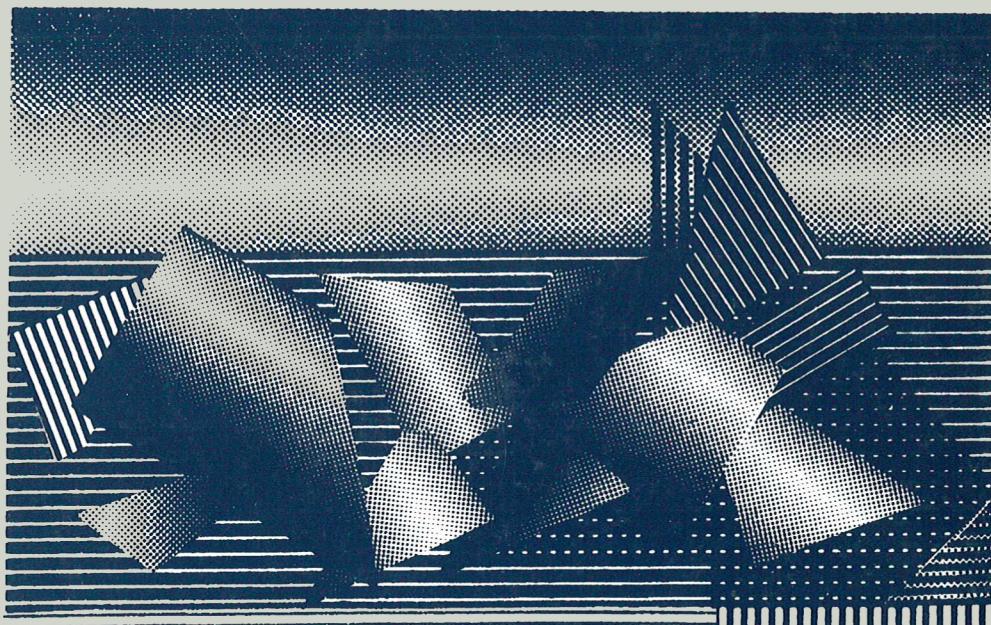


MEMORIA DEL VII FORO NACIONAL DE ESTADÍSTICA

Universidad de las Américas Puebla

7-11 de septiembre de 1992

Diseño: Lidia Lagarda



Análisis de la varianza

AME - CEC-CP - CIMAT - CONACYT - IIMAS-UNAM
INEGI - ITAM -ITESM-CEM - OLIVETTI - UDLA

VII FORO NACIONAL DE ESTADISTICA
7 AL 11 DE SEPTIEMBRE, 1992

SEDE:
UNIVERSIDAD DE LAS AMERICAS, PUEBLA
CHOLULA, PUEBLA

Comité Organizador:

Javier Alagón (ITAM)
Francisco Burguete (ITESM-CEM)
Alberto Castillo (CEC)
Antonio González (UDLA)
Víctor Guerrero (ITAM)
Manuel Mendoza (ITAM)

Comité de Programa y Editorial:

Esteban Burguete (UDLA)
Antonio González (UDLA)
David Sotres (CEC)
Guillermo Zárate (ITAM)

Con el apoyo de las siguientes instituciones:

AME CEC-CP CIMAT CONACYT IIMAS-UNAM INEGI
ITAM ITESM-CEM OLIVETTI UDLA

PRESENTACION

El VII FORO NACIONAL DE ESTADISTICA se llevó a cabo del 7 al 11 de septiembre de 1992 en la Universidad de las Américas, ubicada en Cholula, Puebla. En la organización participaron otras instituciones como: la Asociación Mexicana de Estadística, el Centro de Estadística y Cálculo-Colegio de Postgraduados (Chapingo), el Centro de Investigación en Matemáticas (Guanajuato), el Consejo Nacional de Ciencia y Tecnología, el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la Universidad Nacional Autónoma de México, el Instituto Nacional de Estadística, Geografía e Informática, el Instituto Tecnológico Autónomo de México, el Instituto Tecnológico de Estudios Superiores de Monterrey-Campus Estado de México y la Compañía Olivetti.

En este Foro se contó con la participación de ponentes mexicanos y extranjeros. Los artículos presentados en esta Memoria, son parte de los trabajos que se expusieron en dicho encuentro.

Es importante mencionar que no se realizó una revisión científica ni técnica de los artículos publicados en esta Memoria.

COMITE EDITORIAL

INDICE

pág.

Robustez de las medidas de influencia en regresión. F.J. GIRON, M.L. MARTINEZ	1
Estimación mediante simulaciones de la potencia y el nivel de significancia de 5 pruebas de comparaciones múltiples y del análisis de varianza. O. CAMACHO, R. CAMACHO	11
Independencia condicional y diagramas de influencia probabilista; dos aplicaciones. M. MARTINEZ, J.A. MONTANO, M.M. OJEDA	29
Modelos de componentes de varianza para investigación educativa. M.M. OJEDA, J.A. MONTANO	40
Una propuesta para la estimación en poblaciones finitas basada en la distancia entre las funciones de distribución empírica de la población y de la muestra. A. CASTILLO	51
La aplicación del modelo de alta y baja movilidad en el estudio de la migración. G. VAZQUEZ	61
El estudio de la mortalidad en el pasado mediante la aplicación del modelo de poblaciones estables: Istmo de Tehuantepec, siglo XVIII. J.R. GUTIERREZ, G. VAZQUEZ	75
Extremos bivariados (versión procesos puntuales). H. GUTIERREZ	92
Ilustración de la no robustez de la prueba de t a distribuciones con colas largas usando datos reales. D. SOTRES	103
Un modelo para la estimación de parámetros en gallinas de postura. E. BURGUETE, J.F. BURGUETE, J.G. HERRERA	112
Estandarización no lineal de la longitud de señales que tienen un mismo origen. J.E. ROHEN	119
Cálculo de riesgos de contraer cáncer. J.F. BURGUETE, E. BURGUETE	131
Capacitación en calidad: Administración vs. Estadística. H. GUTIERREZ, R. DE LA VARA	138
Diseño de experimentos en el modelo de Michaelis-Menten. J. GAYTAN, E. RENDON	148

Robustez de las medidas de influencia en regresión

F. J. GIRÓN y M. L. MARTÍNEZ *

RESUMEN

El análisis de las observaciones anómalas (outliers) y su influencia en la estimación y el contraste de hipótesis de los parámetros de regresión y de la varianza común de la fuente de error ha sido un tema de bastante interés en la última década, tanto desde el punto de vista clásico como desde el bayesiano, y se han establecido algunas medidas de la influencia para el modelo de regresión con errores normales.

El principal objetivo de este trabajo es doble. En primer lugar, se demuestra que la mayoría de las medidas usuales para detectar outliers y para medir su influencia se pueden deducir utilizando simples técnicas bayesianas; y después se demuestra que la mayoría de estas medidas son débilmente robustas con respecto a una amplia clase de fuentes de error que incluye el caso normal. En particular, se demuestra que para la clase de los modelos de regresión con errores distribuidos según una mezcla de normales respecto al parámetro de escala, algunas de estas medidas tienen distribuciones independientes de la distribución de mezcla, cuando no existe información a priori.

Palabras clave: medidas de influencia; mixturas de distribuciones; modelos lineales; observaciones influyentes; observaciones anómalas; residuos jackknifed; residuos ordinarios; filtro de Kalman.

Clasificación AMS: 62M20, 62F15, 62J05.

* Departamento de Estadística e I.O. Facultad de Ciencias. Campus de Teatinos s/n. Universidad de Málaga. 29071-MÁLAGA (ESPAÑA).

Este trabajo ha sido subvencionado parcialmente por la *Dirección General de Investigación Científica y Técnica (DGICYT)* como parte del proyecto de referencia PB87-0607-C02-02 y la *Consejería de Educación de la Junta de Andalucía*.

1. INTRODUCCIÓN.

El análisis de residuos, el estudio de observaciones anómalas o influyentes y la definición de medidas de influencia, han sido desarrollados desde diferentes puntos de vista pero quizás de manera más amplia para el modelo general de regresión lineal. Muchas de las medidas de influencia y de los procedimientos para detectar observaciones anómalas se han basado en ideas más o menos intuitivas dependiendo del modelo particular que se estuviera estudiando y de las posibles aplicaciones que éste pudiera tener. No es de asombrar la cantidad de tests que se pueden encontrar en la literatura, tests que aplicados a un mismo problema pueden dar resultados distintos e incluso contradictorios (Barnett y Lewis (1984)).

En este trabajo se consideran estos problemas desde el punto de vista bayesiano intentando dar un enfoque común a todos ellos. Enfoque que aplicaremos, en particular, al modelo de regresión usual, con errores normales, y se demuestra que la mayoría de las medidas usuales para detectar observaciones anómalas y para medir su influencia se deducen de forma natural.

A continuación se estudia la invariancia o robustez de estas medidas en los modelos de regresión cuando se consideran fuentes de error no normales, en concreto, errores distribuidos según una mixtura de normales respecto al parámetro de escala. Se obtiene que las medidas para la detección de observaciones anómalas e influyentes son débilmente robustas para el parámetro de regresión θ , cuando no se tiene información a priori, ya que su distribución es independiente de la distribución de mezcla considerada. Sin embargo, las medidas de influencia para la varianza σ^2 no son robustas.

2. DEFINICIONES BÁSICAS.

Sea $\{f(y|\phi); \phi \in \Phi\}$ un modelo estadístico y sea $\mathbf{y} = (y_1, \dots, y_n)'$ una muestra aleatoria del modelo, donde ϕ es un vector de parámetros de dimensión p , que puede incluir o no parámetros marginales. Notaremos por D al conjunto formado por \mathbf{y} y otras variables concomitantes que puedan presentarse en el modelo. Nos centraremos en dos problemas diferentes pero relacionados que se suelen presentar en un análisis de datos: uno el de decidir cuándo una o más observaciones del conjunto de datos D puedan considerarse como observaciones anómalas; otro problema es el estudiar la posible influencia que estas observaciones u otras puedan tener sobre los estimadores de los parámetros o sobre alguna función de ellos.

En relación con el primer problema, diremos que la observación x_i es inconsistente con el resto de las observaciones $D_{(i)}$, cuando no se puede predecir a partir de éstas. Desde el punto de vista bayesiano esto se refleja en un valor muy pequeño de la densidad a posteriori para x_i

$$p(x|D_{(i)}) = \int p(x|\phi) p(\phi|D_{(i)}) d\phi.$$

Basándonos en esta idea intuitiva damos la siguiente definición.

Definición 1. Un subconjunto de observaciones con subíndices en \mathcal{I} , con $\mathcal{I} = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$, es un *conjunto de observaciones anómalas* si el vector $y_{\mathcal{I}}$ no pertenece a la región M.D.P. de $p(z_1, \dots, z_m | D_{(\mathcal{I})})$ de contenido probabilístico $1 - \alpha$.

El cálculo de las regiones M.D.P. suele ser complicado; sin embargo, en muchas ocasiones, como por ejemplo cuando existe un estadístico suficiente predictivo este cálculo de una región M.D.P. para y_1, \dots, y_m se reduce a encontrar una región M.D.P. para la distribución predictiva $p(t_m | D_{(\mathcal{I})})$, donde $t_m = t_m(y_1, \dots, y_m)$ es un estadístico suficiente.

Dado que el modelo se rige por un determinado vector de parámetros, nos planteamos estudiar cuándo una o más observaciones serán influyentes para la estimación de dicho vector. La definición de observación influyente respecto a la estimación de algún vector de parámetros, está relacionada con la definición de estimador bayes, de aquí que para evitar la ambigüedad recordemos que un estimador bayesiano se puede considerar como una función del espacio de las distribuciones (a priori y(o) a posteriori) en el conjunto de los posibles valores del parámetro; por tanto, un estimador Bayes $\tilde{\phi}$ es una función medible

$$\tilde{\phi}: p(\phi | D) \mapsto \tilde{\phi}(p(\phi | D)) \in \Phi;$$

tal como la media, la moda o cualquier otro estimador que se obtenga, por ejemplo, minimizando una función de pérdida dada, condicionado a su existencia.

Supongamos que $\tilde{\phi}$ es un estimador Bayes bien definido. La siguiente definición nos da un test para decidir cuándo una observación será influyente respecto a $\tilde{\phi}$.

Definición 2. La observación y_i es *influyente respecto a $\tilde{\phi}$* si $\tilde{\phi}(p(\phi | D))$ no pertenece a la región M.D.P. de la distribución a posteriori $p(\phi | D_{(i)})$ de contenido probabilístico $1 - \alpha$.

Las definiciones sobre observaciones influyentes propuestas con anterioridad, como las basadas en la divergencia de Kullback-Leibler entre $p(\phi | D)$ y $p(\phi | D_{(i)})$, permiten ordenar las observaciones en cuanto a su influencia pero no proporcionan un test para poder afirmar si una observación o un subconjunto de ellas es influyente, hecho que si es posible con nuestra definición.

3. APPLICACIÓN AL MODELO DE REGRESIÓN.

Sea el modelo lineal de regresión

$$y_i = \mathbf{x}_i' \boldsymbol{\theta} + u_i;$$

donde y_1, \dots, y_n son variables aleatorias observables, u_1, \dots, u_n son errores independientes normalmente distribuidos con media 0 y varianza σ^2 desconocida, \mathbf{x}_i' son vectores ($k \times 1$) de variables independientes y $\boldsymbol{\theta}$ es un vector ($k \times 1$) de parámetros de regresión desconocidos.

Notaremos con el subíndice \mathcal{I} a cualquier subconjunto de m observaciones, así como al correspondiente subvector o submatriz de cualquier vector o matriz del modelo y con (\mathcal{I}) al resto del vector o submatriz correspondientes. Además si representamos por $\hat{\theta}$ y $\hat{\sigma}^2$ a los estimadores mínimo cuadráticos usuales de θ y σ^2 respectivamente, entonces $\hat{\theta}_{(\mathcal{I})}$ y $\hat{\sigma}_{(\mathcal{I})}^2$ serán los estimadores mínimo cuadráticos cuando se suprime el conjunto de observaciones cuyos subíndices pertenecen a \mathcal{I} .

