (véase Ripley, 1996, página 77).

## 5. Resultados

Vemos que para el vector de medias  $(1, 1)^T$  la FLDF clasifica mejor que FLMN y FLMU cuando  $n=20$  y  $\rho = 0, 0.25$  y cuando  $n = 100$  y  $\rho = 0$ . En los demás casos la FLMU es mejor que la de Fisher y en ocho de los casos es significativa. Para el vector de medias  $(-1, 1)^T$  en siete de quince casos la FLDF es mejor que las funciones lineales producidas por los algoritmos genéticos y en cinco ocasiones la FLDF es mejor de manera significativa.

## 6. Conclusión e investigación adicional

Los algoritmos genéticos pertenecen a la clase de algoritmos de optimización llamados *evolutivos*. Estos algoritmos buscan aleatoriamente soluciones óptimas dentro del espacio de soluciones usando principios de la teoría Darwiniana de la evolución. Con esta idea en mente, es que formulamos la hipótesis de que, iniciando a partir de la FLDF muestral, las funciones FLMN y/o FLMU se *acercarían* a la función lineal discriminante con la tasa de error más pequeña, es decir, a la FLDF  $l(\mathbf{x}) = (\mu_1 - \mu_2)^T \Sigma^{-1} \mathbf{x}$ . Sin embargo, aparentemente este no es el caso.

Cuadro 1: Las primeras tres columnas son los errores de clasificación. Las siguientes dos columnas muestran el nivel de significancia observado (*p-values*) de la prueba de McNemar para comparar a la FLDF con la FLMN y la FLMU.

	$\rho$	FLDF	FLMN	FLMU	FLDF vs FLMN	FLDF vs FLMU
$\mu_1 = (1, 1)^T$	0	0.243	0.247	0.247	0.0000	0.0000
	0.25	0.249	0.251	0.251	0.0004	0.0004
	0.50	0.308	0.310	0.303	0.0000	0.0003
	0.75	0.288	0.287	0.282	0.5000	0.2275
	0.95	0.309	0.299	0.293	0.3707	0.0004
$\mu_1 = (1, 1)^T$	0	0.250	0.246	0.244	0.0392	0.0039
	0.25	0.304	0.285	0.284	0.0002	0.0071
	0.50	0.275	0.274	0.273	0.0071	0.0158
	0.75	0.316	0.309	0.309	0.0008	0.0001
	0.95	0.303	0.298	0.295	0.0001	0.0002
$\mu_1 = (1, 1)^T$	0	0.249	0.252	0.251	0.0018	0.0002
	0.25	0.271	0.265	0.259	0.0000	0.0002
	0.50	0.272	0.272	0.266	0.0003	0.0206
	0.75	0.285	0.280	0.279	0.0001	0.0001
	0.95	0.304	0.301	0.299	0.0082	0.4297
$\mu_1 = (-1, 1)^T$	0	0.259	0.256	0.254	0.0000	0.0013
	0.25	0.212	0.221	0.222	0.0029	0.0000
	0.50	0.152	0.153	0.145	0.2266	0.0011
	0.75	0.085	0.078	0.076	0.5000	0.5000
	0.95	0.003	0.003	0.002	0.1251	0.5000
$\mu_1 = (-1, 1)^T$	0	0.228	0.224	0.226	0.0001	0.0005
	0.25	0.209	0.223	0.221	0.0013	0.0025
	0.50	0.152	0.152	0.151	0.0018	0.0227
	0.75	0.088	0.091	0.093	0.0046	0.0009
	0.95	0.001	0.001	0.003	0.5000	0.2419
$\mu_1 = (-1, 1)^T$	0	0.237	0.238	0.239	0.0484	0.2118
	0.25	0.195	0.191	0.193	0.0003	0.0007
	0.50	0.156	0.155	0.153	0.2118	0.4129
	0.75	0.097	0.100	0.100	0.0003	0.0001
	0.95	0.001	0.001	0.002	0.5000	0.1251

Bajo el supuesto de normalidad multivariada, la FLDF muestral  $\hat{l}(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_c^{-1} \mathbf{x}$  converge en probabilidad a la verdadera FLDF, por lo que el error de clasificación de la FLDF muestral debe ser cercano a la verdadera TOE. De acuerdo a esto, para mejorar el poder de clasificación de la FLDF muestral la opción sigue siendo apelar a su consistencia y entonces disponer de más observaciones.

A pesar de los resultados arrojados por este primer estudio de simulación, consideramos que aún no hemos agotado a los algoritmos genéticos como una alternativa para mejorar a la FLDF muestral. El uso de otros operadores genéticos así como redefinir el espacio de soluciones son, por mencionar algunas, líneas que pensamos merecen investigación adicional.

## Referencias

- Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Analysis. Fourth edition.* Prentice Hall: Upper Saddle River.
- Montano Rivas, J.A. y Cantú Sifuentes, M. (2005). Una Aplicación de los Algoritmos Genéticos en la Discriminación. *Revista Agraria Nueva Época.* Año 1 1, 42-47.
- Ripley, B.D. (1996). *Pattern Recognition and Neural Networks.* Cambridge: University Press.

# **La política monetaria en el periodo 1984-2004. Confrontación teórica-práctica**

**Federico Ricardo Muller Rodríguez**

*Facultad de Economía, Universidad Autónoma de Coahuila*

**Adrián Guerrero Arévalo<sup>1</sup>**

*Facultad de Economía, Universidad Autónoma de Coahuila*

**Mónica Rodríguez Soria**

*Facultad de Economía, Universidad Autónoma de Coahuila*

## **1. Introducción**

Uno de los instrumentos de política económica que el gobierno mexicano puede utilizar, sin requerir la aprobación del Congreso de la Unión para su aplicación, es la política monetaria. Mediante el manejo de la oferta monetaria puede acelerar, mantener o inhibir el crecimiento de la economía, exceptuando el último aspecto, el resultado indudablemente debe manifestarse en mayores inversiones, más ingresos y mejores oportunidades para la sociedad, de ahí que el manejo de ésta herramienta represente una gran responsabilidad para las autoridades monetarias del país por los efectos que tiene sobre el desarrollo.

## **2. Objetivo**

Analizar el comportamiento de la oferta monetaria en México en el período de 1984 a 2004, a través de formación de un modelo estadístico, que permita identificar sus efectos en los agregados macroeconómicos que son claves en el crecimiento económico.

---

<sup>1</sup>acuario\_16@hotmail.com

### **3. Desarrollo**

El eje conductor de la investigación es confrontar las teorías que explican las políticas monetarias y la información sobre el agregado monetario que pública el Banco de México. La parte operativa se trabaja con un modelo econométrico que permite identificar la relación causa-efecto de las siguientes variables:

1. Oferta Monetaria (OM).
2. Inflación.
3. Tasa de Desempleo Abierto (TDA).
4. Tasa de Interés Interbancaria de Equilibrio (TIIE).
5. Certificados de la Tesorería de la Federación (CETES).
6. Índice de Precios al Producto (IPP).
7. Exportaciones.
8. Importaciones.
9. Producto Interno Bruto (PIB).

Con la ayuda del programa *e-views* se despliegan los modelos lineales a base de logaritmos, cada variable se compara con la oferta monetaria, la que opera inicialmente como variable dependiente y después como independiente; y así se llega a una primera discriminación de variables. Posteriormente se convierten los saldos de OM a porcentajes para armonizarlos con las variables que están dadas en tasas, las que son:

- CETES.
- IPP.
- TDA.
- TIIE.

La segunda selección se realiza en función de la volatilidad y de la significancia de las variables. Finalmente las variables obtenidas se someten a las pruebas de Durbin-Watson, autocorrelación y heteroelasticidad.

Con efecto de lo anterior, el Producto Interno Bruto (PIB), la inflación, las exportaciones y la Tasa de Interés Interbancaria de Equilibrio (TIIE) son los agregados macroeconómicos que superaron las pruebas estadísticas a las que fueron sometidos.

Una vez obtenidos los resultados del modelo se contrastan con los postulados teóricos y se permite una conclusión.

## 4. Resultados

Para estudiar relaciones entre las variables mencionadas, utilizamos las siguientes herramientas.

### 4.1. Diagramas de dispersión

Las Figuras 1-2 muestran diagramas de dispersión para estudiar las relaciones entre: PIB y OM (Figura 1 a la izquierda), Inflación y OM (Figura 1 a la derecha), Exportaciones y OM (Figura 2 a la izquierda) y finalmente TIIE y OM (Figura 2 a la derecha).

### 4.2. Modelos Econométricos

Los siguientes modelos fueron ajustados para estudiar las relaciones enunciadas

Dependet Variable: OM

Method: Least Squares

Date: 04/24/05 Time: 23:11

Sample: 1984-2004

Included observations: 21

Variable	Coefficient	Std.Error	t-Statistic	Prob
PIB	5.012551	0.216771	23.12374	0.00000
R-square	0.918552	Mean dependent var	13491693	
Adjusted R-squared	0.918552	S.D.dependent var	12321097	
S.E. ofregression	3516332	Akaike info criterion	33.03018	
Sum squared resid	2.47e+14	Schwarz criterion	33.07992	
Log Likelihood	-345.8169	Durbin-Watson stat	2.117276	

Dependet Variable: LOG(OM)

Method: Least Squares

Date: 04/25/05 Time: 20:48

Sample: 1984-2004

Included observations:21

Variable	Coefficient	Std.Error	t-Statistic	Prob
C	16.44286	0.334689	49.12881	0.0000
INFLACION	-0.017919	0.006673	-2.685246	0.0146
R-square	0.275101	Mean dependent var	15.83000	
Adjusted R-squared	0.236948	S.D.dependent var	1.284228	
S.E. ofregression	1.121810	Akaike info criterion	3.158156	
Sum squared resid	23.91069	Schwarz criterion	3.257635	
Log Likelihood	-31.16064	F-statistic	7.210544	
Durbin-Watson stat	0.860765	Prob(F-statistic)	0.014848	

Dependet Variable: OM

Method: Least Squares

Date: 04/25/05 Time: 20:45

Sample: 1984-2004

Included observations: 21

Variable	Coefficient	Std.Error	t-Statistic	Prob
C	-8240447	2347063	-3.510960	0.0023
EXPORTACIONES	497.7780	47.88642	10.39497	0.0000
R-square	0.850459	Mean dependent var	13491693	
Adjusted R-squared	0.842588	S.D.dependent var	12321097	
S.E. ofregression	4888412	Akaike info criterion	33.73303	
Sum squared resid	4.54e+14	Schwarz criterion	33.83250	
Log Likelihood	-352.1968	F-statistic	108.0554	
Durbin-Watson stat	1.628243	Prob(F-statistic)	0.000000	

Dependet Variable: LOG(OM)

Method: Least Squares

Date: 04/15/05 Time: 00:07

Sample: 1984-2004

Included observations: 21

Variable	Coefficient	Std.Error	t-Statistic	Prob
C	19.63842	0.877616	22.37700	0.0000
LOG(TIIE)	-1.166641	0.261695	-4.458014	0.0003
R-square	0.511240	Mean dependent var	15.83000	
Adjusted R-squared	0.485518	S.D.dependent var	1.204229	
S.E. ofregression	0.921146	Akailke info criterion	2.763995	
Sum squared resid	16.12167	Schwarz criterion	2.863474	
Log Likelihood	-27.02195	F-statistic	19.87389	
Durbin-Watson stat	0.741297	Prob(F-statistic)	0.000270	

## 5. Conclusiones

Una vez obtenidos los resultados del modelo se contrastan con los postulados teóricos y se emiten las siguientes conclusiones:

- La relación entre OM y PIB es positiva; asociación que confirma los postulados teóricos.
- La asociación entre OM e inflación es negativa; resultado que es contrario a los argumentos teóricos.
- La correlación entre OM y exportaciones es positiva, tal como lo marca la teoría.
- La TIIE se asocia de manera negativa con la OM, relación que afirma la teoría.