Si ahora consideramos la distribución a priori no informativa usual para θ y σ^2 , obtenemos, aplicando el lema 1 del apéndice, que la distribución a posteriori conjunta de θ y σ^2 dado $D_{(\mathcal{I})}$, considerando que $n - k - m > 0$, es

$$\theta, \sigma^2 | D_{(\mathcal{I})} \sim NGaI\left(\hat{\theta}_{(\mathcal{I})}, (X'_{(\mathcal{I})} X_{(\mathcal{I})})^{-1}, \frac{n - k - m}{2} \hat{\sigma}_{(\mathcal{I})}^2, \frac{n - k - m}{2}\right).$$

A partir de aquí y teniendo en cuenta las propiedades de la distribución normal gamma-invertida, es fácil obtener las distribuciones marginales a posteriori de θ y σ^2

$$\theta | D_{(\mathcal{I})} \sim t_k(\hat{\theta}_{(\mathcal{I})}, \hat{\sigma}_{(\mathcal{I})}^2 (X'_{(\mathcal{I})} X_{(\mathcal{I})})^{-1}, n - k - m), \quad (1)$$

$$\sigma^2 | D_{(\mathcal{I})} \sim GaI\left(\frac{n - k - m}{2} \hat{\sigma}_{(\mathcal{I})}^2, \frac{n - k - m}{2}\right). \quad (2)$$

Como el vector formado por las restantes m observaciones, digamos \mathbf{z} , se distribuye dado θ y σ^2 , según

$$\mathbf{z} | \theta, \sigma^2 \sim N(X_{\mathcal{I}} \theta, \sigma^2 I_m);$$

entonces a partir de las propiedades de la distribución normal gamma-invertida, la distribución a posteriori conjunta de \mathbf{z} y σ^2 dado $D_{(\mathcal{I})}$ es

$$\mathbf{z}, \sigma^2 | D_{(\mathcal{I})} \sim NGaI\left(X_{\mathcal{I}} \hat{\theta}_{(\mathcal{I})}, V_{\mathcal{I}}, \frac{n - k - m}{2} \hat{\sigma}_{(\mathcal{I})}^2, \frac{n - k - m}{2}\right);$$

donde $V_{\mathcal{I}} = I_m + X_{\mathcal{I}} (X'_{(\mathcal{I})} X_{(\mathcal{I})})^{-1} X'_{\mathcal{I}}$.

La distribución predictiva de \mathbf{z} dado $D_{(\mathcal{I})}$ vendrá entonces dada por

$$\mathbf{z} | D_{(\mathcal{I})} \sim t_m(X_{\mathcal{I}} \hat{\theta}_{(\mathcal{I})}, \hat{\sigma}_{(\mathcal{I})}^2 V_{\mathcal{I}}, n - k - m); \quad (3)$$

y de las propiedades de la distribución t -multivariante se sigue que la región M.D.P. de contenido probabilístico $(1 - \alpha)$ es el elipsoide m -dimensional dado por

$$\left\{ \mathbf{z} \in \mathbb{R}^m; \frac{(\mathbf{z} - X_{\mathcal{I}} \hat{\theta}_{(\mathcal{I})})' V_{\mathcal{I}}^{-1} (\mathbf{z} - X_{\mathcal{I}} \hat{\theta}_{(\mathcal{I})})}{m \hat{\sigma}_{(\mathcal{I})}^2} \leq F(m, n - k - m; 1 - \alpha) \right\};$$

donde $F(m, n - k - m; 1 - \alpha)$ es el percentil $1 - \alpha$ de la distribución F centrada con m y $n - k - m$ grados de libertad.

Entonces de acuerdo con la definición 1, el subconjunto de observaciones cuyos subíndices pertenezcan a \mathcal{I} , $\mathbf{y}_{\mathcal{I}}$, es anómalo si

$$\frac{(\mathbf{y}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}\hat{\theta}_{(\mathcal{I})})' \mathbf{V}_{\mathcal{I}}^{-1} (\mathbf{y}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}\hat{\theta}_{(\mathcal{I})})}{m\hat{\sigma}_{(\mathcal{I})}^2} \geq F(m, n - k - m; 1 - \alpha).$$

Teniendo en cuenta que $\mathbf{r}_{\mathcal{I}} = \mathbf{y}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}\hat{\theta}_{(\mathcal{I})}$ es el vector de errores de predicción que se obtiene aplicando el filtro de Kalman al modelo lineal usual o, como es conocido generalmente, el vector de residuos jackknifed, el test anterior para detectar si el subconjunto de observaciones cuyos subíndices pertenecen a \mathcal{I} es anómalo, es equivalente a

$$DP(\mathcal{I}) = \frac{\mathbf{r}_{\mathcal{I}}' \mathbf{V}_{\mathcal{I}}^{-1} \mathbf{r}_{\mathcal{I}}}{m\hat{\sigma}_{(\mathcal{I})}^2} \geq F(m, n - k - m; 1 - \alpha); \quad (4)$$

donde por $DP(\mathcal{I})$ designaremos a lo que llamaremos la *distancia predictiva* entre $\mathbf{y}_{\mathcal{I}}$ y su predicción $\mathbf{X}_{\mathcal{I}}\hat{\theta}_{(\mathcal{I})}$.

Aplicando identidades matriciales o el filtro de Kalman, que se deduce del lema 1 del apéndice, se tiene que

$$DP(\mathcal{I}) = \frac{\mathbf{e}_{\mathcal{I}}'(I_m - \mathbf{H}_{\mathcal{I}})^{-1} \mathbf{e}_{\mathcal{I}}}{m\hat{\sigma}_{(\mathcal{I})}^2} \geq F(m, n - k - m, 1 - \alpha). \quad (5)$$

donde $\mathbf{e}_{\mathcal{I}} = \mathbf{y}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}\hat{\theta}$ son los residuos ordinarios y $\mathbf{H}_{\mathcal{I}} = \mathbf{X}_{\mathcal{I}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{\mathcal{I}}$.

En el caso particular de considerar una sola observación se tendrá que la observación i -ésima será considerada como anómala si

$$\left| \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{v_i}} \right| \geq t(n - k - 1; 1 - \alpha).$$

Esto demuestra la importancia del análisis de residuos en el estudio individual de observaciones anómalas.

3.1. Medidas de influencia para el vector θ .

A partir de la distribución a posteriori de $\theta | D_{(\mathcal{I})}$, dada por (3) y de las propiedades de la distribución t -Student multivariante, se obtiene fácilmente que las regiones M.D.P. para θ , de contenido probabilístico $1 - \alpha$, son de la forma

$$\left\{ \theta \in \mathbb{R}^k; \frac{(\theta - \hat{\theta}_{(\mathcal{I})})' (\mathbf{X}'_{(\mathcal{I})} \mathbf{X}_{(\mathcal{I})})(\theta - \hat{\theta}_{(\mathcal{I})})}{k\hat{\sigma}_{(\mathcal{I})}^2} \leq F(k, n - k - m; 1 - \alpha) \right\}.$$

Por lo tanto y de acuerdo con la definición 4, el subconjunto de observaciones cuyos subíndices estén en \mathcal{I} , será influyente para θ , con respecto al estimador $\hat{\theta}$ (en este caso la moda de la distribución a posteriori o la media si $n - k - m > 1$) si

$$DB(\mathcal{I}) = \frac{(\hat{\theta} - \hat{\theta}_{(\mathcal{I})})'(X'_{(\mathcal{I})} X_{(\mathcal{I})})(\hat{\theta} - \hat{\theta}_{(\mathcal{I})})}{k \hat{\sigma}_{(\mathcal{I})}^2} \geq F(k, n - k - m; 1 - \alpha); \quad (6)$$

donde $DB(\mathcal{I})$ representa la *distancia bayesiana*, que puede considerarse como una especie de distancia entre el estimador mínimo cuadrático usual, $\hat{\theta}$, y el estimador jackknifed $\hat{\theta}_{(\mathcal{I})}$.

Es de señalar que esta expresión está relacionada con la distancia de Welsch (1982) para la detección de observaciones influyentes.

Podemos dar expresiones más sencillas de $DB(\mathcal{I})$, pero equivalentes, en función de los dos tipos de residuos anteriormente definidos

$$DB(\mathcal{I}) = \frac{r'_\mathcal{I} V_\mathcal{I}^{-1} (I_m - V_\mathcal{I}^{-1}) r_\mathcal{I}}{k \hat{\sigma}_{(\mathcal{I})}^2} = \frac{e'_\mathcal{I} H_\mathcal{I} (I_m - H_\mathcal{I})^{-1} e_\mathcal{I}}{k \hat{\sigma}_{(\mathcal{I})}^2}.$$

Para el caso de una sola observación, la i -ésima, la distancia bayesiana, $DB(i)$, se puede escribir en términos del residuo jackknifed estudiantizado t_i^* y el elemento diagonal h_{ii} de la matriz H en la forma

$$DB(i) = \frac{t_i^{*2} h_{ii}}{k}.$$

Por tanto, la observación i -ésima es influyente para θ si

$$\frac{t_i^{*2} h_{ii}}{k} \geq F(k, n - k - 1; 1 - \alpha).$$

4. ROBUSTEZ DE LAS MEDIDAS.

Si en el modelo de regresión en lugar de considerar los errores u_i independientes se considera que tienen simetría esférica $\forall n$, lo que implica intercambiabilidad, entonces por el teorema dado por Kingman (1972) (véase también Smith (1981)), $u \sim \int N(0, \lambda \sigma^2 I_n) dF(\lambda)$. Esto implica que las u_i son incorreladas si F posee momento de orden 1. Además esta clase de distribuciones incluye a familias importantes como la t-Student.

Sea, entonces, el modelo lineal de regresión

$$y_i = x'_i \theta + u_i;$$

donde y_1, \dots, y_n son variables aleatorias observables, u_1, \dots, u_n son errores que se distribuyen conjuntamente según la mixtura

$$u \sim \int N(0, \lambda \sigma^2 I_n) dF(\lambda)$$

con σ^2 desconocido y $\lambda > 0$, x_i son vectores ($k \times 1$) de variables independientes y θ es un vector ($k \times 1$) de parámetros de regresión desconocidos.

Si suponemos que a priori la distribución sobre (θ, σ^2) es $NGaI(\mathbf{m}_0, \mathbf{C}_0; a_0, p_0)$ y que u se distribuye según la mixtura ya dada, aplicando el lema 2 del apéndice, se tiene que la distribución de $(\theta, \sigma^2 | D)$ viene dada por

$$\int NGaI(\mathbf{m}_n(\lambda), \mathbf{C}_n(\lambda); a_n(\lambda), p_n) dF(\lambda | D)$$

donde

$$\begin{aligned} \mathbf{m}_n(\lambda) &= \mathbf{m}_0 + \mathbf{C}_0 \mathbf{X}' (\lambda \mathbf{I}_n + \mathbf{X} \mathbf{C}_0 \mathbf{X}')^{-1} (\mathbf{y} - \mathbf{X} \mathbf{m}_0) \\ \mathbf{C}_n(\lambda) &= \mathbf{C}_0 - \mathbf{C}_0 \mathbf{X}' (\lambda \mathbf{I}_n + \mathbf{X} \mathbf{C}_0 \mathbf{X}')^{-1} \mathbf{X} \mathbf{C}_0 \\ a_n(\lambda) &= a_0 + \frac{1}{2} (\mathbf{y} - \mathbf{X} \mathbf{m}_0)' (\lambda \mathbf{I}_n + \mathbf{X} \mathbf{C}_0 \mathbf{X}')^{-1} (\mathbf{y} - \mathbf{X} \mathbf{m}_0) \\ p_n &= p_0 + \frac{n}{2} \\ dF(\lambda | D) &\propto f(y_1, \dots, y_n | \lambda) dF(\lambda) \propto t_n(\mathbf{X} \mathbf{m}_0, (\lambda \mathbf{I}_n + \mathbf{X} \mathbf{C}_0 \mathbf{X}') \frac{a_0}{p_0}; 2p_0) dF(\lambda). \end{aligned}$$

A partir de las propiedades de la distribución normal gamma-invertida se tiene que

$$\begin{aligned} \theta | D &\sim \int t_k(\mathbf{m}_n(\lambda), \mathbf{C}_n(\lambda) \frac{a_n(\lambda)}{p_n}, 2p_n) dF(\lambda | D) \\ \sigma^2 | D &\sim \int GaI(a_n(\lambda), p_n) dF(\lambda | D) \end{aligned}$$

Para el caso particular de considerar la distribución no informativa usual a priori sobre el par (θ, σ^2) , tendremos que las distribuciones anteriores se transforman en

$$\begin{aligned} \theta, \sigma^2 | D &\sim \int NGaI\left(\hat{\theta}, \lambda(\mathbf{X}' \mathbf{X})^{-1}, \frac{n-k}{2\lambda} \hat{\sigma}^2, \frac{n-k}{2}\right) dF(\lambda) \\ \theta | D &\sim t_k(\hat{\theta}, \hat{\sigma}^2 (\mathbf{X}' \mathbf{X})^{-1}, n-k) \\ \sigma^2 | D &\sim \int GaI\left(\frac{n-k}{2\lambda} \hat{\sigma}^2, \frac{n-k}{2}\right) dF(\lambda) \end{aligned} \tag{7}$$

Es de destacar que en este caso la distribución de $\theta | D$ no depende de la distribución de mezcla por lo que el estudio de las medidas de influencia y observaciones influyentes se reducirán al caso ya estudiado. Además $dF(\lambda | D) \propto dF(\lambda)$.