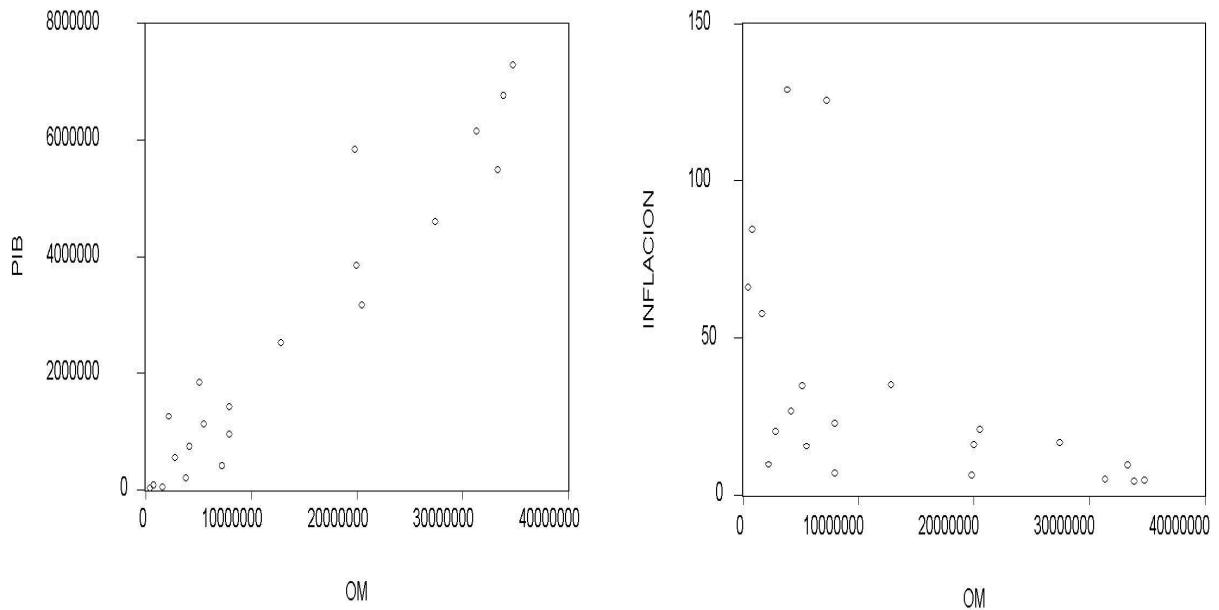


Figura 1: Diagramas de dispersión: PIB-OM (izquierda) Inflación-OM (derecha).

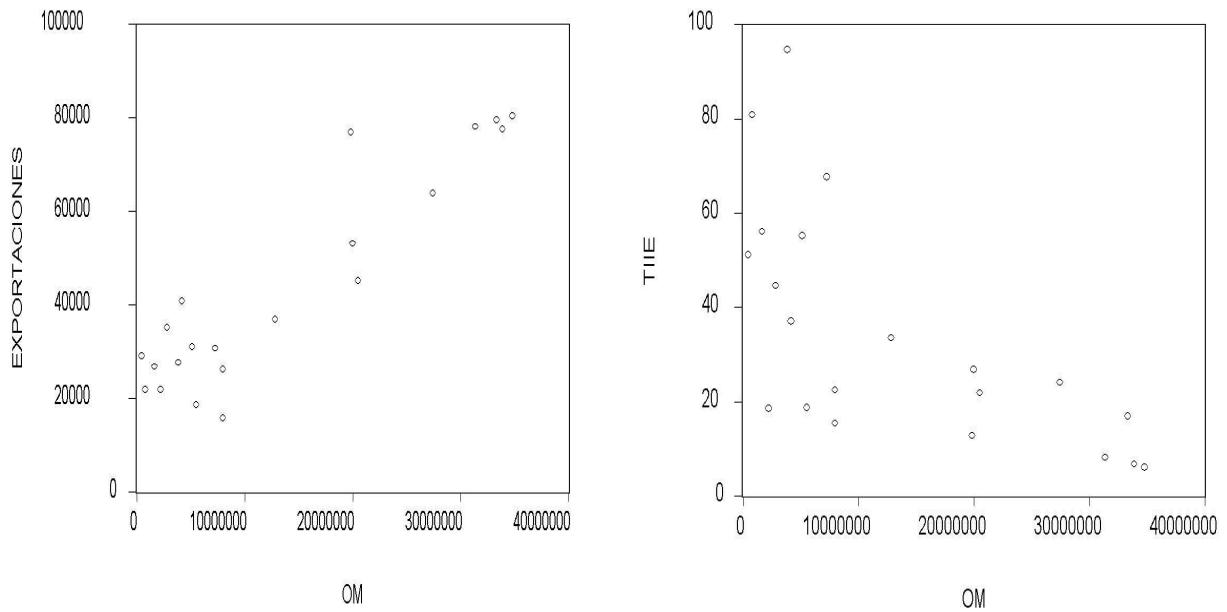


Figura 2: Diagramas de dispersión: Exportaciones-OM (izquierda) TIIE-OM (derecha).

## Referencias

- Case, K. E. (1997). *Principios de Microeconomía*. Prentice Hall. Cuarta Edición.
- Dornbusch, R. y Fischer, S. (1994). *Macroeconomía*. Mc. Graw Hill, Sexta Edición.
- Gujarati, D. N. (1998). *Econometría*. Mc. Graw Hill. Cuarta Edición.
- Parkin, M. y Esquivel, G. (2001). *Macroeconomía: versión para Latinoamérica*. Quinta Edición.



# Análisis Bayesiano de un modelo de regresión para datos circulares

Gabriel Nuñez Antonio<sup>1</sup>

*Instituto Tecnológico Autónomo de México*

Eduardo Gutiérrez Peña<sup>2</sup>

*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas*

## 1. Naturaleza de los datos direccionales

Los datos direccionales aparecen en varias áreas de manera natural y son especialmente comunes en las ciencias biológicas, geofísicas y meteorológicas.

La representación gráfica de este tipo de datos son puntos sobre la circunferencia de un círculo y en general puntos sobre la superficie de una esfera. La aplicación de técnicas lineales convencionales puede producir paradojas, dada la periodicidad inherente del círculo y la diferente topología del círculo y la línea recta.

Los datos direccionales aparecen a menudo en modelos de regresión como la variable de respuesta. Los modelos de regresión propuestos en la literatura para una respuesta circular sufren de problemas que los vuelven imprácticos para el análisis de tales datos. Esta dificultad para aplicar la metodología es especialmente frustrante cuando se contrasta con la metodología de los modelos lineales (generalizados) para el análisis de una respuesta escalar.

## 2. La distribución Normal proyectada

En el caso  $k$ -dimensional una dirección se puede representar como un vector unitario  $\mathbf{U}$  en  $\mathbb{R}^k$ . Si  $\mathbf{U}$  es una dirección aleatoria en  $\mathbb{R}^k$ , su dirección media está determinada por el vector unitario  $\eta = E[\mathbf{U}]/\rho$ , donde  $\rho = \|E[\mathbf{U}]\|$ .

---

<sup>1</sup>gabriel@itam.mx

<sup>2</sup>eduardo@sigma.iimas.unam.mx

Se dice que un vector unitario  $k$ -dimensional  $\mathbf{U} = \mathbf{Y}/R$ , donde  $R = \|\mathbf{Y}\|$ , tiene una distribución Normal proyectada  $k$ -variada,  $NP(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ , si  $\mathbf{Y}$  tiene una distribución  $N_k(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ .

Se debe notar que dado que  $\mathbf{U}$  es un vector unitario en  $\mathbb{R}^k$ , este se puede determinar especificando  $k - 1$  ángulos.

### 3. Modelo de regresión basado en la Normal proyectada

Algunos modelos de regresión propuestos en la literatura para una respuesta circular se pueden revisar, por ejemplo, en Fisher y Lee (1992). Estos autores asumen una distribución von Mises para la variable de respuesta. Más recientemente, en Presnell *et al.* (1998) se propone un modelo basado en proyecciones. Este modelo de regresión normal proyectado trata las respuestas direccionales como proyecciones, sobre el círculo o la esfera unitaria, de los vectores de respuesta en un modelo lineal multivariado.

Especificamente, sea  $\mathbf{U}_i = \mathbf{Y}_i/R$ , con  $R = \|\mathbf{Y}_i\| i = 1, \dots, n$ , un vector unitario donde las  $\mathbf{Y}_i \sim N_k(\boldsymbol{\mu}_i, \boldsymbol{\Lambda})$  son v.a. independientes, con  $\boldsymbol{\mu}_i = \mathbf{B}^t \mathbf{x}_i$  donde  $\mathbf{B}_{k \times p}^t$  es una matriz con  $j$ -ésimo renglón el vector de parámetros  $(\boldsymbol{\beta}^j)$  y  $\mathbf{x}_i$  un vector de  $p - 1$  covariables de dimensión  $p$ . De esta manera, los componentes de  $\boldsymbol{\mu}_i$  resultan ser:

$$\boldsymbol{\mu}_i^j = \mathbf{x}_i^t \boldsymbol{\beta}^j, \quad \forall j = 1, \dots, k$$

El objetivo es realizar inferencias sobre  $\mathbf{B}$  y  $\boldsymbol{\Lambda}$ . Sin embargo,  $\mathbf{B}$  y  $\boldsymbol{\Lambda}$  no son identificables, dado que, para cualquier  $a > 0$ ,  $NP(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = NP(a\boldsymbol{\mu}, \boldsymbol{\Lambda}/a^2)$ .

Una forma de evitar este problema es tomar  $\boldsymbol{\Lambda} = \mathbf{I}$ . Modelos con estructuras más generales de  $\boldsymbol{\Lambda}$  pueden resultar más atractivos, ya que entonces la densidad  $NP(\boldsymbol{\mu}, \boldsymbol{\Lambda})$  puede ser asimétrica y/o bimodal. En el resto del trabajo nos enfocaremos al caso  $\boldsymbol{\Lambda} = \mathbf{I}$ .

## 4. Análisis Bayesiano del Modelo de Regresión

Una revisión del modelo de regresión para variables de respuesta circular desde un enfoque clásico se puede encontrar en Presnell *et al.* (1998). Adicionalmente, algunas aplicaciones se pueden revisar, por ejemplo, en D'Elia *et al.* (1999).

En este trabajo se presenta un análisis Bayesiano del modelo  $\mathbf{Y} \sim N_k(\mathbf{B}^t \mathbf{x}, \mathbf{I})$  usando una inicial conjugada. Se deriva la correspondiente distribución final vía la introducción de variables latentes y la implementación de métodos MCCM (Gibbs sampling - Metropolis Hastings). Hay que mencionar que este análisis es una aplicación del modelo Normal bajo proyecciones analizado en Nuñez-Antonio y Gutiérrez-Peña (2005).

### 4.1. Especificación del modelo

Sea  $\mathbf{Y} \sim N_k(\cdot | \mathbf{B}^t \mathbf{x}, \mathbf{I})$ . Vía coordenadas polares se puede obtener la densidad conjunta de  $(\Theta, R)$ ,  $f_{(\Theta, R)}(\theta, r)$ , con  $R = \|\mathbf{Y}\|$ . De esta forma,  $\Theta \sim NP(\cdot | \mathbf{B}^t \mathbf{x}, \mathbf{I})$ . El problema es realizar inferencias sobre  $\mathbf{B}$  con base sólo en una m.a. de ángulos  $\{\theta_1, \dots, \theta_n\}$ . Si se pudiera observar  $(\Theta_1, R_1), \dots, (\Theta_n, R_n)$  entonces se estaría en condiciones de realizar inferencias sobre  $\mathbf{B}$ . El problema es que sólo se observan las direcciones  $\{\theta_1, \dots, \theta_n\}$ .

La estructura del modelo sugiere tratar los  $R_i = \|\mathbf{Y}_i\|$  no observados como “datos faltantes”. Este enfoque fue el seguido por Presnell *et al.* (1998). Ellos se enfocan en la estimación de máxima verosimilitud de  $\mathbf{B}$  vía el algoritmo EM.

El problema de no tener una muestra  $(\Theta_1, R_1), \dots, (\Theta_n, R_n)$  se puede solucionar si se introducen las variables latentes  $R_1, \dots, R_n$ . Así, el modelo para los datos completos será el modelo normal multivariado usual.

En el modelo normal multivariado usual, si  $\mathbf{D}_n = (\mathbf{y}_1, \dots, \mathbf{y}_n)$  es una m.a. de  $N_k(\mathbf{y} | \mathbf{B}^t \mathbf{x}, \mathbf{I})$  y

$$\boldsymbol{\beta}^j \sim N_p(\cdot | \boldsymbol{\beta}_0^j, \lambda_0^j \mathbf{I}), \quad \forall j = 1, \dots, k,$$

entonces

$$f(\boldsymbol{\beta}^j | \mathbf{D}_n) = N_p(\cdot | \boldsymbol{\beta}_F^j, \boldsymbol{\Lambda}_F^j), \quad \forall j = 1, \dots, k.$$

## 4.2. Inferencias vía MCCM

Sea  $R$  una variable latente definida en  $[0, \infty)$ . Como  $\mathbf{Y} \sim N_k(\cdot | \boldsymbol{\mu} = \mathbf{B}^t \mathbf{x}, \mathbf{I})$  se tiene que

$$f(\boldsymbol{\theta}, r | \boldsymbol{\mu}) = N_k(r \mathbf{u} | \boldsymbol{\mu}, \mathbf{I}) | J |,$$

donde  $|J|$  es el Jacobiano de la transformación  $\mathbf{y} \rightarrow (\boldsymbol{\theta}, r)$ . De lo anterior, la condicional completa de  $\boldsymbol{\beta}^j$  está dada por

$$f(\boldsymbol{\beta}^j | \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n, \mathbf{r}) = N_p(\cdot | \boldsymbol{\beta}_F^j, \boldsymbol{\Lambda}_F^j),$$

donde  $\mathbf{r} = (r_1, \dots, r_n)$  es un vector  $n$ -dimensional.

Para simular un vector aleatorio  $\mathbf{R} = (R_1, \dots, R_n)$  de  $f(\mathbf{r} | \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n, \boldsymbol{\beta}^j)$  hay que notar que: De  $f(\boldsymbol{\theta}, r | \boldsymbol{\mu} = \mathbf{B}^t \mathbf{x})$  la densidad condicional de  $R$  queda determinada por

$$f(r | \boldsymbol{\theta}, \boldsymbol{\mu}) \propto r^{p-1} \exp\{-\frac{1}{2}r^2 + \mathbf{u}^t \boldsymbol{\mu}\} I_{(0, \infty)}(r),$$

y que los  $R_i$  dado  $\boldsymbol{\Theta}_i$ ,  $i = 1, \dots, n$ , son independientes.

Así, se puede simular  $R_i$  de  $f(r | \boldsymbol{\theta}_i, \mathbf{B})$  y en esta forma se puede obtener un vector aleatorio  $\mathbf{R}$  de  $f(\mathbf{r} | \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n, \mathbf{B})$ . Este paso es llevado a cabo vía un algoritmo de Metropolis-Hastings.

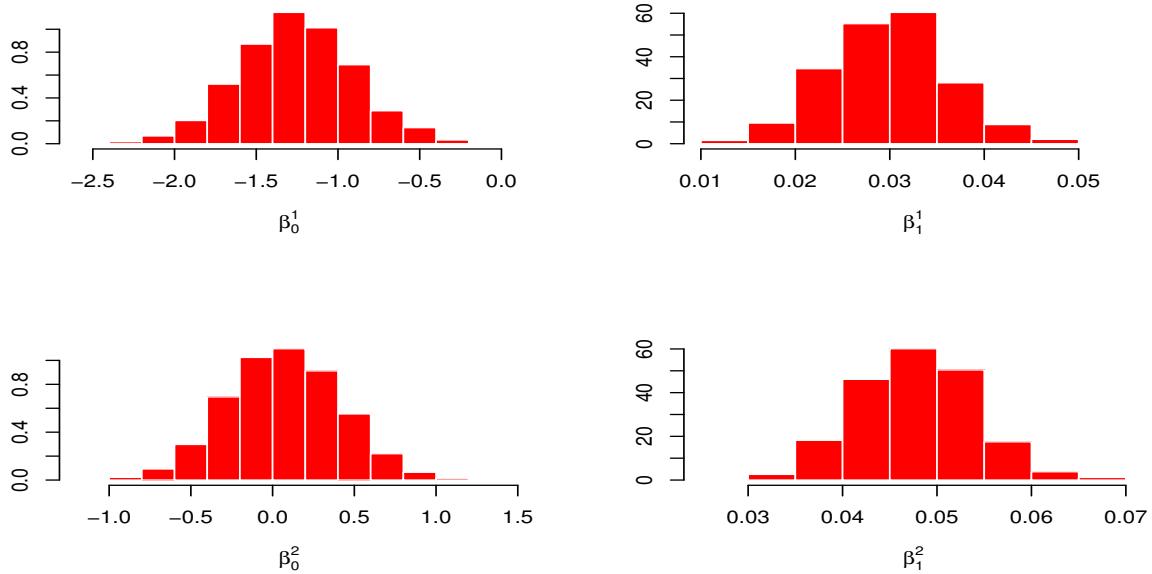
Finalmente, se pueden usar las condicionales completas

$$f(\boldsymbol{\beta}^j | \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n, \mathbf{r}) \text{ y}$$

$$f(\mathbf{r} | \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n, \mathbf{B})$$

en un Gibbs sampler para obtener muestras de la distribución final

$$f(\mathbf{B}, \mathbf{r} | \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n).$$



**Ejemplo.** Los datos para este ejemplo se tomaron de la Tabla 1 de Fisher y Lee (1992) y corresponden a la dirección y distancia recorridas por cierta especie de caracoles de mar, después de ser cambiados de la altura a la cual normalmente viven. Presnell *et al.* (1998) ajustan un modelo bajo la normal proyectada y obtienen los siguientes estimadores (así como sus correspondientes errores estándar):

$$\hat{\beta}^1 = \begin{pmatrix} \hat{\beta}_0^1 \\ \hat{\beta}_1^1 \end{pmatrix} = \begin{pmatrix} -1.228 (0.423) \\ 0.030 (0.008) \end{pmatrix}$$

y

$$\hat{\beta}^2 = \begin{pmatrix} \hat{\beta}_0^2 \\ \hat{\beta}_1^2 \end{pmatrix} = \begin{pmatrix} 0.157 (0.451) \\ 0.049 (0.2) \end{pmatrix}$$

En la figura anterior se muestran las distribuciones finales de los vectores  $\beta^1$  y  $\beta^2$  usando la metodología descrita en este trabajo.

## 5. Comentarios finales

En este trabajo se presenta un análisis Bayesiano completo de un modelo de regresión para datos direccionales basado en el modelo lineal Normal Multivariado bajo proyección. Aunque la versión de la distribución Normal proyectada presentada aquí es simple, es bastante flexible y fácil de analizar vía *Gibbs sampler*. En general, la distribución Normal proyectada se puede considerar como una buena alternativa para modelar datos direccionales. Consideramos que la metodología presentada en este trabajo ofrece las bases para un análisis Bayesiano completo de datos direccionales en el contexto de regresión.

## Referencias

- D'Elia, A. (1999). Analyzing Longitudinal Circular Data by Projected Models, *Proceedings of the 52nd Session of International Statistical Institute*, Helsinki - Finlandia, pp. 249-250.
- Fisher, N. I. y Lee, A. J. (1992). Regression models for an angular response. *Biometrics*, 48, 665-667.
- Nuñez-Antonio, G., Gutiérrez-Peña, E. (2005). A Bayesian Analysis of Directional data using the Projected Normal Distribution. *Journal of the Applied Statistics*, 32, 10, 995-1001.
- Presnell, B., Morrisson, S.P. & Littel, R.C. (1998). Projected multivariate linear model for directional data. *Journal of the American Statistical Association*, 443, 1068-1077.

# **Componentes principales y su relación al análisis de estabilidad con aplicación agronómica**

**Emilio Padrón Corral**

*Centro de Investigación en Matemáticas Aplicadas*

**Ignacio Méndez Ramírez**

*Instituto de Investigación en Matemáticas Aplicadas y en Sistemas*

**Armando Muñoz Urbina**

*Universidad Autónoma Agraria Antonio Narro*

**Félix de Jesús Sánchez Pérez<sup>1</sup>**

*Centro de Investigación en Matemáticas Aplicadas*

## **1. Introducción**

Una de las principales razones para cultivar genotipos en un amplio rango de ambientes es estimar su estabilidad. Esto implica estabilidad de algunas características que pueden ser económicamente importantes como rendimientos y calidad. La selección por estabilidad por muchos años se ha realizado con el modelo de Eberhart y Russell (1966), el cual utiliza el coeficiente de regresión ( $B_i$ ) pero posteriormente considera la suma de cuadrados de desviación ( $S_{di}^2$ ) como una segunda medida.

Carballo y Márquez (1970) indican que existen seis situaciones posibles en que puedan ser clasificados los genotipos según los valores que pueden tomar estos dos parámetros de estabilidad. Otras metodologías se han generado para describir patrones de interacción genotipo-ambiente.

Westcott (1985) señala que el análisis de componentes principales se relaciona con la regresión lineal, en la cual la estimación de mínimos cuadrados de los coeficientes de regresión es equivalente a extraer el primer componente principal del comportamiento de los genotipos.

Kempton (1984) señala que el primer componente principal maximiza la variación entre

---

<sup>1</sup>fel1925@yahoo.com

genotipos. El segundo componente principal es el eje en ángulo recto con el primero, el cual maximiza la variación restante, cuando la mayoría de la variación de la respuesta varietal es explicada por los dos primeros componentes principales, una gráfica de las variedades con estos dos ejes proporciona una descripción exitosa de los datos. Con respecto a los objetivos generales de este trabajo se encuentran:

- Identificar genotipos de cártamo con alto rendimiento y estabilidad.
- Identificar los ambientes más relevantes que influyen para separar los genotipos por su sensibilidad ambiental.

## 2. Materiales y Métodos

El material genético utilizado en este experimento está constituido por 22 genotipos de cártamo en nueve ambientes:

1. Venecia, Durango en 1987 (V1), 1988 (V2), 1989 (V3).
2. Buenavista, Coahuila en 1987 (V4), 1988 (V5), 1989 (V6).
3. Ocampo, Coahuila en 1987 (V7), 1989 (V8).
4. Muzquiz, Coahuila en 1989 (V9).