4.1. Estudio de las observaciones anómalas.

Para este estudio no precisaremos qué distribución de mezcla se está considerando, pues la distribución predictiva para las observaciones que se obtiene no depende de ella y por lo tanto, los resultados serán análogos a los del modelo usual de regresión.

Atendiendo a los resultados expuestos en el apartado anterior la distribución a posteriori de θ dado $D_{(\mathcal{I})}$, considerando que $n - k - m > 0$, es

$$\theta | D_{(\mathcal{I})} \sim t_k(\hat{\theta}_{(\mathcal{I})}, \hat{\sigma}_{(\mathcal{I})}^2 (X'_{(\mathcal{I})} X_{(\mathcal{I})})^{-1}, n - k - m)$$

Si representamos por z al vector formado por las restantes m observaciones, se tiene que

$$z, \sigma^2 | \lambda, D_{(\mathcal{I})} \sim NGaI \left(X_{\mathcal{I}} \hat{\theta}_{(\mathcal{I})}, \lambda V_{\mathcal{I}}, \frac{n - k - m}{2\lambda} \hat{\sigma}_{(\mathcal{I})}^2, \frac{n - k - m}{2} \right),$$

donde $V_{\mathcal{I}} = I_m + X_{\mathcal{I}} (X'_{(\mathcal{I})} X_{(\mathcal{I})})^{-1} X'_{(\mathcal{I})}$.

Y teniendo en cuenta las propiedades de la distribución normal gamma-invertida, obtenemos la distribución predictiva para z

$$z | \lambda, D_{(\mathcal{I})} \sim t_m(X_{\mathcal{I}} \hat{\theta}_{(\mathcal{I})}, \hat{\sigma}_{(\mathcal{I})}^2 V_{\mathcal{I}}, n - k - m),$$

que es una distribución independiente de λ y por lo tanto

$$z | D_{(\mathcal{I})} \sim t_m(X_{\mathcal{I}} \hat{\theta}_{(\mathcal{I})}, \hat{\sigma}_{(\mathcal{I})}^2 V_{\mathcal{I}}, n - k - m),$$

que coincide con (3), la obtenida en el caso del modelo lineal usual, y por tanto la *distancia predictiva* $DP(\mathcal{I})$, test para las observaciones anómalas es el ya dado en (4). Por tanto, como ya habíamos anunciado esta medida es débilmente robusta.

4.2. Medidas de influencia para el vector θ .

Puesto que esta medida se basa en la distribución a posteriori de $\theta | D_{(\mathcal{I})}$ y, según hemos visto en (7), ésta es independiente de la distribución de mezcla considerada, se tiene, por tanto, que la *distancia bayesiana*, $DB(\mathcal{I})$, coincide con la del caso del modelo de regresión usual (6).

5. APÉNDICE.

En este apartado damos algunos resultados generales que permiten calcular (con cierta facilidad) las distribuciones a posteriori de los parámetros de los modelos que se consideran en este trabajo.

Lema 1. Si \mathbf{X}_1 y \mathbf{X}_2 son vectores aleatorios de dimensión k_1 y k_2 , respectivamente y v es una variable aleatoria no negativa, tales que

$$\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2, v \sim N(\mathbf{A}\mathbf{x}_2 + \mathbf{b}; v\mathbf{S})$$

y

$$(\mathbf{X}_2, v) \sim NGaI(\mathbf{m}, \mathbf{V}; a, p)$$

entonces

$$\mathbf{X}_2, v | \mathbf{X}_1 = \mathbf{x}_1 \sim NGaI(\mathbf{m}', \mathbf{V}'; a', p')$$

donde

$$\mathbf{m}' = \mathbf{m} + \mathbf{V}\mathbf{A}'(\mathbf{S} + \mathbf{A}\mathbf{V}\mathbf{A}')^{-1}(\mathbf{x}_1 - \mathbf{A}\mathbf{m} - \mathbf{b})$$

$$\mathbf{V}' = \mathbf{V} - \mathbf{V}\mathbf{A}'(\mathbf{S} + \mathbf{A}\mathbf{V}\mathbf{A}')^{-1}\mathbf{A}\mathbf{V}$$

$$a' = a + \frac{1}{2}(\mathbf{x}_1 - \mathbf{A}\mathbf{m} - \mathbf{b})'(\mathbf{S} + \mathbf{A}\mathbf{V}\mathbf{A}')^{-1}(\mathbf{x}_1 - \mathbf{A}\mathbf{m} - \mathbf{b})$$

$$p' = p + \frac{k_1}{2}$$

Aplicación al modelo de regresión

Si $\mathbf{y}_{(T)} | \theta, \sigma^2 \sim N(\mathbf{X}_{(T)}\theta, \sigma^2 I_{n-m})$ y si suponemos que a priori la distribución sobre (θ, σ^2) es $NGaI(\mathbf{m}_0, \mathbf{C}_0; a_0, p_0)$ entonces, aplicando el lema 1 y haciendo tender $C_0 \rightarrow 0$, $a_0 \rightarrow 0$, $p_0 \rightarrow -k/2$ (equivalente a considerar la a priori no informativa), con $n - k - m > 0$, se tiene que

$$\theta, \sigma^2 | \mathbf{D}_{(T)} \sim NGaI\left(\hat{\theta}_{(T)}, (\mathbf{X}'_{(T)}\mathbf{X}_{(T)})^{-1}, \frac{n-k-m}{2} \hat{\sigma}^2_{(T)}, \frac{n-k-m}{2}\right).$$

Lema 2. Si \mathbf{X}_1 y \mathbf{X}_2 son vectores aleatorios de dimensión k_1 y k_2 , respectivamente y v es una variable aleatoria no negativa, tales que

$$\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2, v \sim \int N(\mathbf{A}(\boldsymbol{\lambda})\mathbf{x}_2 + \mathbf{b}(\boldsymbol{\lambda}); v\mathbf{S}(\boldsymbol{\lambda}))dF(\boldsymbol{\lambda})$$

y

$$(\mathbf{X}_2, v) \sim NGaI(\mathbf{m}, \mathbf{V}; a, p)$$

entonces

$$\mathbf{X}_2, v | \mathbf{X}_1 = \mathbf{x}_1 \sim \int NGaI(\mathbf{m}(\mathbf{x}_1, \boldsymbol{\lambda}), \mathbf{V}(\boldsymbol{\lambda}); a(\mathbf{x}_1, \boldsymbol{\lambda}), p')dF(\boldsymbol{\lambda} | \mathbf{x}_1)$$

donde

$$m(\mathbf{x}_1, \lambda) = \mathbf{m} + \mathbf{V}\mathbf{A}'(\lambda)(\mathbf{S}(\lambda) + \mathbf{A}(\lambda)\mathbf{V}\mathbf{A}'(\lambda))^{-1}(\mathbf{x}_1 - \mathbf{A}(\lambda)\mathbf{m} - \mathbf{b}(\lambda))$$

$$\mathbf{V}(\lambda) = \mathbf{V} - \mathbf{V}\mathbf{A}'(\lambda)(\mathbf{S}(\lambda) + \mathbf{A}(\lambda)\mathbf{V}\mathbf{A}'(\lambda))^{-1}\mathbf{A}(\lambda)\mathbf{V}$$

$$a(\mathbf{x}_1, \lambda) = a + \frac{1}{2}(\mathbf{x}_1 - \mathbf{A}(\lambda)\mathbf{m} - \mathbf{b}(\lambda))'(\mathbf{S}(\lambda) + \mathbf{A}(\lambda)\mathbf{V}\mathbf{A}'(\lambda))^{-1}(\mathbf{x}_1 - \mathbf{A}(\lambda)\mathbf{m} - \mathbf{b}(\lambda))$$

$$p' = p + \frac{k_1}{2}$$

y

$$dF(\lambda | \mathbf{x}_1) \propto t_{k_1}(\mathbf{x}_1 | \mathbf{A}(\lambda)\mathbf{m} + \mathbf{b}(\lambda); \frac{a}{p}(\mathbf{S}(\lambda) + \mathbf{A}(\lambda)\mathbf{V}\mathbf{A}'(\lambda))^{-1}, 2p) dF(\lambda)$$

6. REFERENCIAS.

- BARNETT, V. and LEWIS, T. (1984). *Outliers in Statistical Data*. New York: Wiley
- GIRÓN, F. J., MARTÍNEZ, M. L. y MORCILLO, C. (1992). A Bayesian Justification to the Analysis of Residuals and Influence Measures. *Bayesian Statistics 4*, 651–660. Oxford University Press: Oxford.
- KINGMAN, J. F. C. (1972). On random sequences with spherical symmetry. *Biometrika*, **59**, 492–494.
- SMITH, A. F. M. (1981). On random sequences with centred spherical symmetry. *Jour. Roy. Stat. Soc. B*, **43**, 208–209.
- WELSCH, R.E. (1982). Influence functions and regression diagnostics. *Modern Data Analysis* (R. L. Launer and A. F. Siegel, eds.). New York: Academic Press, 149–169.

ESTIMACION MEDIANTE SIMULACIONES DE LA POTENCIA Y EL NIVEL
DE SIGNIFICANCIA DE 5 PRUEBAS DE COMPARACIONES MULTIPLES
Y DEL ANALISIS DE VARIANZA

Osvaldo Camacho Castillo*

Ramiro Camacho Castillo**

RESUMEN

Generalmente, el primer paso del análisis estadístico de los datos de un experimento, es realizar una prueba de F para determinar si existen o no diferencias significativas entre medias de tratamientos; después de esto es necesario un análisis a fondo para entender como es dicha respuesta.

En caso de que el análisis de varianza sea significativo, es decir, que la hipótesis nula de igualdad entre las unidades experimentales sea rechazada con un α fijado de antemano, las opciones con que cuenta un investigador son: 1. Conjuntos planeados de contrastes entre medias o combinaciones lineales de medias de tratamientos; 2. Ajuste de funciones de respuesta, utilizando regresión y 3. Procedimientos de comparaciones múltiples por pares de medias.

De estos procedimientos los usados con mayor frecuencia y por lo mismo frecuentemente mal aplicados, son los procedimientos de comparaciones múltiples.

El objetivo de este trabajo es estimar la potencia y el nivel de significancia del análisis de varianza y las pruebas de comparaciones múltiples de medias de Duncan, Tukey, DMS, Scheffè y SNK, mediante la técnica de simulación.

* Facultad de Ingeniería, Universidad de Guadalajara, Campus Tecnológico, Av. Revolución s/n, Guadalajara Jal.

** Facultad de Ingeniería, UNAM, División de Estudios de Postgrado, Cd. Universitaria, D.F.

REVISIÓN BIBLIOGRÁFICA

Después de rechazar la igualdad de la u.e. mediante la prueba de F queda el problema de decidir cuál o cuáles tratamientos ocasionan el rechazo de H_0 . Esta puede rechazarse porque uno de los efectos de los tratamientos es diferente a los demás o bien porque son varios los que difieren.

Numerosos estudios se han dedicado a tratar de contestar esta pregunta y hasta el momento no hay acuerdo completo entre los estadísticos sobre el mejor procedimiento.

Los primeros trabajos para el establecimiento de las pruebas de comparaciones múltiples se realizaron en la década de los veintes, Irwin (1927), realizó el primer intento utilizando la distribución de distancias entre la n -ésima y la $(n-1)$ -ésima estadística de orden, Student(1927), propuso que se utilizara la amplitud como criterio para el rechazo de observaciones en análisis rutinarios, Fisher (1935) propuso el uso de la prueba de F y pruebas de t individuales, ésta es conocida como la prueba de la Diferencia Mínima Significativa, señalándose una probabilidad alta de inflación del error.

Newman (1939) a sugerencia de Student, formuló la primer prueba de comparaciones múltiples para el problema del análisis de varianza basada en las tablas de rangos estandarizados y elaboró la tabla para $\alpha = 0.05$ y $\alpha = 0.01$. Estos trabajos constituyeron la base del estudio de las comparaciones múltiples, sin embargo, fue hasta finales de los cuarentas y principios de la década de los cincuentas cuando Duncan, Scheffé y Tukey establecieron los fundamentos y desarrollaron las pruebas que llevan sus nombres.

MATERIALES Y MÉTODOS

Se realizaron simulaciones de experimentos utilizando el diseño experimental conocido como "completamente aleatorizado", considerando los siguientes factores y niveles:

- I Número de tratamientos (3,6,9)
- II Número de repeticiones (4,7,10)
- III Valores de medias (μ , $\mu + 0.5\sigma$, $\mu + \sigma$, $\mu + 2\sigma$, $\mu + 2\sigma$, $\mu + 3\sigma$).

Para cada experimento simulado, se generó la matriz de resultados $Y = \{y_{ij}\}$ donde $y_{i,j}$ es el valor observado de la respuesta en el

i-ésimo tratamiento en la repetición j, con las siguientes características:

$$Y_{i,j} \sim N(\mu + i\tau, \sigma^2) \quad \text{para } j = 1, 2, \dots, r$$

con t tratamientos, r repeticiones y j = 1, 2, ..., r

τ	$E(y_{1j})$	$E(y_{2j})$	$E(y_{tj})$
0 σ	$\mu + 0$	$\mu + 0$	$\dots, \mu + 0t$
0.5(10)	$\mu + 5$	$\mu + 10$	$\dots, \mu + 5t$
1(10)	$\mu + 10$	$\mu + 20$	$\dots, \mu + 10t$
2(10)	$\mu + 20$	$\mu + 40$	$\dots, \mu + 20t$
3(10)	$\mu + 30$	$\mu + 60$	$\dots, \mu + 30t$

Se generaron 10,000 experimentos para cada una de las 45 corridas (combinaciones de los niveles de los factores), dando un total de 450,000 experimentos a los cuales se les aplicó la metodología del Análisis de Varianza y de las pruebas de comparaciones múltiple consideradas.

Se registró el número de rechazos por cada $\mu + i\tau$ en los 10,000 experimentos de cada corrida, con lo cual se obtuvieron estimaciones del nivel de significancia y la potencia de las pruebas en las condiciones evaluadas.