El diseño experimental utilizado fue el de Bloques al Azar con tres repeticiones, la parcela experimental constó de cuatro surcos de tres metros de longitud con 0.08 m y 0.10 m de distancia entre surcos y plantas, respectivamente.

Las fechas de siembra se determinaron considerando las establecidas para los cultivos de otoño-invierno. Se sometió a análisis de estabilidad el rendimiento de grano ajustado al 8 % de humedad y trasformando a t/ha, efectuándose además el análisis por componentes principales.

## 2.1. Resultados y Discusión

En el Cuadro 1, se presenta la clasificación de los 22 genotipos de cártamo según los parámetros de estabilidad propuestos por Eberhart y Russell, de los 22 genotipos de cártamo 16 fueron clasificados como estables (a), tres presentaron buena respuesta a todos los ambientes pero fueron inconsistentes (b), dos responden mejor en ambientes desfavorables: uno consistente (c) y el otro inconsistente (d), y uno responde mejor en buenos ambientes pero es inconsistente (f). Los genotipos 6, 8, 14 y 22 fueron los que presentaron mayor rendimiento y estabilidad. La estimación de los coeficientes de regresión variaron de  $B_i = 0.511$  (genotipo 20) a  $B_i = 2.696$  (genotipo 11), lo que indica que los genotipos mostraron considerable variación en su respuesta a los ambientes que favorecieron un más alto rendimiento de grano, en este caso el ambiente de Buenavista, Coahuila 1989 (Cuadro 2).

La Figura 1, muestra que el primer componente principal, el ambiente Buenavista, Coahuila 1989 (V6) tuvo mayor peso que los ambientes de bajo rendimiento (Cuadro 3). Esto sugiere que el primer componente principal discrimina los genotipos que se comportan relativamente mejor en ambientes con alto rendimiento (genotipos 7 y 11) de aquellos que se comportan mejor en ambientes de bajo rendimiento (genotipos 19 y 20) esto coincide con la clasificación obtenida utilizando la técnica de Eberhart y Russell, ya que los genotipos 7 clasificado como (b) y 11 clasificado como (f) responden mejor en buenos ambientes y los genotipos 19 y 20 clasificados como (c) y (d), respectivamente, responden mejor en ambientes desfavorables.

Así, el primer componente principal, separa a los genotipos de acuerdo con su sensitividad ambiental. Similarmente el segundo componente principal, separa a los genotipos que se comportaron mejor en el ambiente Venecia, Durango 1988 (V2) de aquellos que se comportaron mejor en otros ambientes. Los genotipos y ambientes con altos valores en el primer componente principal (positivos o negativos) tienen grandes interacciones y aquellos cercanos a cero tienen pequeñas interacciones. Así, los genotipos 11, 18, 20, 21 pueden ser considerados inconsistentes como fueron clasificados en el análisis de estabilidad de Eberhart y Russell, y los genotipos más cercanos al origen en el primer componente principal pueden ser considerados como estables y consistentes.

Sánchez (1995) indica que los valores de los genotipos para cada ambiente en particular puede verse proyectando los puntos de las observaciones sobre el vector que representa los ambientes,

así por ejemplo, los genotipos 11 y 7 fueron los que presentaron mayor rendimiento en el ambiente V6 y los genotipos 21, 20 y 18 fueron los de más alto rendimiento en el ambiente V2, el vector señala en la dirección de los valores mayores para el ambiente (Figura 1).

Con respecto a la varianza acumulada (Cuadro 3) los dos primeros componentes principales describieron el 73 % de la suma de cuadrados total de rendimiento varietal dentro de ambientes.

Cuadro 1. Clasificación del rendimiento de 22 genotipos de cártamo según su estabilidad (Método Eberhart y Russell).

Genotipo	Media (ton/ha)	<i>Estabilidad</i> <sup>1</sup>	Genotipo	Media (ton/ha)	<i>Estabilidad</i> <sup>1</sup>
1	1.755	a	12	1.655	a
2	1.918	a	13	1.806	a
3	1.926	a	14	2.115	a
4	1.720	a	15	1.889	a
5	1.872	a	16	1.836	a
6	2.024	a	17	1.630	a
7	2.107	b	18	1.810	b
8	2.081	a	19	1.929	c
9	1.747	a	20	1.736	d
10	1.923	a	21	2.255	b
11	2.794	f	22	2.265	a

<sup>1</sup> Clasificación Carballo y Márquez

Cuadro 2. Rendimiento promedio e índice ambiental en los 9 ambientes de evaluación.

Ambiente	Rendimiento promedio (ton/ha)	Indice ambiental
Buenavista, Coah. 89 (V6)	4.229	2.283
Venecia, Dgo. 89 (V3)	2.656	0.710
Ocampo, Coah. 89 (V8)	2.247	0.301
Buenavista, Coah. 88 (V5)	1.822	-0.124
Venecia, Dgo. 88 (V2)	1.780	-0.166
Venecia, Dgo. 87 (V1)	1.673	-0.273
Ocampo, Coah. 87 (V7)	1.052	-0.894
Muzquiz, Coah. 89 (V9)	1.051	-0.895
Buenavista, Coah. 87 (V4)	1.001	-0.945
Media General	1.946	

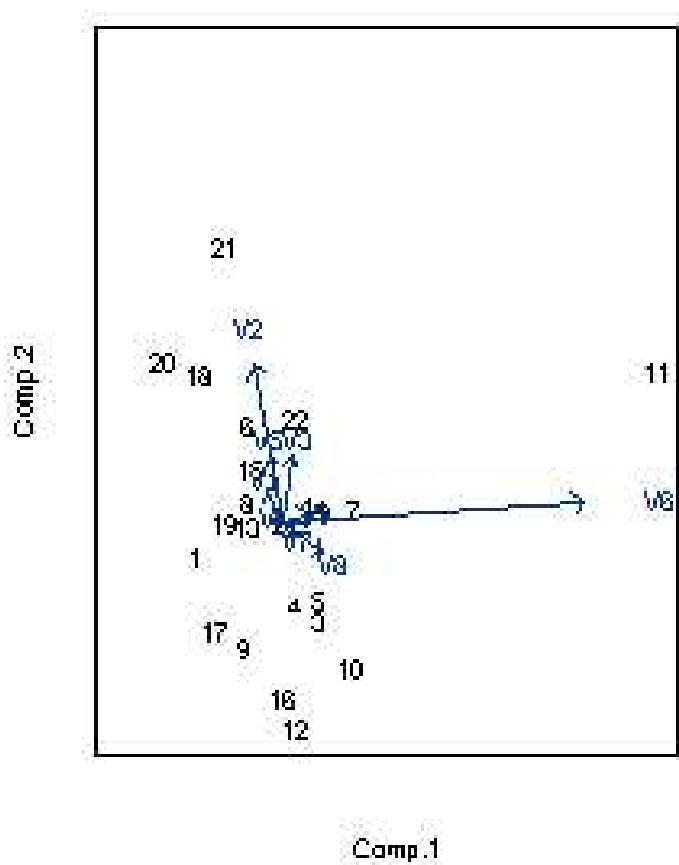


Figura 1: Gráfica de componentes principales para los genotipos de cártamo.

Cuadro 3. Ponderaciones de los nueve ambientes en las componentes principales y la proporción de varianza acumulada.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	Varianza acumulada
CP1	-	-	-	-	-	0.98	-	0.13	-	0.586
CP2	0.18	0.82	0.36	-	0.35	0.10	-	-0.16	-	0.732
CP3	-	0.34	-0.76	-	0.21	-	-	0.50	-	0.828
CP4	0.10	-0.34	-0.17	-0.20	0.76	-	0.12	-0.37	0.21	0.891
CP5	-	-0.20	0.47	0.18	0.41	-0.11	0.13	0.70	-	0.940
CP6	-0.92	0.11	-	-0.24	0.11	-	-0.19	-	-	0.969
CP7	-	-	0.14	-0.62	-0.23	-0.11	0.28	0.18	0.62	0.982
CP8	0.27	-0.13	0.10	-0.44	-	-	-0.80	0.16	-0.13	0.993
CP9	-	-	-	-0.52	-	-	0.43	-	-0.71	1.000

### 3. Conclusiones

1. El método de parámetros de estabilidad propuesto por Eberhart y Russell, permitió detectar genotipos estables con alto rendimiento o adaptados a ambientes favorables o desfavorables.
2. El análisis de Componentes Principales estuvo acorde con los resultados obtenidos con el análisis de estabilidad de Eberhart y Russell para detectar genotipos estables o adaptados a ambientes favorables o desfavorables permitiendo además, detectar que los ambientes tuvieron mayor impacto en explicar el comportamiento de los genotipos.
3. Es importante utilizar las dos metodologías para tener una mejor interpretación de los resultados.

### Referencias

- Eberhart, S.A. and Russell, W.A. (1996). Stability parameters for comparing varieties *Crop Science* **6**: 36-40.
- Carballo, C.A. y Márquez, F. (1970). Comparación de variedades de maíz del Bajío y la Mesa Central por rendimiento y estabilidad. *Agrociencia*. **5**: 129-146.

Kempton, R.A. (1984). The use of the bi-plots in interpreting variety by environment interactions *Journal of Agricultural Science*. Cambridge. **103**: 123-135.

Sánchez, J.J. (1995). El análisis biplot en clasificación *Revista Fitotecnia Mexicana* **18**:188-203.

Westcott, B. (1985). Some methods of analysis genotypoe-environment interaction. *Heredity*. **56**: 243-253.



# Una prueba para normalidad basada en la propiedad de la cerradura de convoluciones

**José A. Villaseñor A.**<sup>1</sup>

*Colegio de Postgraduados*

**Elizabeth González Estrada**<sup>2</sup>

*Colegio de Postgraduados*

## 1. Introducción

Varios de los métodos estadísticos clásicos como son el análisis de varianza, el análisis de regresión y la prueba de t se basan en la hipótesis de que las observaciones se distribuyen normalmente. Sin embargo, cuando los datos no satisfacen dicha hipótesis, la validez de las conclusiones se puede ver seriamente afectada, conduciendo a decisiones equivocadas. Por lo tanto, al hacer uso de esta clase de métodos es necesario contar con una prueba para normalidad que preserve el tamaño de prueba y que sea lo más potente posible.

En este trabajo se presenta una prueba de bondad de ajuste para la distribución normal basada en la propiedad de cerradura de sumas de variables aleatorias (v.a's) normales así como una comparación entre la potencia de esta nueva prueba y la potencia de la prueba  $W$  de Shapiro y Wilk (1965) la cual es una de las mejores pruebas para normalidad (Thode, 2002).

## 2. Cerradura de sumas de variables aleatorias normales

El siguiente teorema establece que la familia de la distribución normal es cerrada bajo convoluciones, es decir, que la suma de v.a's normales independientes también es una v.a. normal.

**Teorema:**  $X$  y  $Y$  son v.a's i.i.d.  $N(\mu, \sigma^2)$  si y sólo si la v.a.  $S = aX + bY$  tiene distribución

---

<sup>1</sup>jvillasr@colpos.mx

<sup>2</sup>egonzalez@colpos.mx

normal con parámetros:  $(a + b)\mu$  y  $(a^2 + b^2)\sigma^2$ , donde  $a$  y  $b$  son cualesquiera constantes.

Note que cuando  $a = b = 1$ , la v.a.  $Z = X + Y$  tiene distribución  $N(2\mu, 2\sigma^2)$  y entonces se tiene que

$$\begin{aligned} F_Z(z) &= P[Z \leq z] = P\left[\frac{Z - 2\mu}{\sqrt{2}\sigma} \leq \frac{z - 2\mu}{\sqrt{2}\sigma}\right] \\ &= \Phi\left(\frac{z - 2\mu}{\sqrt{2}\sigma}\right) \end{aligned}$$

lo cual implica que  $\Phi^{-1}(F_Z(z)) = \frac{z - 2\mu}{\sqrt{2}\sigma}$ , donde  $\Phi(\cdot)$  denota a la función de distribución normal estándar.

Note que la correlación lineal entre  $q(Z) \equiv \Phi^{-1}(F_Z(Z))$  y  $Z$  es igual a uno ya que  $q(Z)$  es una función lineal de  $Z$  con pendiente  $\frac{1}{\sqrt{2}\sigma} > 0$ .

## 2.1. La prueba

Sean  $X_1, X_2, \dots, X_n$  una muestra aleatoria (m.a.) de tamaño  $n$  y sea  $Z_k = X_i + X_j$ ,  $i < j$ ,  $j = 1, 2, \dots, n$ . Para probar la hipótesis siguiente:

$H_0 : X_1, X_2, \dots, X_n$  es una m.a.  $N(\mu, \sigma^2)$  donde  $\mu \in \Re$  y  $\sigma^2 > 0$  son desconocidos

se propone la estadística de prueba

$$R = \frac{\sum_{k=1}^m \left( \widehat{q(Z_k)} - \widehat{\overline{q(Z)}} \right) \left( Z_k - \bar{Z} \right)}{\sqrt{\sum_{k=1}^m \left( \widehat{q(Z_k)} - \widehat{\overline{q(Z)}} \right)^2 \sum_{k=1}^m \left( Z_k - \bar{Z} \right)^2}}$$

donde  $m = n(n-1)/2$ ,  $\widehat{q(Z_k)} \equiv \Phi^{-1}(\widehat{F_Z(Z_k)})$  y  $\widehat{F_Z(Z_k)}$  denota a la función de distribución empírica de las  $Z'$ s.

Note que la estadística  $R$  es el coeficiente de correlación muestral entre  $q(Z)$  y  $Z$ . Además, la distribución de  $R$  bajo  $H_0$  no depende de los parámetros desconocidos  $\mu$  y  $\sigma^2$  ya que  $R$  es invariante bajo transformaciones de escala y localidad, por lo que la distribución de  $R$  bajo  $H_0$  solo depende del tamaño de muestra  $n$ .

La regla de decisión es la siguiente: se rechaza la hipótesis nula de normalidad si  $R < R_\alpha$  donde el valor crítico  $R_\alpha$  es tal que  $P(R < R_\alpha | H_0) = \alpha$ .

### 3. Estudio de simulación

La distribución de  $R$  se calculó por simulación de Monte Carlo para tamaños de muestra  $n = 20, 30$  y  $50$  usando 20000 m.a. de la distribución Normal estándar. De la distribución estimada de  $R$  dado  $n$  se obtuvieron los valores críticos siguientes para el tamaño de prueba  $\alpha = 0.05$ :

<b>n</b>	<b>R<sub>0.05</sub></b>
20	0.9845
30	0.9891
50	0.9933

Para realizar el estudio de potencia se consideraron distribuciones alternativas de cola pesada, asimétricas y con soporte finito. Las potencias de las pruebas  $R$  y  $W$  de Shapiro-Wilk se estimaron usando 10000 m.a. de tamaño 50 de cada alternativa y  $\alpha = 0.05$ .

Las alternativas de cola pesada consideradas fueron la distribución  $t$  con 1, 2 y 4 grados de libertad (g.l.) y las distribuciones estables simétricas con parámetros de estabilidad 1.3, 1.5 y 1.7. Recuerde que la distribución  $t$  con 1 g.l. es igual a la distribución Cauchy y que la distribución Normal es una distribución estable con parámetro de estabilidad igual a 2. En la Tabla 1 se presentan las potencias calculadas contra estas alternativas. En esta situación la prueba  $R$  resulta ser más potente que la prueba  $W$  en todos los casos.

En la Tabla 2 se presentan las potencias obtenidas cuando se consideraron las distribuciones asimétricas Gumbel, Estable, Lognormal, Ji-cuadrada, Gama y Weibull. Los números que aparecen resaltados corresponden a la prueba más potente. La prueba  $R$  es más potente que la prueba  $W$  contra las distribuciones Gumbel y Estable asimétrica. Ninguna de las dos pruebas es uniformemente más potente bajo las alternativas Ji-cuadrada y Gama; sin embargo, sus potencias no son muy diferentes.

En la Tabla 3 se observa que la prueba  $W$  detecta mejor que la prueba  $R$  las distribuciones Beta. Recuerde que la distribución Beta(1,1) es igual a la distribución Uniforme(0,1).

## 4. Conclusiones

La potencia de la prueba de bondad de ajuste para normalidad propuesta es mayor que la potencia de la prueba de Shapiro-Wilk en la mayoría de las distribuciones alternativas consideradas en el estudio de simulación realizado.

Similarmente a la prueba de Shapiro-Wilk, la prueba obtenida también es invariante, es decir, no es necesario estimar los parámetros de la distribución Normal para poder realizar la prueba ya que la distribución bajo  $H_0$  de la estadística de prueba no depende de dichos parámetros.

## Referencias

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality: complete samples. *Biometrika*. **52** 591-611.

Thode, H. C. (2002). *Testing for normality*. New York: Marcel Dekker, Inc.

**Tabla 1. Potencias estimadas de las pruebas R y W contra alternativas de cola pesada con n=50 y  $\alpha=0.05$**

Alternativa	R	W
t(1)	1.00	0.99
t(2)	0.86	0.81
t(4)	0.47	0.37
stab(1.3)	0.93	0.91
stab(1.5)	0.81	0.76
stab(1.7)	0.59	0.50

**Tabla 2. Potencias estimadas de las pruebas R y W contra alternativas asimétricas con n=50 y  $\alpha=0.05$**

Alternativa	R	W	Alternativa	R	W
Gumbel(1)	0.73	0.66	exp(1)	1.00	1.00
Gumbel(2)	0.74	0.66	gamma(2,1)	0.93	0.95
Gumbel(4)	0.76	0.66	gamma(4,2)	0.70	0.66
Gumbel(2,2)	0.73	0.66	gamma(5,1)	0.59	0.55
stab(1.5,.5)	0.87	0.80	gamma(5,2)	0.64	0.56
stab(1.7,.5)	0.62	0.56	Weibull(2)	0.38	0.41
Inorm	1.00	1.00	Weibull(4)	0.04	0.05
chisq(4)	0.93	0.95	Weibull(2,2)	0.38	0.42
chisq(6)	0.80	0.81	Weibull(2,2)	0.38	0.42

**Tabla 3. Potencias estimadas de las pruebas R y W contra alternativas con soporte finito con n=50 y  $\alpha=0.05$**

Alternativa	R	W
beta(1,1)	0.13	0.86
beta(2,1)	0.67	0.88
beta(3,2)	0.10	0.26



# Extensión de la prueba $t$ para observaciones intercambiables

José A. Villaseñor A.<sup>1</sup>

*Colegio de Postgraduados*

Eduardo Gutiérrez G.<sup>2</sup>

*Instituto Politécnico Nacional*

## 1. Caso de independencia

Primeramente recordaremos la prueba  $t$  para el caso de independencia.

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de variables  $N(\mu, \sigma^2)$  y el contraste de hipótesis

$$H_0 : \mu = \mu_0, \sigma^2 > 0,$$

$$H_1 : \mu \neq \mu_0, \sigma^2 > 0,$$

en donde,  $\mu_0$  es una constante conocida, el parámetro  $\sigma^2$  es desconocido, aquí el espacio paramétrico bajo la hipótesis nula es  $\omega = \{\theta = (\mu, \sigma^2) | \mu = \mu_0, \sigma^2\}$  y el espacio paramétrico  $\Omega = \mathbb{R} \times \mathbb{R}^+$ .

Para la prueba del contraste de hipótesis se utiliza el método de pruebas de la razón de verosimilitudes generalizada, para lo cual se obtienen los estimadores de máxima verosimilitud de los parámetros  $(\mu, \sigma^2)$  bajo todo  $\Omega = \mathbb{R} \times \mathbb{R}^+$ ,

$$(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, S_{n-1}^2),$$

y bajo  $\omega$

---

<sup>1</sup>jvillasr@colpos.mx

<sup>2</sup>guge610926@yahoo.com.mx

$$(\mu_0, \widehat{\sigma}^2) = \left( \mu_0, \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu_0)^2 \right).$$

Aplicando el método de la razón de verosimilitudes generalizada y el resultado de que  $\bar{X}$  y  $S_{n-1}^2$  son independientes, se obtiene el estadístico de prueba

$$T = \frac{\bar{X} - \mu_0}{\frac{S_{n-1}}{\sqrt{n}}}. \quad (1)$$

Donde  $T$  tiene una distribución  $t$ -student con  $n - 1$  grados de libertad y la hipótesis nula se rechaza cuando

$$T < -k \text{ o } T > k,$$

con la constante crítica  $k$  igual al cuantil  $t_{1-\frac{\alpha}{2}}(n-1)$ .

## 2. Caso de variables intercambiables

En general, cuando las variables aleatorias  $X_1, X_2, \dots, X_n$  son dependientes la prueba anterior se complica enormemente, y su solución depende de la matriz de varianzas y covarianzas, más aún, en la literatura no se tienen pruebas para estos casos. Pero si relajamos un poco la dependencia y trabajamos con variables intercambiables se puede demostrar el siguiente resultado.

**Teorema 1** *Sea  $X_1, X_2, \dots, X_n$  variables aleatorias intercambiables e idénticamente distribuidas normales con parámetros  $\mu, \sigma^2$  y con covarianzas homogéneas,  $c \geq 0$ , entonces para el contraste de hipótesis*

$$H_0 : \mu = \mu_0, \sigma^2 > 0, c \geq 0,$$

$$H_1 : \mu \neq \mu_0, \sigma^2 > 0, c \geq 0,$$

en donde,  $\mu_0$  es una constante conocida, la prueba es la misma que en el caso de independencia ( $c = 0$ ). Es decir, las pruebas para los casos de independencia e intercambiabilidad coinciden.

## Demostración

Sean  $X_1, X_2, \dots, X_n$  variables aleatorias intercambiables normalmente distribuidas con parámetros  $\mu, \sigma^2$  y  $c$ . La prueba para el contraste de hipótesis

$$\begin{aligned} H_0 &: \mu = \mu_0, \sigma^2 > 0, c \geq 0; \\ H_1 &: \mu \neq \mu_0, \sigma^2 > 0, c \geq 0, \end{aligned}$$

en donde,  $\mu_0$  es una constante conocida, el parámetro  $\sigma^2$  es desconocido, el espacio paramétrico bajo la hipótesis nula es

$$\omega = \{\theta = (\mu, \sigma^2, c) | \mu = \mu_0, \sigma^2 > 0, c \geq 0\}$$

y el espacio paramétrico general  $\Omega = \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$ , se busca en base a un estadístico de prueba, que tenga una representación similar a la estadística de prueba dada en (1), para independencia. La demostración se simplifica por medio de las siguientes subsecciones.  $\square$

### 2.1. Comprobación de la inexistencia de los EMV para $\mu, \sigma^2$ y $c$

Primeramente se demuestra que la matriz de varianzas y covarianzas de las variables aleatorias  $X_1, X_2, \dots, X_n$ , está dada por

$$\Sigma = c\mathbf{J} - (c - \sigma^2)\mathbf{I}.$$

En donde,  $\mathbf{J}$  es la matriz de unos de orden  $n \times n$ , mientras que  $\mathbf{I}$  se refiere a la matriz identidad del mismo orden.