Para la realización del trabajo se requirió diseñar e implementar un programa computacional para lo cual se eligió el lenguaje "PASCAL" y para compilarlo se uso la versión 6.0 de TURBO PASCAL.

Se ejecuta 10,000 veces una subrutina, dentro de la cual se simula un experimento cada vez, se lleva a cabo el ANVA y las 5 pruebas de comparaciones múltiples.

Todo lo anterior se lleva a cabo 45 veces una por cada corrida, lo que nos lleva a crear 45 archivos de resultados.

RESULTADOS Y CONCLUSIONES

Para la estimación del nivel de significancia se utilizó la proporción de experimentos en los que hubo al menos un rechazo de H_0 aplicando la prueba considerada, dado que no existía ningún efecto de los tratamientos.

$$\hat{\alpha} = \frac{n}{N}$$

donde n es el número de rechazos en N experimentos simulados con igualdad en los efectos de los tratamientos.

Los resultados se muestran en las gráficas 1 a 6, de donde se puede concluir que:

La prueba de la Diferencia Mínima Significativa (DMS) y la de Duncan presentan una alta probabilidad de cometer el error tipo I, ya que considerando experimentos con tres tratamientos para un α fijado de 0.01, el $\hat{\alpha}$ es para la DMS de 0.026 y de 0.02 para la Duncan, incrementándose substancialmente al considerar experimentos con mayor número de tratamientos, al grado de que para un $\alpha = 0.05$, con experimentos de 9 tratamientos, la probabilidad de cometer el error tipo I es mayor al 0.5.

Las pruebas de Tukey, SNK y el ANVA, presentan muy pocas diferencias entre $\hat{\alpha}$ y α , mientras que la prueba de Scheffé resultó sobre-protectora contra el riesgo de cometer error tipo I ya que $\hat{\alpha}$, siempre fue menor que el α fijado siendo ésto más marcado en los experimentos con muchos tratamientos.

En general podemos decir que las pruebas de DMS y Duncan tienen alta probabilidad de presentar diferencias significativas inexistentes entre efectos de tratamientos, contrariamente la prueba de Scheffé se protege en exceso de cometer este error.

La forma más confiable de tener un nivel de significancia fijado de antemano es usar las pruebas de Tukey o la de SNK (estos resultados son válidos para todos los niveles de los factores estudiados con $\alpha = 0.01$ y $\alpha = 0.05$).

Los resultados de la simulaciones se agruparon según la diferencia entre las medias teóricas que se evaluaron, obteniéndose las funciones de potencia de cada prueba para el número de tratamientos y repeticiones considerados, se calcularon intervalos de confianza para las estimaciones de la potencia, mediante el método de la aproximación de la distribución binomial a la normal, estos intervalos debido al tamaño de muestra ($>10,000$) son muy estrechos (del orden de 10^{-4}) por lo se optó por utilizar la estimación puntual.

Se presentan 18 gráficas (de la 7 a la 25) de las funciones de potencia. Dichas gráficas en el eje de las abscisas muestran los valores de las diferencias de medias (p) en los experimentos simulados, mientras que en el eje de las ordenadas, se observa la probabilidad estimada $\delta(p)$ de que se rechace la hipótesis nula.

De las funciones de potencia se pueden extraer algunas conclusiones generales las cuales son válidas tanto para $\alpha=0.05$ como para $\alpha=0.05$, así pues, tenemos que:

I. En lo que se refiere al nivel de significancia se encontró que, en las pruebas DMS y Duncan, un número grande de tratamientos provoca un nivel de significancia real alto, lo cual está de acuerdo con Vargas(1982). Como ejemplo de esto, mencionaremos que para 9 tratamientos, 10 rep. y $\alpha=0.05$ se encontró que la DMS trabaja con un nivel de significancia real de 0.52 mientras que Duncan lo hace con 0.33.

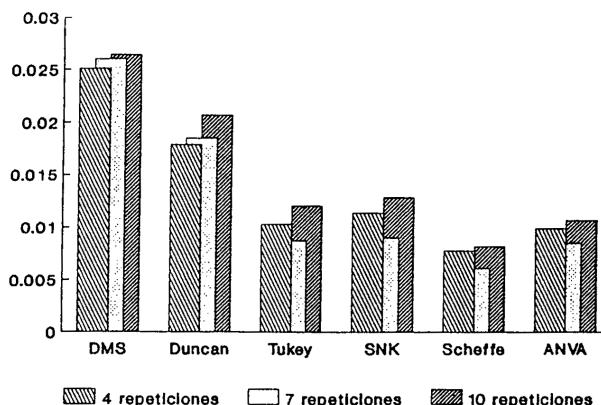
II. La prueba de Scheffé resultó la más conservadora, ya que, si bien , para tres tratamientos los valores estimados para los niveles de significancia son cercanos a los teóricos, para más tratamientos el nivel de significancia real es menor, a manera de ejemplo, para 9 tratamientos 10 repeticiones y $\alpha=0.05$. se encontró que $\hat{\alpha}=0.003$, esto significa que Scheffé sólo rechaza una diferencia cuando ésta es muy grande. Lo que concuerda con Vargas(1982).

III. El análisis de varianza y las pruebas SNK y Tukey tuvieron un valor estimado bastante cercano al valor teórico en todos los casos.

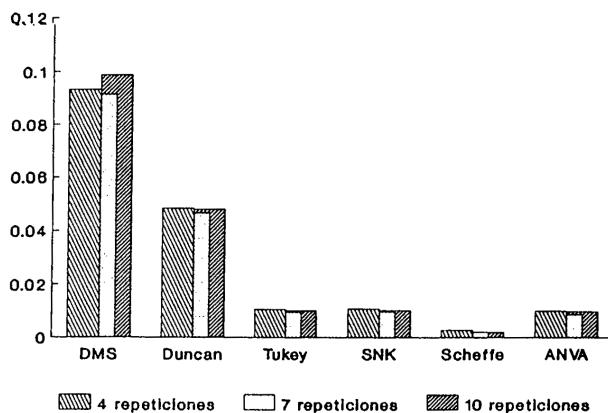
IV. En lo referente a la potencia de las pruebas, un número alto de repeticiones, como era de esperarse, trae como consecuencia un aumento en la potencia, lo cual es válido para todas las pruebas.

V. Se encontró también que, al igual que con el número de repeticiones, un incremento en el número de tratamientos redonda en un aumento de la potencia de las pruebas, aunque esto no sucede con la misma intensidad para cada una de ellas así pues, mientras DMS y Duncan, experimentan un fuerte aumento en la potencia Scheffé apenas si es afectada por este factor. Las pruebas restantes (Tukey y SNK) se mantienen en niveles intermedios.

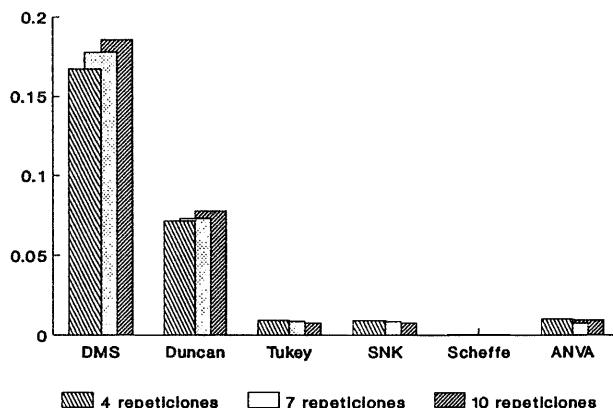
VI. En lo particular, la prueba de Scheffé resultó ser la menos potente de todas, ya que sólo rechazó para diferencias muy grandes.



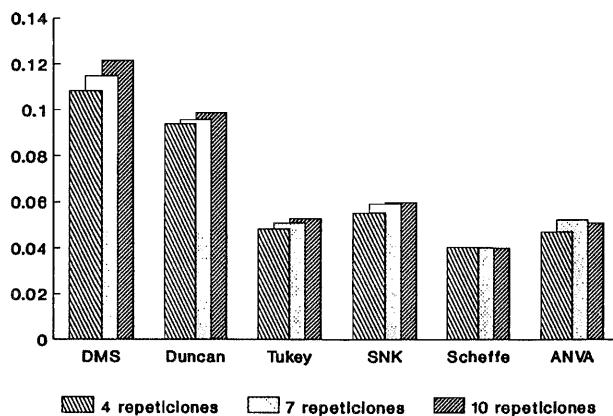
Graf. 1. Nivel de significancia ($\hat{\alpha}$) de las pruebas de comparaciones múltiples, y el análisis de varianza con experimentos simulados de 3 tratamientos de efectos similares, con $\alpha = 0.01$.



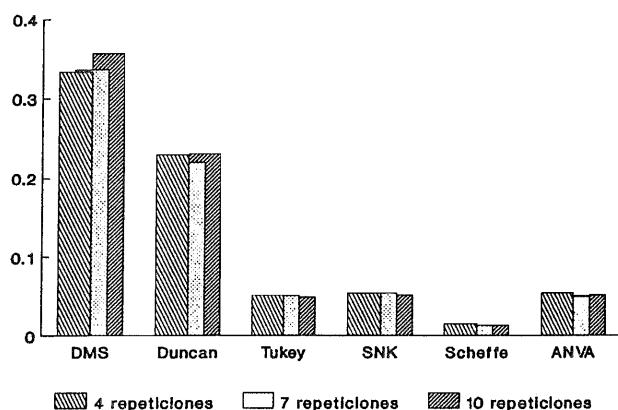
Graf. 2. Nivel de significancia ($\hat{\alpha}$) de las pruebas de comparaciones múltiples, y el análisis de varianza con experimentos simulados de 6 tratamientos de efectos similares, con $\alpha = 0.01$.



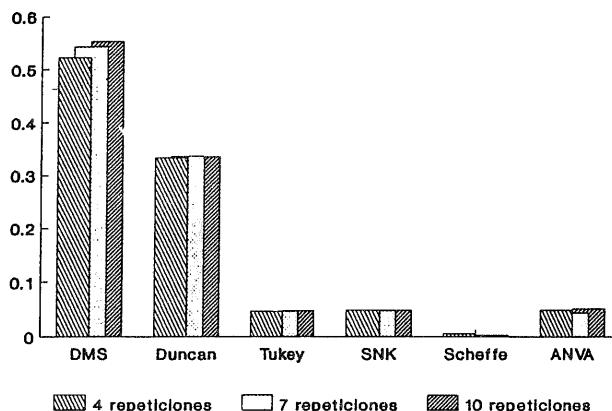
Graf. 3. Nivel de significancia ($\hat{\alpha}$) de las pruebas de comparaciones múltiples, y el análisis de varianza con experimentos simulados de 9 tratamientos de efectos similares, con $\alpha = 0.01$.



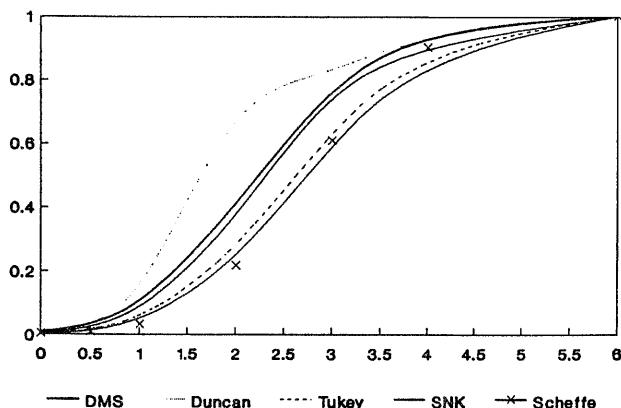
Graf. 4. Nivel de significancia ($\hat{\alpha}$) de las pruebas de comparaciones múltiples, y el análisis de varianza con experimentos simulados de 3 tratamientos de efectos similares, con $\alpha = 0.05$.



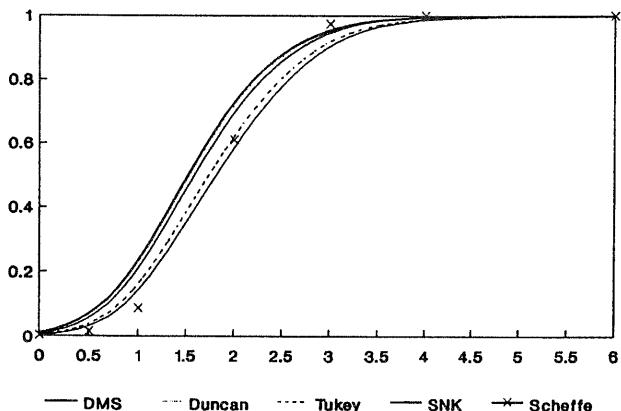
Graf. 5. Nivel de significancia ($\hat{\alpha}$) de las pruebas de comparaciones múltiples, y el análisis de varianza con experimentos simulados de 6 tratamientos de efectos similares, con $\alpha = 0.05$.



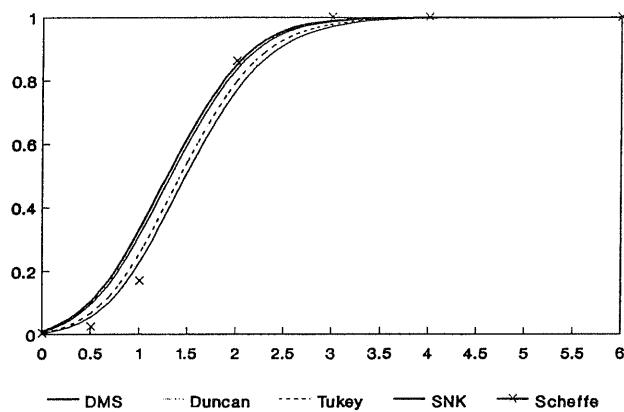
Graf. 6. Nivel de significancia ($\hat{\alpha}$) de las pruebas de comparaciones múltiples, y el análisis de varianza con experimentos simulados de 9 tratamientos de efectos similares, con $\alpha = 0.05$.