Posteriormente, se demuestra que los estimadores de máxima verosimilitud para los parámetros  $\mu, \sigma^2$  y  $c$ , no existen. Esto se demuestra al calcular los estimadores de máxima verosimilitud y comprobar que no son acotados.

## 2.2. Transformación de las variables aleatorias

Se demuestra que en el caso de variables aleatorias intercambiables normalmente distribuidas con parámetros  $\mu, \sigma^2$  y  $c$ , y se tiene que

$$(a) \bar{X} \sim N\left(\mu, \frac{c(n-1)+\sigma^2}{n}\right) \text{ y } S_{n-1}^2 \sim \Gamma\left(\frac{n-1}{2}, 2\frac{\sigma^2-c}{n-1}\right).$$

(b)  $\bar{X}$  y  $S_{n-1}^2$  siguen siendo independientes. De tal forma que bajo  $H_0$ .

$$(c) \frac{\bar{X}-\mu_0}{\sqrt{\frac{c(n-1)+\sigma^2}{n}}} \sim N(0, 1), \frac{(n-1)S_{n-1}^2}{\sigma^2-c} \sim \Gamma\left(\frac{n-1}{2}, 2\right) = \chi_{n-1}^2 \text{ y se prueba que son independientes.}$$

Para verificar los tres incisos, primeramente se calcula el determinante de la matriz de varianzas y covarianzas de las variables aleatorias  $X_1, X_2, \dots, X_n$ . Posteriormente, con la matriz de covarianzas se lleva a cabo una transformación, basada en la matriz de Helmer y con las nuevas variables se pueden concluir los primeros incisos. El inciso (c) se comprueba con las propiedades de variables aleatorias.

## 2.3. Estadística de prueba y su distribución

Con base en los resultados de la subsección 2.2 y la estadística de prueba dada en (1), se define la estadística de prueba para variables intercambiadas y se obtiene su distribución.

En este caso la estadística de prueba estará dada por:

$$T^* = \frac{\frac{\bar{X}-\mu_0}{\sqrt{\frac{c(n-1)+\sigma^2}{n}}}}{\sqrt{\frac{\frac{(n-1)S_{n-1}^2}{\sigma^2-c}}{n-1}}}.$$

La cual se probará que tiene una distribución  $t$  central con  $n - 1$  grados de libertad y un factor de escala, el cual se puede obtener explícitamente al simplificar la expresión anterior, obteniendo

$$T^* = \frac{\bar{X} - \mu_0}{\frac{S_{n-1}}{\sqrt{n}}} \sqrt{\frac{\sigma^2 - c}{c(n-1) + \sigma^2}} = \frac{\bar{X} - \mu_0}{\frac{S_{n-1}}{\sqrt{n}}} \sqrt{\frac{1 - \rho}{\rho(n-1) + 1}}.$$

De tal forma que

$$T^* = \frac{\bar{X} - \mu_0}{\frac{S_{n-1}}{\sqrt{n}}} h(\rho) \sim t_{n-1}. \quad (2)$$

en donde,  $\bar{X}$  y  $S_{n-1}^2$  y son la media y varianza muestrales y son independientes, y  $h(\rho) = \sqrt{\frac{1-\rho}{\rho(n-1)+1}}$  con  $0 \leq \rho < 1$ , es el factor de escala que resulta ser una función del coeficiente de correlación.

## 2.4. Búsqueda de la constante crítica

Se mencionó que el estadístico de prueba se busca de forma similar al dado en (1). Luego, al iniciar la búsqueda de la constante crítica  $k$ , para el tamaño de la prueba, se empleará la estadística dada en (2), junto con su distribución

$$\begin{aligned} 1 - \alpha &= \max_{0 \leq \rho < 1} P(-k < T < k | H_0) \\ &= \max_{0 \leq \rho < 1} P\left(-k < \frac{T^*}{h(\rho)} < k | H_0\right) \\ &= \max_{0 \leq \rho < 1} P(|T^*| < kh(\rho) | H_0) \\ &= \max_{0 \leq \rho < 1} F_{|T^*|}(kh(\rho)) \\ &= F_{|T^*|}\left(k \left( \max_{0 \leq \rho < 1} h(\rho) \right)\right). \end{aligned}$$

Es decir,

$$1 - \alpha = F_{|T^*|} \left( k \left( \max_{0 \leq \rho < 1} h(\rho) \right) \right). \quad (3)$$

#### 2.4.1. Acotación de la función de escala

En esta parte, se tiene que demostrar que la función de escala,  $h(\cdot)$ , es monótona decreciente. Por consiguiente, el máximo de la función de escala se obtiene cuando  $\rho = 0$ , es decir, cuando  $h(0) = 1$ .

De la monotonía de la función de escala y (3), resulta

$$1 - \alpha = F_{|T^*|} \left( k \left( \max_{0 \leq \rho < 1} h(\rho) \right) \right) = F_{|T^*|} (kh(0)) = F_{|T^*|} (k).$$

#### 2.4.2. Regla de decisión

De la subsubsección 2.4.1 se concluye que la prueba no rechaza  $H_0$  cuando  $-k < T < k$ , donde  $k$  es tal que si  $\alpha \in (0, 1)$ , entonces

$$P(-k < T^* < k | H_0) \leq 1 - \alpha.$$

Con  $k = t_{1-\frac{\alpha}{2}}(n-1)$  el  $1 - \frac{\alpha}{2}$  cuantil de la distribución  $t$ -student con  $n-1$  grados de libertad.

Con lo cual se concluye que se tiene la misma prueba para el caso de intercambiabilidad e independencia.

## Referencias

- Billingsley, P. (1979). *Probability and Measure*. Wiley, New York.
- Casella, G. and Berger, R.L. (1990). *Statistical inference*. Duxbury press, Belmont, California.
- Chung, K. L. (1968). *A course in probability theory*. Harcourt, Brace and World, Inc.
- Feller, W. (1971). *An introduction to probability theory and its applications* V.2 (Second edition). Wiley, New York.
- Halperin, M. (1963). Approximations to the non-central t, with applications, *Technometrics*, **Vol. 5, No. 3**, pp 295-305.
- Herstein, I. N. and Winter, D. J. (1988). *A primer on Linear Algebra*, MacMillan publishing company, United States of America.
- Kraemer, H. C. and Paik, M. (1979). A central t approximation to the noncentral t-distribution. *Technometrics*, **Vol. 21, No. 3**, pp 357-360.
- Maltsev, A. I. (1972). *Fundamentos de álgebra lineal*. Editorial Mir, Moscú.
- Mood, A.M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the theory of statistics*, (third edition), McGraw Hill, Singapore.
- Owen, D. B. (1968). A Survey of properties and applications of the noncentral t-distribution, *Technometrics*, **Vol. 10, No. 3**, pp 445-473.
- Rohatgi, V. K. (1984). *Statistical inference*, Wiley, New York.
- Searle, S. R. (1982). *Matrix algebra useful for statistics*, Wiley, New York.
- Wilks, S. S. (1962). *Mathematical statistics*. Wiley, New York.



# Pruebas de hipótesis para procesos Gaussianos

**José A. Villaseñor A.**<sup>1</sup>

*Colegio de Postgraduados*

**Eduardo Gutiérrez G.**<sup>2</sup>

*Instituto Politécnico Nacional*

## 1. Introducción

En finanzas, las variables aleatorias  $X(t)$  de un proceso Gaussiano pueden representar el valor diario del índice de la bolsa de valores al tiempo  $t$  ó los precios al cierre diarios de una acción. Lo único que se conoce de ellas, es que bajo ciertas hipótesis, provienen de la misma distribución y que son dependientes; entonces se desea encontrar una prueba para el contraste de hipótesis:

$$H_0 : p \leq p_0 \text{ vs } H_1 : p > p_0 \quad (1)$$

en donde,  $p = P[X(t) > q_0]$  con  $q_0$  y  $p_0$  constantes definidas anticipadamente.

Si se hacen algunas consideraciones sobre las covarianzas, de tal forma que se tenga un proceso débilmente estacionario, es posible construir una prueba para este contraste de hipótesis.

Con base en una realización finita  $X_1, X_2, \dots, X_n$  del proceso Gaussiano  $X(t)$  donde las  $X_i$  son v.a.'s dependientes e idénticamente distribuidas con parámetros  $\mu$ ,  $\sigma^2$  y covarianzas homogéneas,  $c$ , a continuación se describe una prueba estadística para el contraste de hipótesis (1).

Por la normalidad de las variables aleatorias, el contraste de hipótesis (1), es equivalente al contraste siguiente:

---

<sup>1</sup>jvillasr@colpos.mx

<sup>2</sup>guge610926@yahoo.com.mx

$$\begin{aligned} H_0 : \frac{q_0 - \mu}{\sigma} &\geq \Phi^{-1}(1 - p_0), \quad c \in R \\ H_1 : \frac{q_0 - \mu}{\sigma} &< \Phi^{-1}(1 - p_0), \quad c \in R \end{aligned} \tag{2}$$

en donde  $\Phi(x)$  es la función de distribución normal estándar.

Entonces se propone una prueba basada en la estadística:

$$T = \frac{q_0 - \bar{X}}{S}, \tag{3}$$

en donde  $\bar{X}$  y  $S = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$  son los estimadores de momentos para  $\mu$  y  $\sigma$ .

La prueba rechaza cuando  $T < k_\alpha$ , donde  $k_\alpha$  es tal que  $P(T < k_\alpha | H_0) \leq \alpha$ , para una  $\alpha \in (0, 1)$  dada.

La distribución de la estadística de prueba (3) puede ser obtenida de la distribución conjunta de  $\bar{X}$  y  $S$ .

Las v.a's  $X_1, X_2, \dots, X_n$  tienen densidad conjunta:

$$f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det(\Sigma)}} \exp \left\{ -\frac{1}{2} (\underline{\mathbf{x}} - \underline{\mu})' \Sigma^{-1} (\underline{\mathbf{x}} - \underline{\mu}) \right\},$$

con vector de medias  $\underline{\mu} = \mu \mathbf{1}$  y matriz de varianzas y covarianzas  $\Sigma$ , con covarianzas  $cov(X_i, X_j) = c$  con  $i \neq j$ .

Note que  $\Sigma = c\mathbf{J} - (c - \sigma^2)\mathbf{I}$ , en donde  $\mathbf{J}$  es la matriz de unos de orden  $n \times n$  e  $\mathbf{I}$  es la matriz identidad del mismo orden. De donde,

$$\det(\Sigma) = \sigma^2(\sigma^2 - c)^{n-1} \left[ \frac{c}{\sigma^2}(n-1) + 1 \right].$$

Por lo tanto para que la matriz  $\Sigma$  sea positiva definida es necesario suponer que  $c > 0$ .

## 2. Transformación de variables

Sea  $\Gamma$  la matriz ortonormal de Helmert. Mediante la transformación  $\underline{Y} = \Gamma \underline{X}$  del vector de observaciones es posible mostrar que en esta situación las estadísticas  $\bar{X}$  y  $S_{\underline{X}}^2$  son independientes.

**Teorema 2.1:** Si  $\underline{X} \sim N^{(n)}(\mu \underline{1}, cJ - (c - \sigma^2)I)$  y  $\underline{Y} = \Gamma \underline{X}$  donde  $\Gamma$  es la matriz ortonormal de Helmert entonces:

$$\underline{Y} \sim N^{(n)}(\sqrt{n}\mu \underline{e}_1, D(c(n-1) + \sigma^2, (\sigma^2 - c), \dots, (\sigma^2 - c)))$$

en donde,  $D$  es una matriz diagonal de orden  $n$ ,  $\underline{e}'_1 = (1, 0, 0, \dots, 0)$  y

$$\Gamma = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ -\frac{1}{\sqrt{2(1)}} & \frac{1}{\sqrt{2(1)}} & 0 & \cdots & 0 \\ -\frac{1}{\sqrt{3(2)}} & -\frac{1}{\sqrt{3(2)}} & \frac{2}{\sqrt{3(2)}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{\sqrt{n(n-1)}} & -\frac{1}{\sqrt{n(n-1)}} & -\frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{n-1}{\sqrt{n(n-1)}} \end{pmatrix}.$$

### 2.1. Propiedades de las variables transformadas $\underline{Y} = \Gamma \underline{X}$

Las siguientes propiedades son consecuencia del Teorema 2.1.

**Propiedad 1.** Las  $Y_i$  tienen distribución normal y son independientes.

**Propiedad 2.**  $\bar{X} = Y_1/\sqrt{n}$ .

**Propiedad 3.**  $S_{\underline{X}}^2 = \sum_{i=2}^n Y_i^2$ .

**Propiedad 4.**  $\bar{X}$  y  $S_{\underline{X}}^2$  son independientes.

### 3. Distribución de la estadística de prueba

De las propiedades anteriores se sigue

$$\bar{X} \sim N\left(\mu, \frac{c(n-1) + \sigma^2}{n}\right), \quad (4)$$

$$S_{\mathbf{x}}^2 \sim \Gamma\left(\frac{n-1}{2}, 2\frac{\sigma^2 - c}{n}\right). \quad (5)$$

Utilizando los resultados (4) y (5) se obtiene el siguiente teorema.

**Teorema 3.1:** Sea  $\underline{\mathbf{X}} \sim N^{(n)}(\mu \mathbf{1}, \Sigma)$  con  $\Sigma = c\mathbf{J} - (c - \sigma^2)\mathbf{I}$ , entonces la estadística de prueba  $T = \frac{q_0 - \bar{X}}{S}$ , tiene una distribución  $t$  no central con parámetro de no centralidad  $\sqrt{n} \left( \frac{\mu - q_0}{\sigma \sqrt{\rho(n-1) + 1}} \right)$  donde  $\rho = \frac{c}{\sigma^2}$ .

### 4. Valores críticos

Del Teorema 3.1 se puede obtener que

$$T = h(\rho) \left( \frac{W}{\sqrt{U}} \right),$$

donde  $W \sim N\left(\frac{\sqrt{n}(\mu - q)}{\sqrt{c(n-1) + \sigma^2}}, 1\right)$ ,  $U \sim \chi_{n-1}^2$  con  $U$  y  $W$  variables independientes y

$$h(\rho) = -\sqrt{\frac{\rho(n-1) + 1}{1-\rho}}.$$

Usando lo anterior, si  $Z \sim N(0, 1)$  es una variable aleatoria independiente de  $U$  tenemos que

$$\begin{aligned}
& P(T < k_\alpha \mid H_0) \\
& \leq P \left[ h(\rho) \left( \frac{\sqrt{n}(\mu - q_0)}{\sqrt{c(n-1) + \sigma^2}} - Z \right) / \sqrt{U} < k_\alpha \mid H_0 \right] \\
& \leq P \left[ \frac{\sqrt{n} \frac{q_0 - \mu}{\sigma} + Z}{\sqrt{U}} < k_\alpha \mid H_0 \right] \\
& \leq P \left[ \frac{\sqrt{n}\Phi^{-1}(1-p_0) + Z}{\sqrt{U}} < k_\alpha \right],
\end{aligned}$$

ya que cuando  $H_0$  es verdadera  $\frac{q_0 - \mu}{\sigma} \geq \Phi^{-1}(1-p_0)$ .

El valor crítico  $k_\alpha$  debe satisfacer:  $P \left[ \frac{\sqrt{n}\Phi^{-1}(1-p_0) + Z}{\sqrt{U}} < k_\alpha \right] = \alpha$  para valores dados de  $n, p_0$  y  $\alpha$ . Los valores críticos  $k_\alpha$  son obtenidos por simulación de Monte Carlo.

## 5. Aplicaciones

Bajo el supuesto de  $c > 0$ , las  $X_i$  pueden ser representadas por un modelo lineal del siguiente tipo:

$$X_i = \mu + W + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

donde  $\mu$  es un parámetro de localización y  $W, \varepsilon_1, \dots, \varepsilon_n$  son v.a's independientes con  $W \sim N(0, c)$ , y  $\varepsilon_1, \dots, \varepsilon_n$  iid  $\sim N(0, \tau^2)$  con  $\tau^2 = \sigma^2 - c > 0$ .

## Variables aleatorias intercambiables

Una generalización de las variables aleatorias iid son las variables aleatorias *intercambiables*. Las variables aleatorias  $X_1, \dots, X_n$  son llamadas *simétricamente dependientes* o *variables intercambiables*, si cualquier permutación de cualquier subconjunto de ellas de tamaño  $m$  ( $m \leq n$ ) tiene la misma distribución conjunta.

En el caso de que las variables aleatorias  $X_1, \dots, X_n$  sean intercambiables, se puede ver que su matriz de varianzas y covarianzas es del tipo  $\Sigma = c\mathbf{J} - (c - \sigma^2)\mathbf{I}$ .

En particular, la prueba puede ser aplicada a datos apareados los cuales aparecen en medicina. Aquí se supone que cada paciente tienen dos medidas tomadas antes y después del tratamiento. Con base en este hecho se formula el supuesto de que las medidas tomadas “antes de” y “después de” en un tratamiento dado son intercambiables. Así los pares (Antes de, Despues de) y (Despues de, Antes de) tienen la misma distribución.

## Referencias

Billingsley, P. (1979). *Probability and Measure*. Wiley, New York.

Box, G. E. P, Jenkins, G. M. and Reinsel, G. C.(1994). *Time series analysis. forecasting and control*. Prentice Hall International, Inc., USA.

Brockwell, P. J. and Davis, R.A. (1996). *Introduction to time series and forecasting*. Springer-Verlag, New-York, Inc.

Casella, G. and Berger, R.L. (1990). *Statistical inference*. Duxbury press, Belmont, California.

Chatfield, C. (1999). *The analysis of time series an introduction* (fifth edition), Chapman and Hall/CRC, UK.

Chung, K. L. (1968). *A course in probability theory*. Harcourt, Brace and World, Inc.

Feller, W. (1971). *An introduction to probability theory and its applications* V.2 (Second edition). Wiley, New York.

Halperin, M. (1963). Approximations to the non-central t, with applications, *Technometrics*, **Vol. 5, No. 3**, pp 295-305.

Herstein, I. N. and Winter, D. J. (1988). *A primer on Linear Algebra*, MacMillan publishing company, United States of America.

Kraemer, H. C. and Paik, M. (1979). A central t approximation to the noncentral t-distribution. *Technometrics*, **Vol. 21, No. 3**, pp 357-360.

Maltsev, A. I. (1972). *Fundamentos de álgebra lineal*. Editorial Mir, Moscú.

Mood, A.M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the theory of statistics*, (third edition), McGraw Hill, Singapore.

Owen, D. B. (1968). A Survey of properties and applications of the noncentral t-distribution, *Technometrics*, **Vol. 10, No. 3**, pp 445-473.

Pankrants, A. (1983) *Forecasting with univariate Box-Jenkins models (concepts and cases)*, Wiley, New York.

Parzen, E. (1972). *Procesos estocásticos*, Paraninfo, Madrid-España.

Rohatgi, V. K. (1984). *Statistical inference*, Wiley, New York.

Ross, S. M. (1996). *Stochastic Processes* (second edition), Wiley, New York.

Searle, S. R. (1982). *Matrix algebra useful for statistics*, Wiley, New York.

Wilks, S. S. (1962). *Mathematical statistics*. Wiley, New York.



# Comparing tests of multinormality. A Monte Carlo study

Alexander von Eye<sup>1</sup>

*Michigan State University, 107 D Psychology Building, East Lansing, MI 48824-1116, USA*

## 1. Problem

In an earlier article in the memorias of the foro, von Eye and Bogat (2004) presented two new tests of multinormality (see also von Eye and Bogat, 2004; von Eye and Gardiner, 2004). These tests allow one to identify those sectors in the multivariate space in which there are more (or fewer) cases than one would expect based on the null hypothesis that the variables that span the multivariate space are normally distributed. In addition, these tests allow one to come to an overall evaluation of this null hypothesis. Although these tests use Pearson  $X^2$  statistics to test the null hypothesis, and although the characteristics of Pearson's  $X^2$  are well known, it is largely unknown whether they are sensitive to various forms of violation of multinormality. There exists one study, in which it was shown that these tests are sensitive to symmetry violations (von Eye, Bogat, and von Eye, 2005). The sensitivity to other violations, for instance skewness or kurtosis is unknown.

Therefore, a simulation study was performed in which various transformations were performed on random data. The new tests and the well known Mardia tests of multivariate skewness and kurtosis were used to examine the thus transformed data. In the following sections, the simulation study and its results are presented.

## 2. The tests used in the simulation

The simulation included Mardia's (1970, 1980) tests of multivariate skewness and kurtosis, and von Eye *et al.* (2005) overall and sector tests.

---

<sup>1</sup>voneye@msu.edu

(A) Mardia's measure of multivariate skewness is

$$b_{1d} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N r_{ij}^3,$$

where  $r_{ij}^3$  is the *Mahalanobis angle* (Mardia, 1980, p. 310) between the vectors  $x_i - \bar{x}$  and  $x_j - \bar{x}$ . The limiting distribution of  $Nb_{1d} / 6$  is a  $\chi^2$  distribution with  $df = d(d + 1)(d + 2)/6$ , and  $d$  is the number of variables that span the multivariate space.

(B) Mardia's measure of multivariate kurtosis is

$$b_{2d} = \frac{1}{N} \sum_{i=1}^N r_i^4,$$

where  $r_i^4$  is the *Mahalanobis distance* of case  $x_i$  from its mean,  $\bar{x}$ . The limiting distribution of  $\sqrt{N} \frac{(b_{2d} - d(d + 2))}{\sqrt{8d(d + 2)}}$  is normal. For both  $b_{1d}$  and  $b_{2d}$ , the cases under study are assumed to be *iid*.

(C) von Eye *et al.* (2005) overall and sector tests. For these, the score ranges of each variable first are split into two or more segments. Second, the thus segmented variables are crossed, and the probability of each of the resulting sectors is estimated under the assumption of a multinormal distribution, using an algorithm proposed by Genz (1992; for more detail, see von Eye and Gardiner, 2004). Finally, for each sector, the expected frequency is calculated and compared with the corresponding observed frequency, under the null hypothesis  $E[o_{i,j,\dots,k}] = e_{i,j,\dots,k}$ , where  $o_{i,j,\dots,k}$  is the observed and  $e_{i,j,\dots,k}$  the expected frequency for sector  $i,j,\dots,k$ , with  $df = 1$ .