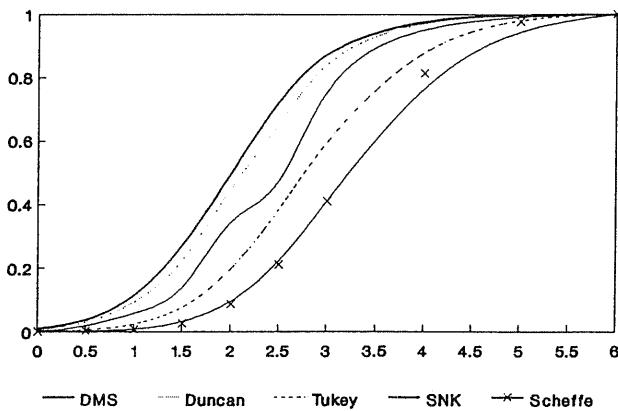
Graf. 7. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 3 tratamientos y 4 repeticiones, $\alpha=0.01$.



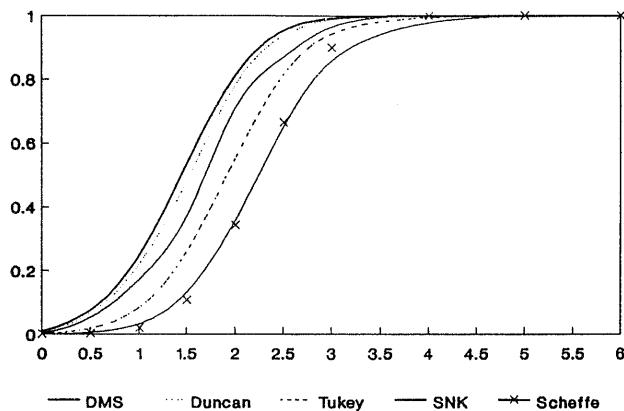
Graf. 8. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 3 tratamientos y 7 repeticiones, $\alpha=0.01$.



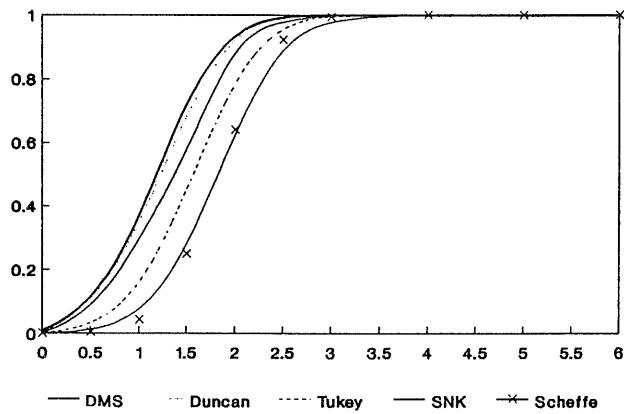
Graf. 9. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 3 tratamientos y 10 repeticiones, $\alpha=0.01$.



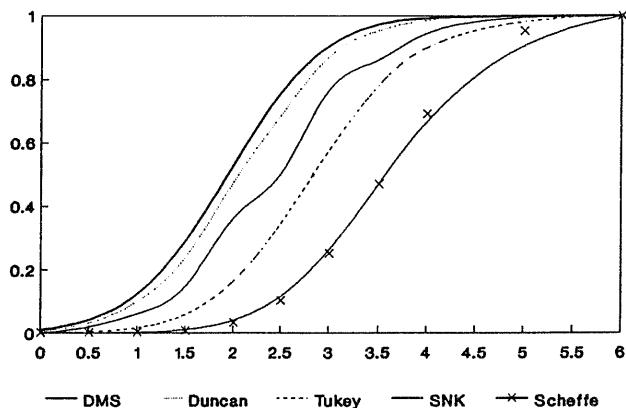
Graf. 10. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 6 tratamientos y 4 repeticiones, $\alpha=0.01$.



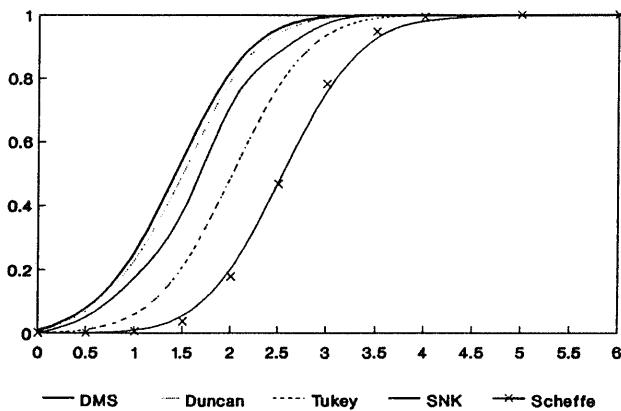
Graf. 11. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 6 tratamientos y 7 repeticiones, $\alpha=0.01$.



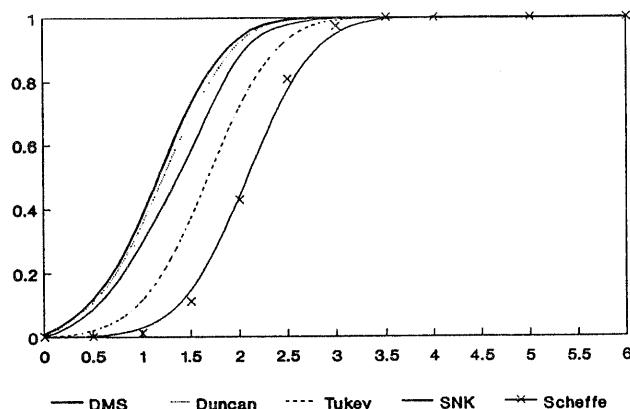
Graf. 12. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 6 tratamientos y 10 repeticiones, $\alpha=0.01$.



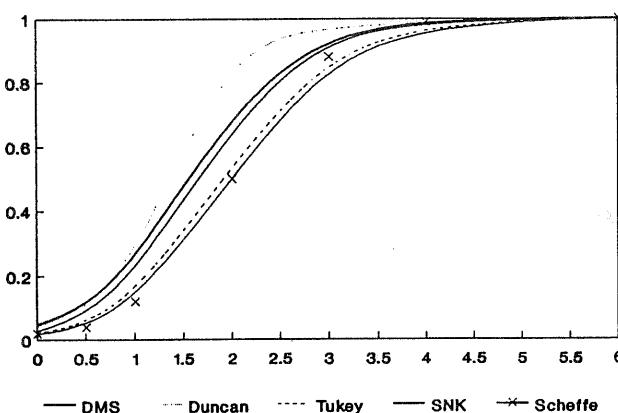
Graf. 13. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 9 tratamientos y 4 repeticiones, $\alpha=0.01$.



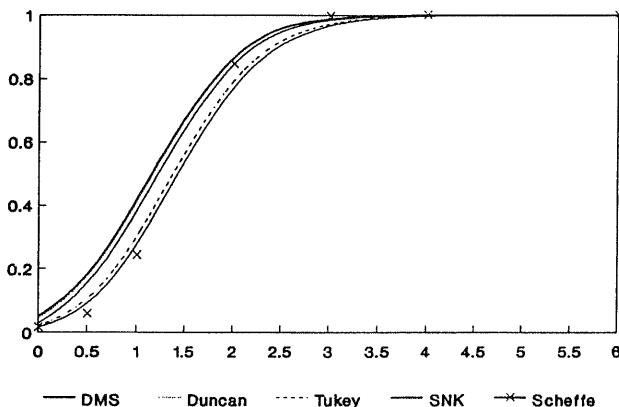
Graf. 14. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 9 tratamientos y 7 repeticiones, $\alpha=0.01$.



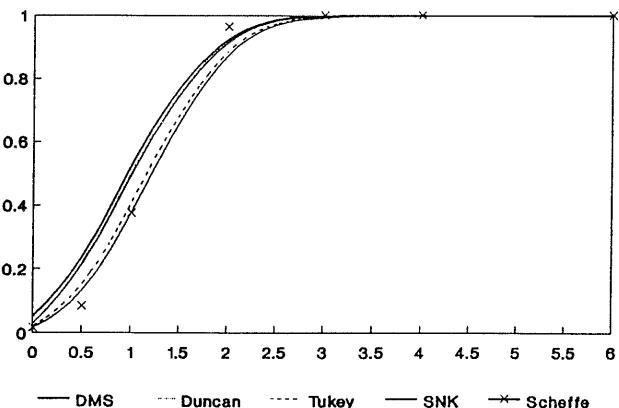
Graf. 15. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 9 tratamientos y 10 repeticiones, $\alpha=0.01$.



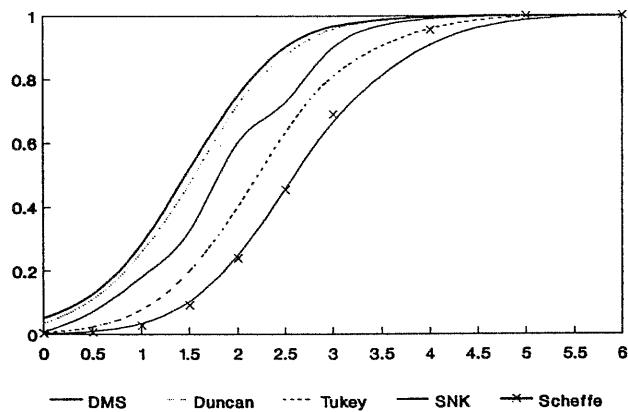
Graf. 16. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 3 tratamientos y 4 repeticiones, $\alpha=0.05$.



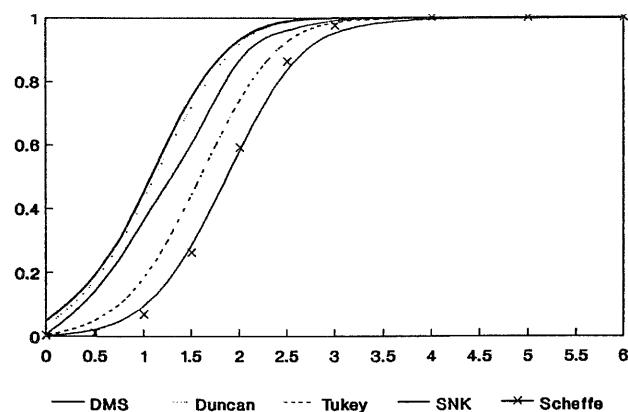
Graf. 17. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 3 tratamientos y 7 repeticiones, $\alpha=0.05$.



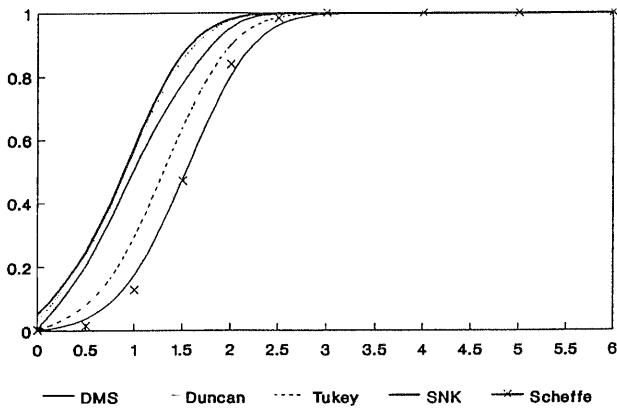
Graf. 18. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 3 tratamientos y 10 repeticiones, $\alpha=0.05$.



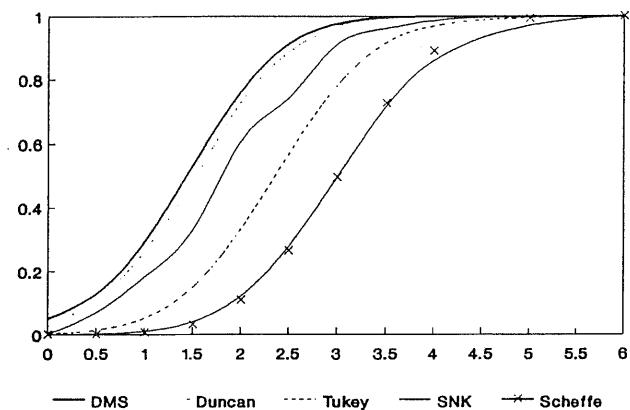
Graf. 19. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 6 tratamientos y 4 repeticiones, $\alpha=0.05$.



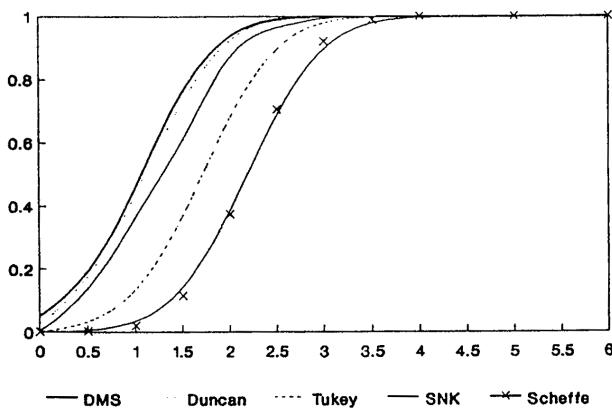
Graf. 20. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 6 tratamientos y 7 repeticiones, $\alpha=0.05$.



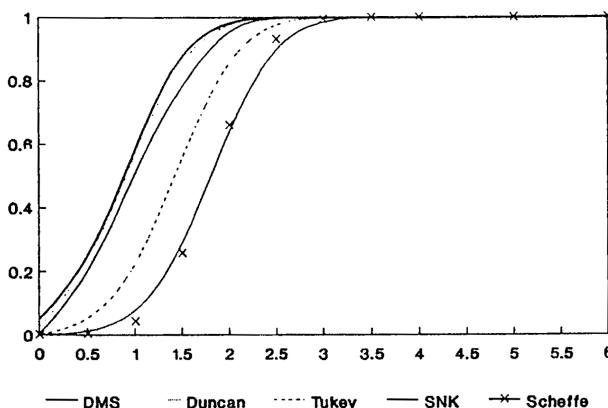
Graf. 21. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 6 tratamientos y 10 repeticiones, $\alpha=0.05$.