(D) To perform the comparison, one can use the Pearson

$$X^2 = \frac{(o_{i,j,\dots,k} - e_{i,j,\dots,k})^2}{e_{i,j,\dots,k}}$$

under  $df = 1$ . Summed, the  $X^2$  components yield the test statistic for the overall test. This test has  $df = (\prod_{l=1}^d c_l) - 2d - d_{cov} - 1$ , where  $d$  is the number of variables that span the

multivariate space,  $c_j$  is the number of segments of the  $j$ th variable, and  $d_{\text{cov}} = \binom{d}{2}$ , because, in the typical case, all correlations are taken into account.

### 3. The simulation

In the simulation, five variables were varied. The first was *Type of Distribution*. Specifically, the following five distributions were created (for more detail, see von Eye, Bogat, and von Eye, 2005): (1) normally distributed variates; (2) uniformly distributed variates, ranging from 0 to 1; (3) logarithmically-transformed variates (from a uniform distribution) (4) uniformly distributed variates that were subjected to the substitute of the inverse Laplace transformation; and (5) uniformly distributed variates that were subjected to the cube root transformation.

In addition to *type of distribution* (i), the following data characteristics were varied:

- (ii) The *sample size* varied from 50 to 1500, in steps of 50.
- (iii) The *number of segments* of each variable varied from 2 to 5, in steps of 1. The number of segments was always the same for all variables.
- (iv) The *correlation*,  $\rho$ , among variables. Specifically, variate  $x_j + 1$  was correlated with variate  $x_j$  by  $x_j + 1 = 0.5 \cdot x_j + x_j + 1 \cdot \rho$ .  $\rho$  assumed the four values 0, 0.1, 0.2, and 0.3.
- (v) The *number of variables* varied from 2 to 5, in steps of 1.

For each of the 9600 simulated conditions, it was scored whether the four comparison tests (A,B,C,D) identified deviations from multinormality or not.

### 3.1. Models and Results

(A) For *skewness*, the following logit model was estimated:

$$\begin{aligned} \text{skewness} = & \alpha + \beta_i D + \beta_j N + \beta_k \text{TRANSFORM} + \beta_l \text{RHO} \\ & + \beta_{ij} (D \times N) + \beta_{ik} (D \times \text{TRANSFORM}) + \beta_{il} (D \times \text{RHO}) \\ & + \beta_{jk} (N \times \text{TRANSFORM}) + \beta_{jl} (N \times \text{RHO}) \\ & + \beta_{kl} (\text{TRANSFORM} \times \text{RHO}), \end{aligned}$$

with  $D$  = number of variables,  $N$  = sample size,  $\text{TRANSFORM}$  = type of distribution, and  $\text{RHO}$  = correlation. For this model, the McFadden  $R^2$  was estimated to be 0.71. In the following sections, we report a selection of results.

Number of variables. The number of violations increased with the number of variables.

Type of distribution. The type of distribution showed the expected effects; see Figure 1.

Sample size by type of distribution. Only for very small samples, the number of distributions flagged as violating multinormality was smaller.

(B) For *kurtosis*, the same logit model was estimated as for skewness. The McFadden  $R^2$  was 0.85.

Type of distribution. The expected effects resulted. See Figure 2.

(C) For the *Sector Test*, a linear model was estimated with all main effects and all first-order interactions;  $R^2 = 0.84$ .

Number of variables. number of violations increased with number of variables.

Sample size. Number of violations increased with sample size.

Number of segments. number of violations increased with umber of segments.

Type of Distribution. The largest numbers of deviant sectors were found for the uniform and the inverse Laplace-transformed distributions, followed by the log-transformed distribution. The number of deviant sectors found for the normal and the cube root-transformed distributions was minimal.

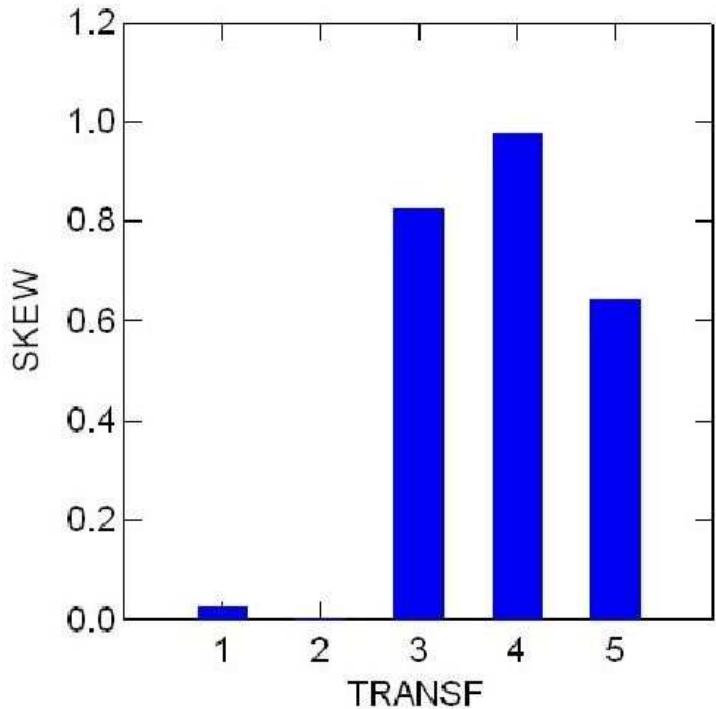


Figura 1: Type of distribution-effect on skewness identification, with 1 = normal, 2 = uniform, 3 = log-transformed, 4 = inverse Laplace-transformed, and 5 = cube root-transformed.

Number of Variables by Sample Size. The sample size effect is weakest for 2 variables and strongest for 5 variables.

Number of Segments by Number of Variables. The number of variables-effect becomes strong only for 4 or more variables.

Number of Variables by Type of Distribution. The number of variables-effect is stronger for the uniform and the inverse Laplace-transformed distributions than for the remaining three distributions.

Number of Segments by Sample Size. The number of segments effect is stronger for larger samples.

Sample size by type of Transformation. There is no sample size effect for the normal and the cube root-transformed distributions. Number of Segments by Type of Distribution. The number of segments-effect is strongest for the uniform and the inverse Laplace-transformed distributions.

(D) For the *overall  $X^2$  test*, a no-interaction model was estimated, because empty subtables

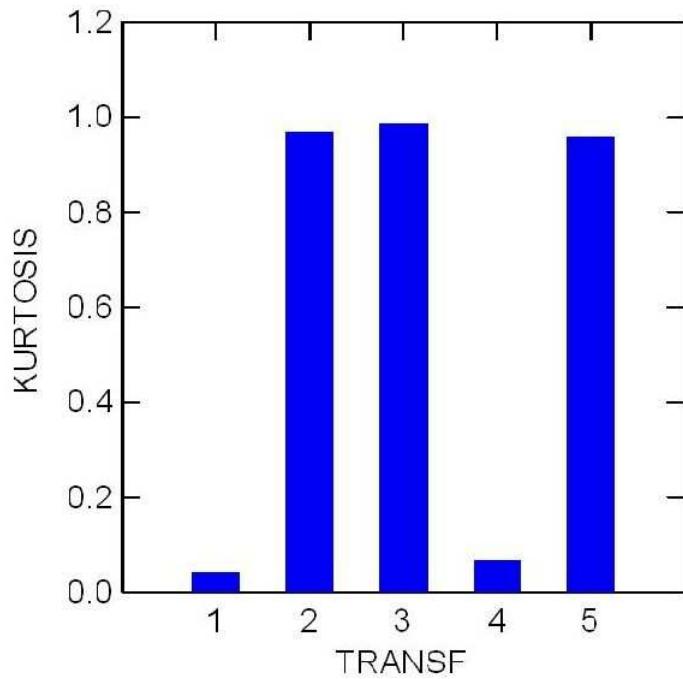


Figura 2: Type of distribution-effect on kurtosis identification, with labels as in Figure 1.

prevented the model with all pairwise interactions from converging. The model also included the main effect of the number of segments; the McFadden  $R^2$  was 0.78.

Number of variables. The number of violations increased with the number of variables.

Number of segments. The number of violations increased slightly with the number of segments.

Type of distribution. See Figure 3.

## 4. Discussion

We now ask, which test to use in the concrete data analysis. Based on the simulation results, we note that, if kurtosis or skewness are of particular interest, the specialized tests perform well. The  $X^2$  overall and the sector tests are *omnibus to all types of violations simulated in this study*, including skewness and kurtosis (and axial symmetry; see von Eye, Bogat and von Eye, 2005). Therefore, they are preferable when specific violations are not targeted. When the location of violations needs to be identified, the sector test is currently the only choice.

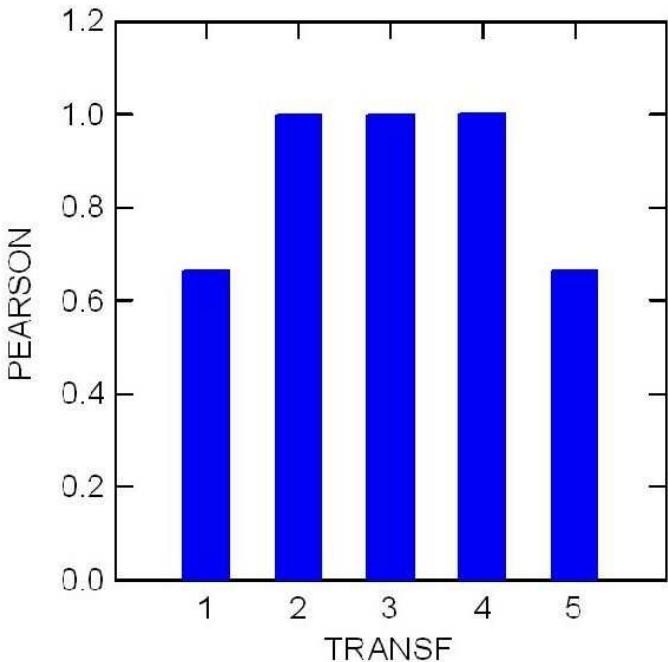


Figura 3: Type of distribution-effect on identification of multinormality violations by the overall  $X^2$  test, with labels as in Figure 1

## References

- Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1, 141 - 149.
- Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519 - 530.
- Mardia, K.V. (1980). Tests of univariate and multivariate normality. In P.R. Krishnaiah (Ed.), *Handbook of statistics* (vol. 1; pp 279 - 320). Amsterdam: North Holland.
- von Eye, A., and Bogat, G.A. (2004). Testing the assumption of multivariate normality. *Psychology Science*, 46, 243 - 258.
- von Eye, A., and Bogat, G.A. (2005). Identifying sectors of deviations from multinormality. In Contreras Cristán, A., Domínguez Molina, A., and Anaya Izquierdo, K. (eds.), *Memorias*

*del XIX Foro Nacional de Estadística.* INEGI, Mexico.

von Eye, A., Bogat, G.A., and von Eye, M. (2005). Multinormality and symmetry: A comparison of two statistical tests. *Psychology Science* (in press).

von Eye, A., and Gardiner, J.C. (2004). Locating deviations from multivariate normality. *Understanding Statistics*, 3, 313 - 331.

Esta publicación consta de 941 ejemplares y se terminó de imprimir en el mes de agosto de 2006 en los talleres gráficos del **Instituto Nacional de Estadística, Geografía e Informática**  
Av. Héroe de Nacozari Sur Núm. 2301, Puerta 11, Nivel Acceso  
Fracc. Jardines del Parque, CP 20270  
Aguascalientes, Ags.  
**México**