Graf. 22. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 9 tratamientos y 4 repeticiones, $\alpha=0.05$.



Graf. 23. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 9 tratamientos y 7 repeticiones, $\alpha=0.05$.



Graf. 24. Función potencia de las pruebas de comparaciones múltiples para diferentes valores del parámetro p (diferencias de las medias de la respuesta en desviaciones estandard en los experimentos simulados), con 9 tratamientos y 10 repeticiones, $\alpha=0.05$.

BIBLIOGRAFIA

- Burden, R. L. & J. D. Faires (1985), Análisis Númerico, Grupo Editorial Iberoamérica, México.
- Fisher, R.A. 1974, The Design of Experiments, Hafner Press, N.Y.
- Gerald. C. F. (1987) Análisis Numérico, Representaciones y Servicios de Ingenieria S. A., México.
- Infante G., S y G.P. Zarate de L. (1984) Métodos Estadísticos para Investigadores, Ed. Trillas, México.
- John, P. W.M. (1971) Statistical Design and Analysis of Experiments, The Macmillan Company, New York.
- Kennedy, W.J. & J. E. Gentle (1980) Statistical Computing, Marcel Dekker, Inc., New York.
- Martínez G., A. (1988) Diseños Experimentales, Ed. Trillas, México.
- Mendenhall, W.; R.L. Scheaffer y D.D. Wackerly, Estadística Matemática con Aplicaciones, Grupo Editorial Iberoamérica. México.
- Patel, J. K.; C.H. Kapadia & D.B. Owen, Handbook of Statistical Distributions, Marcel Dekker, Inc. New York.
- Pérez V., O. (1990) Algunas consideraciones para la selección de un procedimiento de comparaciones múltiples de medias, Tesis de Maestría en Ciencias, Centro de Estadística y Cálculo, Colegio de Postgraduados, Montecillos, México.
- Rohatgi, V. K. (1976) An Introduction to Probability Theory and Mathematical Statistical, Jhon Wiley and Sons, N.Y.
- Shanon, R.E. (1976) Systems Simulation the Art and Science, Prentice Hall, Inc. Englewood Cliffs, New Jersey.
- Spiegel, M. R. (1970) Estadística, Mc Graw Hill, México.
- Steel, R.G.D. & J.H. Torrie (1980), Bioestadística Principios y Procedimientos, Ed. McGraw Hill. Bogotá, Colombia.
- Vargas Ch., D. (1982), El estado actual de las pruebas de comparaciones múltiples, Tesis Profesional, Facultad de Ciencias, Universidad Nacional Autónoma de México. México.

INDEPENDENCIA CONDICIONAL Y DIAGRAMAS DE
INFLUENCIA PROBABILISTA; DOS APLICACIONES.

M. MARTINEZ MORALES, J.A. MONTANO RIVAS, M.M. OJEDA RAMIREZ¹

RESUMEN

En este trabajo se presentan dos ejemplos de la aplicación de los diagramas de influencia probabilista a problemas estadísticos. Estos diagramas son particularmente útiles en estudios que involucran interrelaciones entre variables "dependientes" e "independientes"¹ auxiliando al estadístico en la formulación de hipótesis de independencia y facilitando la interpretación de los resultados.

INTRODUCCION

Los diagramas de influencia probabilista (Barlow y Braganca, 1990) tienen sus antecedentes en los diagramas con nodos de decisión propuestos por Miller (Citado en Howard y Matheson, 1984). Posteriormente Shachter (1986) desarrolló otro método para el análisis de diagrámas de influencia. Estos diagrámas, según nuestro punto de vista, no sustituyen a los métodos estadísticos formales empleados en el análisis de sistemas o procesos en los que existen interrelaciones entre varias variables; sin embargo,

¹ LABORATORIO DE INVESTIGACION Y ASESORIA ESTADISTICA, UNIVERSIDAD VERACRUZANA; AV. XALAPA CAMACHO. XALAPA, VERACRUZ. CP. 91000. (LINAE), ESQ. FAC. AVILA

constituyen un valioso auxiliar en la interpretación de los resultados, como se muestra a través de un par de ejemplos.

En términos simples, se define un diagrama de influencia probabilista como una gráfica acíclica dirigida en la cual:

- (i) los nodos representan variables aleatorias y los arcos dirigidos indican una posible dependencia entre ellos.
- (ii) a cada nodo se asocia una función de probabilidad condicional que depende de los valores de las variables en los nodos antecesores adyacentes.

Dada una gráfica acíclica dirigida con las probabilidades condicionales asociados a los nodos, existe una función única de probabilidad conjunta correspondiente al conjunto de variables aleatorias representadas en la gráfica. También se puede demostrar que la ausencia de un arco que conecte a dos nodos en el diagrama indica que las variables asociadas a estos nodos son condicionalmente independientes dados los estados de todos los otros nodos antecesores adyacentes.

INDEPENDENCIA CONDICIONAL

A. P. Dawid (1979) ha sostenido que el concepto de independencia condicional es de suma importancia para la teoría estadística. En particular él ha mostrado que conceptos como los de suficiencia y ancilaridad, esquemas de decisión y de muestreo

secuencial pueden derivarse del concepto básico de independencia condicional. En seguida presentamos de manera un tanto informal la definición de independencia condicional y su aplicación a problemas concretos de análisis estadístico.

Definición. Sean X , Y , Z variables aleatorias. Decimos que Y es condicionalmente independiente de X dado Z , si la distribución condicional de Y dados (X, Z) es igual a la distribución condicional de Y dado Z . (Se denota $Y \perp\!\!\!\perp X|Z$).

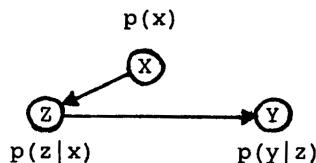
La interpretación intuitiva de este concepto, sugerida por Barlow y Bragance, es que si Z esta dada, entonces X no proporciona información adicional alguna sobre Y .

Dawid (1979) ofrece expresiones para la independencia condicional equivalentes entre sí, en términos de las funciones de densidad correspondientes. Empleando la notación obvia, tenemos que Y es condicionalmente independiente de X dado Z si se cumplen:

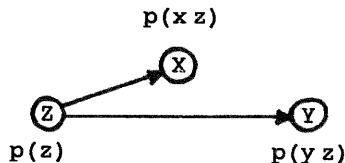
$$(i) p(y,x|z) = p(x|z)p(y|z); \text{ o equivalentemente si:}$$

$$(ii) p(y|x,z) = p(y|z).$$

Si Y es condicionalmente independiente de X
(equivalentemente: si X es condicionalmente independiente de Y)
el diagrama de influencia correspondiente es:



o, equivalentemente:



Puede verificarse fácilmente que, dada la independencia condicional de X y Y dado Z, ambos diagramas conducen a la misma distribución conjunta de Y, X, Z. Lo importante es que no exista un arco conectando los nodos representantes de Y y X.

En numerosos estudios estadísticos están presentes varias variables, existiendo interrelaciones más o menos complejas entre ellas. En muchos casos, es posible distinguir conjuntos de variables "dependientes" y variables "independientes"; es decir, se presume una asociación (o un "efecto") asimétrica entre las variables. Es en esta situación que es posible representar esas interrelaciones por medio de diagramas de influencia como los descritos, emplearlos para formular claramente las hipótesis del investigador y, posteriormente, apoyar la interpretación de los resultados del análisis.

A continuación se describen dos ejemplos relativamente sencillos donde se ilustra la aplicación de estos diagramas.

EJEMPLO I

Se realizó un experimento en los viveros de Xalapa para determinar cuál combinación de sustratos y fórmulas de fertilización producía el mayor porcentaje de crisantemos que

cumplieran con la norma de calidad exigida por ciertos compradores. La variable asociada con la calidad era el diámetro de la flor; considerándose de calidad aceptable si $d \geq 15$ cms. y no aceptable en caso contrario. Sea Y la variable definida por :

$$Y = \begin{cases} 1 & \text{si } d \geq 15 \\ 0 & \text{si } d < 15 \end{cases}$$

Se eligieron cuatro distintos tipos de sustratos, ($T=1, 2, 3, 4$) y dos fórmulas de fertilización ($F=1, 2$).

Se desea saber qué efectos tienen las variables T y F sobre Y . En términos de diagramas de influencia probabilística se tienen varias posibilidades.

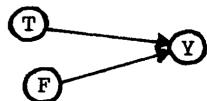
(1)



$$\begin{matrix} Y \perp\!\!\! \perp T \\ Y \perp\!\!\! \perp F \end{matrix}$$

Las variables T y F no tiene efecto alguno sobre la variable Y (Y es independiente de T y F).

(2)



$$Y \not\perp\!\!\! \perp T | F$$

$$Y \not\perp\!\!\! \perp F | T$$

T y F afectan ambas a Y independientemente (no existe interacción).

(3)



$$Y \perp\!\!\!\perp F | T$$

(a)

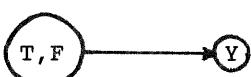


$$Y \perp\!\!\!\perp T | F$$

(b)

Y es condicionalmente independiente de F(o T) dado T(o F).

(4)



$$Y \not\perp\!\!\!\perp T | F$$

$$Y \not\perp\!\!\!\perp F | T$$

$$(T \not\perp\!\!\!\perp F)$$

T y F afectan conjuntamente a Y (existe interacción, no se pueden separar los efectos de T y F).

En el estudio original de este problema (Montano, A., Ojeda, M. Reporte Interno, LINAE) se empleó el análisis de varianza convencional para determinar cuál era la combinación sustrato-fertilizante que daba el mejor rendimiento, tomando como variable de respuesta al diámetro y largo del tallo medidos en una escala de intervalo. De este estudio se concluyó que el mejor tratamiento era la combinación F = 1 y T = 1.

Discretizado el problema mediante el empleo de la variable Y realizamos un análisis cualitativo basado en las distribuciones

condicionales estimadas de $Y|T,F$; $Y,F|T$ y $Y,T|F$; ya que comparando estas distribuciones podrá inferirse si Y es condicionalmente independiente de T dado F o viceversa.

A continuación se muestran los resultados obtenidos en el experimento.

$F = 1$

Y	T				TOTAL
	1	2	3	4	
0	1	3	2	1	7
1	3	1	2	3	9
TOTAL	4	4	4	4	16

$F = 2$

Y	T				TOTAL
	1	2	3	4	
0	3	3	3	3	12
1	1	1	1	1	4
TOTAL	4	4	4	4	16

TABLA I. Presentación de los cuatro diferentes sustratos (T) con los dos tipos de fertilización (F) para los dos niveles de calidad (Y) aceptables en el mercado.

Puede observarse que para $F=2$, el efecto de T es nulo; esto es, se observa la misma respuesta para $T = 1, 2, 3, 4$, dado $F=2$. Cuando $F=1$, T sí parece tener un efecto diferenciador. Los datos sugieren probar la hipótesis:

$$H_0: Y \perp\!\!\! \perp T|F \quad \text{vs} \quad H_1: Y \not\perp\!\!\! \perp T|F$$

Kullback (1968) proporciona una prueba para hipótesis de independencia condicional como ésta para tablas de contingencia. El estadístico de prueba tiene una distribución (aproximadamente) Chi-cuadrada. Para nuestro ejemplo el valor del estadístico fue de $\chi^2 = 2.89$ con 6 grados de libertad. El resultado es no significativo al nivel =0.05, lo que nos indica no rechazar H_0 , es decir la hipótesis de independencia condicional de Y y T dado F. Esto es que lo determinante es la fórmula de fertilización. En términos gráficos H_0 se representa por el diagrama 3(b).

EJEMPLO II

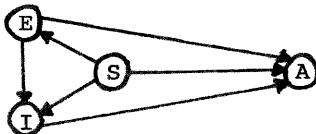
En una encuesta de opinión levantada en mayo de 1991 en la ciudad de Xalapa, interesaba, entre otras cosas, ver el efecto de algunas variables "independientes" sobre la opinión que se tenía sobre el aborto. Uno de los autores aplicó en aquella ocasión un modelo lineal generalizado para analizar los datos (Ojeda M.M., Modelos Lineales Generalizados Para Analizar Datos en Grupos, Reporte Interno, LINAE). A continuación se describe la reformulación y análisis del problema empleado diagramas de influencia probabilista y el concepto de independencia condicional.

Sean A, E, I, S las variables que representan la opinión sobre el aborto, escolaridad, ingreso mensual y sexo

respectivamente. (Por razones de espacio no se reproduce la tabla de frecuencias ($2 \times 6 \times 4 \times 2$) de este estudio; están a la disposición de quien se interese).

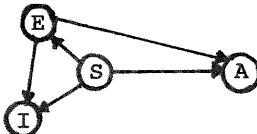
Ahora bien, el punto de partida, la hipótesis inicial es que las variables "independientes" (E, S, I) afectan a la variable A y existen también interrelaciones entre ellas.

El diagrama (hipotético) inicial sería pues el siguiente:



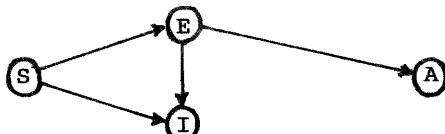
Aplicando repetidamente la prueba de Kullbak (v. g. para determinar si $A \perp\!\!\!\perp I | (E, S)$) se fue reduciendo al diagrama de la siguiente manera:

(1) $A \perp\!\!\!\perp I | E, S$



Esto es, desaparece el arco uniendo a los nodos I con A. (El ingreso no "afecta" la opinión sobre el aborto dadas la escolaridad y el sexo).

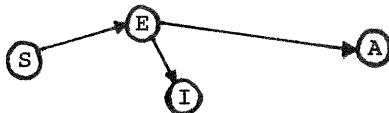
(2) $A \perp\!\!\!\perp I, S | E$



Dada la escolaridad, el sexo no "afecta" la opinión sobre el aborto. Es decir, dentro de cada nivel de escolaridad no hay diferencia entre las respuestas de los sexos. Adicionalmente se probó que I es condicionalmente independiente de S dado E:

(3)

$I \perp\!\!\!\perp S | E$



La conclusión, clara en el diagrama, es que A es condicionalmente independiente de S e I, dado E. En otras palabras, si conocemos la escolaridad, el sexo y el ingreso no agregan información acerca de la opinión sobre el aborto. En concreto, la gente con mayor escolaridad tiende a dar una opinión negativa sobre permitir el aborto deseado.

Estos ejemplos, relativamente sencillos, ilustran una aplicación práctica del concepto de independencia condicional y los respectivos diagramas de influencia probabilística. Consideramos que su empleo es de gran utilidad en cierto tipo de estudios estadísticos. Aunque aquí los empleamos en variables discretas, su uso puede extenderse a variables continuas usando las correlaciones parciales como indicadores de dependencia condicional. En el LINAE se llevan a cabo actualmente varios estudios que ameritan el empleo de esta técnica.

REFERENCIAS

Barlow, R.E. and Braganca-Pereira, C.A. (1990). Conditional independence and probabilistic influence diagrams. En Topics in Statistical Dependence. Lecture Notes-Monograph Series. Institute of Mathematical Statistics.

Dawid, A.P. (1979). Conditional independence in statistical theory. J.R. Statist. Soc. B 41, 1-31.

Howard, R.A. and Matheson, J.E. (1984). Influence diagrams. En Readings in the Principles and Applications of Decision Analysis. Howard and Matheson, eds. Strategic Decision Group, Menlo Park, CA.

Kullback, S, (1968) Information Theory and Statistics. Dover Pub. Inc. New York.

Schachter, R. (1986). Evaluating influence diagrams. Oper. Res. 34, 871-882.

MODELOS DE COMPONENTES DE VARIANZA PARA INVESTIGACION EDUCATIVA.

M.M. OJEDA ¹
J.A. MONTANO

RESUMEN:

Se presenta la metodología de los modelos de componentes de varianza balanceados, en los casos univariado y multivariado, extendiéndolos a una formulación llamada multivariada doble. Se propone una prueba para verificar el efecto de grupo en el caso multivariado.

SUMMARY:

The components of variance models methodology in balanced univariate and multivariate cases are presented. A formulation named double multivariate is introduced. A test for checking effect of groups in the multivariate case is proposed.

KEYWORDS: Components of Variance Models, Double Multivariate Linear Model formulation, Educational Research.

¹ LABORATORIO DE INVESTIGACION Y ASESORIA ESTADISTICA, (LINAEE)
FAC. DE ESTADISTICA, UNIVERSIDAD VERACRUZANA. AV. XALAPA Y
AVILA CAMACHO, XALAPA VER.

1.- INTRODUCCION.

El investigador en el área educativa enfrenta con bastante frecuencia la necesidad de realizar un estudio analítico basado en modelar una respuesta Y (por ejemplo rendimiento escolar o calificación para un test), considerando como factor diferencial el grupo. La investigación se diseña considerando que estudiar la totalidad de grupos (N) es prohibitiva, por tiempo y por recursos requeridos, y por tanto k grupos son seleccionados aleatoriamente. En cada grupo seleccionado, a la vez se obtiene una muestra aleatoria de n individuos, porque no se requiere estudiar al grupo completo. Así la muestra queda constituida por $m = nk$ individuos. Sea el modelo a considerar el siguiente

$$Y_{ij} = \mu + a_i + e_{ij} \quad (1)$$

donde Y_{ij} es la respuesta del individuo j -ésimo ($j = 1, 2, \dots, n$) del grupo i -ésimo ($i = 1, 2, \dots, k$), y μ es una media general. Las hipótesis adicionales que se asocian a este modelo son las siguientes:

- i) $e_{ij} \sim NI(0, \sigma^2_e)$
 - ii) $a_i \sim NI(0, \sigma_a^2)$
 - iii) $Cov(a_i, e_{ij}) = 0$
- (2)

La primera se refiere a que los errores aleatorios e_{ij} , se distribuyen de acuerdo a una normal, independientes con media cero y varianza σ^2 ; los efectos de grupos a se postulan ser asimismo una variable aleatoria con media cero y varianza σ_a^2 , y distribuirse independientemente. El tercer supuesto plantea que la covarianza, y por tanto la correlación, entre los errores y los efectos de grupo, es nula. Bajo estas consideraciones, este modelo, llamado de componentes de varianza, nos permite estudiar la varianza de la respuesta Y , descomponiéndola en dos partes, σ^2 del error aleatorio y σ_a^2 de los efectos de grupo; es decir

$$\sigma_y^2 = \sigma_a^2 + \sigma^2 \quad (3)$$

El uso de estos modelos es bastante generalizado en el área de diseños experimentales, principalmente en el área de genética [Searle (1971)], pero recientemente se usa para investigaciones en el área educativa [Keesling (1976), Keesling y Wiley (1974)]. En este artículo presentamos una revisión de los resultados importantes, necesarios de considerar al usar este tipo de modelos y algunas extensiones, que se desarrollan en las secciones 3 y 4. Además se presentan algunas recomendaciones metodológicas para su uso.

2.- EL ANALISIS DE VARIANZA.

La tabla usual de análisis de varianza [Searle (1971)] contiene los estimadores de los componentes de la varianza, que resultan ser:

$$(a) \hat{\sigma}^2 = CM_{(DENTRO)} = \frac{SC_{(DENTRO)}}{k(n-1)} \quad (4)$$

$$(b) \hat{\sigma}_a^2 = \frac{1}{n} (CM_{(ENTRE)} - CM_{(DENTRO)})$$

Con $CM_{(ENTRE)} = [SC_{(ENTRE)} / (k-1)]$, y SC y CM representan la suma de cuadrados y cuadrado medio respectivamente. Puede ser mostrado fácilmente que $\hat{\sigma}_a^2$ y $\hat{\sigma}^2$ son estimadores insesgados de σ_a^2 y σ^2 respectivamente [ver Searle (1971)], pero un problema importante es que existe una probabilidad diferente de cero de obtener una estimación de la varianza del efecto de grupos negativa; es decir $P[\hat{\sigma}_a^2 < 0] \geq 0$. Una forma simple de resolver este problema es restringiendo este valor a valores positivos (restricted estimation).

La prueba de hipótesis que se plantea en este caso sería

$$H_0: \sigma_a^2 = 0 \quad vs \quad H_a: \sigma_a^2 > 0 \quad (5)$$

El rechazo de esta hipótesis nos llevaría a sostener que hay diferencias entre grupos. El procedimiento de prueba sería:

1) Calcular el estadístico

$$F_c = \frac{CM_{(ENTRE)}}{CM_{(DENTRO)}}$$

2) Considerando que $F_c \sim F_{(k-1, k(n-1))}$ bajo H_0 , entonces; rechazar H_0 : con un nivel de significancia α si $F_c > F_{(k-1, k(n-1))}^{(1-\alpha)}$ donde $F_{(a, b)}^{(1-\alpha)}$ es un valor de tablas de la distribución F con a grados de libertad en el numerador y b en el denominador.

Las consideraciones sobre la importancia de las suposiciones para este caso son análogas al caso del modelo con efectos fijos, pero esta prueba es bastante robusta para tamaños de muestra relativamente pequeños ($n = 8$) [Lindman (1974) pag. 119-120].

El enfoque multivariado para el mismo problema descrito en la introducción se plantea considerando que hay k muestras independientes n -variadas de una distribución normal, con vector de medias μl_n^t , donde b^t indica el transpuesto de b , y matriz de covarianzas

$$V = \sigma^2 I_n + \sigma_a^2 l_n^t l_n \quad (6)$$

En esta notación matricial I_n representa la matriz idéntica de orden n y l_n es un vector columna de n unos. La estructura de covariación, que se corresponde con los incisos ii) y iii) de las expresiones en (2), nos plantea que Y_{ij} tiene covarianza σ_a^2 dentro

de grupo, e igual a cero si los correspondientes individuos son de diferentes grupos; es decir:

$$\text{COV } (Y_{ij}, Y_{i'j'}) = \begin{cases} 0 & \text{si } i \neq i' \\ \sigma_\alpha^2 & \text{si } i=i' \text{ y } j \neq j' \\ \sigma_\alpha^2 + \sigma_e^2 & \text{si } i=i', j=j' \end{cases} \quad (7)$$

Este enfoque permite incorporar algunos resultados generales del Análisis de Varianza Multivariado [Mardia et. al (1979)] a este tipo de problemas. Nosotros sólo hemos presentado la formulación puesto que a través de ella se obtiene un avance conceptual para comprender la siguiente sección. El lector interesado puede revisar Tiao y Tan (1966) y Keesling (1976).

3.- EL PROBLEMA EN EL CASO MULTIVARIADO.

Consideremos el mismo problema de la introducción, pero asumamos que el investigador está interesado en estudiar conjuntamente varias respuestas; es decir, $\underline{Y}_{ij}^t = (Y_{ij}^{(1)}, \dots, Y_{ij}^{(p)})$. Los componentes del modelo (1) ahora se convierten en vectores de orden p y las varianzas se transforman en matrices de varianzas y covarianzas (Σ_a y Σ). Los supuestos se extienden al caso de normales p -variadas [Mardia et al (1979)], y entonces tenemos que

$$\sum_y = \sum_a + \sum \quad (8)$$

con lo que ahora estamos interesados en estudiar los componentes de varianzas y covarianzas. El modelo para el caso multivariado sería:

$$\tilde{Y}_{ij} = \mu + \tilde{a}_i + \tilde{e}_{ij} \quad (9)$$

Nuevamente el procedimiento usual del análisis de varianza multivariado nos provee los cálculos necesarios para probar la hipótesis de interés. Sean las matrices S_a y S definidas como sigue:

$$S_a = \left(\frac{n}{k} \right) \sum_{i=1}^k (\tilde{Y}_{i\cdot} - \tilde{Y}\dots)(\tilde{Y}_{i\cdot} - \tilde{Y}\dots)^t \quad (10)$$

$$S = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^n (\tilde{Y}_{ij} - \tilde{Y}_{i\cdot})(\tilde{Y}_{ij} - \tilde{Y}_{i\cdot})^t$$

donde $\tilde{Y}_{i\cdot}$ es el vector de medias del i -ésimo grupo $\tilde{Y}\dots$, es el vector de medias global.

Los estimadores de máxima verosimilitud que permiten estudiar la variación y la covarianza entre y dentro de grupos, que han sido propuestos y estudiados extensivamente [Jöresong (1970), Cronbach (1976)], son:

$$i) \hat{V} = \left(\frac{n}{n-1} \right) \quad (11)$$

$$ii) \hat{V}_s = \left(\frac{1}{k} \right) (S_s - \hat{V})$$

Este problema multivariado se puede adaptar usando el enfoque matricial presentado en la sección 3, considerando la generalización natural. Supongamos que los k grupos son una muestra de una distribución normal $(n \times p)$ -dimensional, con vector de medias $\underline{\mu}$ y la matriz de covarianzas.

$$V_{(np)} = \underline{1}_n \underline{1}_n^t \otimes \Sigma_a + I_n \otimes \Sigma \quad (12)$$

donde $C = A \otimes B$ indica el producto directo de las matrices A y B ; es decir $((c_{ij})) = a_{ij}B$. Note que la expresión en (12) es la generalización natural de la expresión en (6).

El modelo llamado "multivariado doble", puede ser escrito en notación matricial por:

$$\underline{Y}_i^* = \underline{1}_n \otimes \underline{\mu} + \underline{1}_n \otimes \underline{a}_i + \underline{e}^* \quad (13)$$

donde $\underline{Y}_i^* = (\underline{y}_{i1}^t, \dots, \underline{y}_{in}^t)^t$ es un vector de orden $(n \times p)$ y \underline{e}^* es el análogo para los errores aleatorios.

El principal supuesto para la estructura de covarianza en

este modelo puede ser expresado por:

$$\text{Cov}(\underset{\sim}{Y}_{1j}, \underset{\sim}{Y}_{1j'}) = \Sigma_a \quad (14)$$

La prueba de hipótesis que interesaría fundamentalmente para este caso es la extensión natural al caso multivariado de la presentada en la sección 2. La metodología para realizar la prueba ha sido reportada en la literatura, pero no integrada para este caso. Seber (1984), pag. 92-95, presenta una revisión de los resultados generales. Consideramos la hipótesis $H_0: \Sigma_a = 0$ contra $\Sigma_a > 0$. Para probarla proponemos los siguientes pasos.

1.- Probar $H_0: \sigma_j^2 = 0$ vs $H_a: \sigma_j^2 > 0$

Para $j = 1, 2, \dots, p$, usando el procedimiento presentado en la sección 2. Si al menos una de ellas se rechaza, entonces se puede concluir que hay efectos de grupo en la o las variables que resultaran con componentes de varianza distintos de cero. Si todas las variables resultan con efecto de grupo, interesaría probar la hipótesis de efecto conjunto. Esta hipótesis implicaría un nivel de análisis multivariado.

2.- El nivel multivariado de análisis implicaría probar $H_0: \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$, que se sigue de obtener

$$\hat{\Lambda} = \frac{|\Sigma_a|}{m\hat{\sigma}_1^2, \dots, m\hat{\sigma}_p^2}$$

donde $\hat{\sigma}_j^2$ es el estimador de la varianza de grupos asociada a la j -ésima variable, y usando el resultado que nos garantiza que $-r\log \hat{\Lambda}$ se distribuye como una ji-cuadrada con $r = m - (2p+11)/6$ grados de libertad, podemos implementar la regla de decisión. Las tablas de los puntos críticos se presentan en Seber (1984) pag. 612.

Si H_0 en 2 es rechazada entonces se tiene evidencia para decir que la contribución de grupo a las respuestas Y_1, Y_2, \dots, Y_p es conjunta, no independiente, y por tanto es razonable pensar que los factores que determinan el grupo están influyendo en las respuestas, tanto individualmente como de manera conjunta.

Veámos cuál es la significancia conjunta de esta prueba:

$$P(\text{Rech } H_{0j} \mid j = \overline{1, p} \mid H_{0j} \text{ cierto } j = \overline{1, p})$$

$$= P(F_{cj} > F_{(k-1, k(n-1))}^{1-\gamma} \mid j = \overline{1, p} \mid \sigma_j = 0, j = \overline{1, p})$$

$$= P(\min_{j=1, p} F_{cj} > F_{(k-1, k(n-1))}^{1-\gamma} \mid \sigma_j = 0, j = \overline{1, p})$$

$$= P[\bigcap_{d=1}^p (F_{cj} > F_{(k-1, k(n-1))}^{1-\gamma} \mid \sigma_j = 0)]$$

$$\leq \sum_{j=1}^p P(F_{c_j} > F_{(k-1, k(n-1))}^{1-\gamma} \mid \sigma_j = 0)$$

$$= \sum_{j=1}^p \gamma = p\gamma = \alpha \quad \text{por lo que obtenemos } \gamma = \alpha/p$$

Con esto queda determinado que:

$$P(\text{Rech } H_{o_j} \mid j = \overline{1, p} \mid H_{o_j} \text{ cierta} \mid j = \overline{1, p}) \leq \alpha.$$

REFERENCIAS:

- Cronbach L. J. (1976) Research on classroom and Schools Formulation of Questions, Design and Analysis; stanford Evaluation Consortium.
- Mardia K. V. Kent and Bibby J. M. (1979) Multivariate Analysis; Academic Press; London.
- Seber G. A. F. (1984) Multivariate observations; Wiley, New York.
- Keesling J. W. and Wiley D. E. (1974) Regression Models for hierarchical Data; Psichometric Society Spring Meeting, Stanford University.
- Keesling J. W. (1976) Components of Variance Models in Multilevel Analysis; paper presented at a Conference on Methodology for Aggregating Data in Educational Research, Stanford University.
- Jöresong K. G. (1970) A General Method for Analysis of Covariance Structures; Biometrika 57; pp. 239-51.
- Ruiseco M., Ojeda M.M., Luna F., Ortiz I. y Viveros C. (1991) Estudio sobre Información Extracurricular en Estudiantes de la Universidad Veracruzana; Reporte Técnico, LINAE U.V..
- Tiao G. C. and Tan W. Y. (1966) Bayesian Analysis of Random Effect Models in the Analysis of Variance, II, Effect of Autocorrelated Errors; Biometrika 53; pp. 4777-95.

UNA PROPUESTA PARA LA ESTIMACION EN POBLACIONES FINITAS BASADA EN
LA DISTANCIA ENTRE LAS FUNCIONES DE DISTRIBUCION EMPIRICA DE LA
POBLACION Y DE LA MUESTRA

Dr. Alberto Castillo Morales¹

RESUMEN

De manera semejante a como una muestra finita se representa por su función de distribución empírica, una población finita se puede representar por su función de distribución.

La distancia de Kolmogorov entre las funciones de distribución de la población y la muestra se desconoce porque se desconocen los valores que toman los elementos de la población que no fueron seleccionados para la muestra. Los estadísticos de rango permiten obtener la distancia entre la población y cada una de las muestras posibles. Los valores de las distancias permiten clasificar a las muestras por su semejanza con la población, y las frecuencias con que ocurren los valores de distancia producen su distribución de probabilidades.

Conociendo la distribución de las distancias de las muestras a la población, para una combinación específica de tamaños de población y de muestra, se puede hacer una banda que cubra a la distribución empírica de la población. Esta banda será en general de una amplitud que limita su uso, pero el concepto se puede extender para

¹Centro de Estadística y Cálculo, Colegio de Postgraduados, Montecillo.
Edo. México. 56230, Km. 35.5 Carretera México-Texcoco.

obtener bandas menos amplias con probabilidades establecidas. Una vez que se tienen bandas de confianza, es posible utilizar el procedimiento para estimar la media de la población sin necesidad de utilizar la varianza de la muestra como medida de la variabilidad, sino la distribución de probabilidades generada por el proceso de aleatorización.

Con este punto de vista se puede definir correctamente el concepto de muestra típica, y se usa la probabilidad generada por la aleatorización para estimar la distribución de la población y su media.

PALABRAS CLAVE: Aleatorización, muestreo, distribución empírica, distancia, estimación, muestra típica.

NOTACION

Se considera que la población P consta de N elementos distintos etiquetados

$$P = \{ X_1, X_2, X_3, \dots, X_N \}.$$

La función de distribución de la población es

$$F_N(X) = \begin{cases} 0 & \text{si } X < X_{(1)} \\ I/N & \text{si } X_{(i)}, \leq X < X_{(i+1)} \text{ para } i=1,2,\dots,N-1 \\ 1 & \text{si } X \geq X_{(N)}, \end{cases}$$

donde $X_{(i)}$, se refiere al $i^{\text{ésimo}}$ estadístico de orden en la población, esto es, $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(N-1)} \leq X_{(N)}$.

La muestra S está compuesta de n elementos tomados de la población. Así,

$S = \{ Y_1, Y_2, \dots, Y_n \}$

es una muestra si S está contenido en P . La muestra S de n elementos genera una función de distribución empírica F_n . Nótese que para cada Y_i en la muestra, hay una X_j en la población tal que $Y_i = X_j$.

MEDIDA DE DISTANCIA

Dada una población y finita una muestra, o sus correspondientes distribuciones, se puede obtener la distancia de Kolmogorov (Gibbons, J.D., 1971) dada por

$$D = \max, \{ |F_N(X_j) - F_n(Y_j)| ; j=1, \dots, N \}. \quad (1)$$

Por facilidad de enumeración, se puede trabajar con los estadísticos de orden, o todavía más, con los estadísticos de rango con respecto a los valores poblacionales, tanto de la población como de la muestra, y se obtiene:

$$D = \max, \{ |F_N(j) - F_n(j)| ; j=1, \dots, N \}, \quad (2)$$

donde j representa al punto donde ocurre el j ésimo salto de tamaño $1/N$ de la distribución de la población.

En general, la distancia D está acotada entre cero y uno, pero en poblaciones finitas de N elementos distintos, para cualquier S que sea una muestra de n elementos distintos

$$D(S) \leq (N-n)/N.$$

También se puede ver que si $n < N$ no puede ocurrir una muestra con distancia cero a la población.

La cota máxima para la distancia entre una muestra y la población indica que para poblaciones finitas con elementos distintos, las muestras que se basan en una pequeña proporción de la población

pueden resultar distantes de ella. Se puede ordenar a las muestras de acuerdo con su distancia a la población para decidir cuáles son menos distantes y por ello más típicas (Casiàn, M. y Castillo, A. 1990, Kruskal, W. y Mosteller, F. 1979a, 1979b), con respecto a las funciones de distribución empíricas.

CLASIFICACION DE LAS MUESTRAS

En la Tabla 1. del Apèndice se puede ver que la distancia dada por la acotación máxima efectivamente ocurre, y como era de esperarse, hay una concentración de frecuencias para muestras con distancias intermedias, para terminar con pocas muestras para las distancias extremas, tanto las menores como las mayores.

Las muestras con mayores distancias serán las que más se aparten de la distribución de la población, y corresponden a las muestras donde las observaciones se aglomeran hacia un solo lado de la población.

Para tener la seguridad de obtener una muestra cuya función de distribución empírica se aleje poco de la población, se debe pensar en tamaños de muestra muy superiores a lo que la práctica del muestreo acostumbra, ya que la distancia máxima debe ser pequeña, esto es, $(N-n)/N$ debe ser pequeña, digamos q, forzando que $n \geq N(1-q)$. Si se recuerda que q es un número entre 0 y $(N-n)/N$, y que una distancia pequeña, dependiendo del problema puede ser del orden de .4. se tienen muestras del 60% de la población.

Por ejemplo, para una población de 10 elementos distintos, si se desea tener distancia máxima de .2 se necesita tomar muestras con 8 elementos, para distancia máxima de .3 se reduce a 7 elementos y

para distancia màxima de .4 se necesitan 6 elementos.

Si se desea que la distancia entre la población y la muestra sólo rebase a una cota previamente establecida para las distancias con cierta probabilidad de ocurrencia, es necesario considerar la distribución de probabilidades de las muestras de acuerdo con sus distancias a la población. La distribución está dada, para el caso de muestreo irrestricto aleatorio, por la frecuencia relativa que se obtiene de las frecuencias dadas en la Tabla 1 del Apèndice.

En el ejemplo de una población con 10 elementos, si se desea que la distancia entre la población y la muestra no rebase a .2 con cierta probabilidad de ocurrencia, digamos .8, se necesita, para el mismo ejemplo, tener muestras de 7 elementos, que dan una probabilidad de .833 y para que la distancia entre la muestra y la población no rebase .3 con probabilidad .8 se pueden tomar muestras de tamaño 5, para tener .921 de probabilidad. En estos casos no se tiene un tamaño de muestra que produzca la probabilidad deseada con exactitud debido a que la distribución de probabilidades de las muestras es discreta.

Puede pensarse que la medida de distancia, por basarse en el màximo, exagera las diferencias entre las muestras y la población. Si una muestra presenta una distancia grande, parecería que la medida que da lugar a la distancia ocurre en sólo un punto de la distribución de la población, estando, por lo demás, cerca de ella. Este no es el caso, pues las diferencias entre F_N y F_n se incrementan punto a punto, y si se presenta una distancia grande en $X_{(k)}$, quiere decir que una distancia parecida se presenta en $X_{(k-1)}$ o/y en $X_{(k+1)}$.

En ocasiones se tienen limitaciones que forzan la selección de cierto tamaño de muestra. En estos casos, para tener una idea de que tan grande es una distancia, conviene revisar las frecuencias con que ocurren los diferentes valores de distancias en la combinación específica de N y n . Las frecuencias acumulativas pueden ser buenos indicadores, pues dan el porcentaje de muestras que son mayores o iguales a una distancia dada. En el caso del ejemplo con 10 elementos, para muestras con 6 elementos, el 0.95% de las muestras están a distancia igual o mayor que .4 de la población, el 4.7% de las muestras dista .33 o más, el 9.5% está a distancia igual o mayor que .3 el 18.1% dista .26 o más y el 29.5% dista .23 o más de la población.

CONSTRUCCION DE BANDAS DE CONFIANZA

El conocimiento las frecuencias con las que ocurren de las distancias de la población a las muestras, permite que a partir de una muestra se pretenda conocer la función de distribución de la población. El procedimiento que se sugiere es tomar F_{n-c} y F_{n+c} para acotar a F_n , buscando un valor apropiado para c .

Una primera opción para determinar el valor de c , consiste en forzar que entre F_{n-c} y F_{n+c} se encuentre F_n . Para ello, se toma la distancia máxima como valor para c .

Debido a lo grande que puede resultar c , se puede buscar que F_n esté dentro de las bandas dadas por F_{n-c} y F_{n+c} con cierta probabilidad p . La probabilidad de ocurrencia de las muestras está dada por el esquema de aleatorización. Si cada muestra tiene igual probabilidad de ocurrir, la Tabla 1 da la frecuencia, de la cual se

obtienen la probabilidad al dividir entre la frecuencia total. Para encontrar el valor apropiado de c, sólo se tiene que tomar el valor de la distancia, en la tabla de frecuencias apropiada, que garantice que 100p% de las muestras tienen una F_n que está a una distancia menor o igual que c de F_n . Por ejemplo, para la población que se ha venido utilizando de 10 elementos distintos y $n=6$, con probabilidad .9524 se cubre a F_n usando $c=.3$, y para $c=.2667$ se tiene una probabilidad de .9048.

Notese que sólo se ha supuesto igual probabilidad para cada una de las muestras posibles, además, se utiliza la distribución de probabilidades generada por la aleatorización como base para formular las bandas, en este caso, de confianza.

La diferencia fundamental entre este método de estimación y los usuales en muestreo, es que se utiliza la distribución de probabilidades generada por la aleatorización, pero no al nivel de muestra, sino tal como se genera y define. Con respecto a los nuevos métodos basados en superpoblaciones, se evita complicar el modelo, y se utilizan sólo las probabilidades generadas por el proceso de elección.

ESTIMACION DE LA MEDIA

No es muy común que se desee estimar a la función de la distribución empírica de la población. Con mayor frecuencia se desea estimar a la media o al total de la población. Para ello, se puede pensar en utilizar a las funciones que definen a las bandas de confianza y calcular los valores medios correspondientes. Surge un problema, pues ninguna de F_{n-c} ni F_{n+c} es una función de

distribución, debido a que la primera no llega a 1 y la segunda no empieza en 0.

Se sugiere la solución que consiste en acotar a la población con un valor mínimo y un valor máximo. La suposición de estos valores permite que las funciones que forman las bandas arriba y abajo de la función de distribución empírica de la muestra sean funciones de distribución. En muchos de los problemas de muestreo donde las unidades se eligen por sorteo, se tiene una idea bastante clara de los valores máximo y mínimo que deben tomar los valores de la variable de interés, por lo que se adicionan supuestos que no lesionan o complican el modelo.

Si las cotas mínima y máxima para los valores de X son X_a y X_m , respectivamente, para un valor c , tal que $(j-1)/n \geq c < j/n$, se tiene:

$$F_i(X) = \begin{cases} \max \{ 0, F_{n-c} \} & \text{si } X < X_m \\ 1 & \text{si } X \geq X_m, \end{cases}$$

y

$$F_d(X) = \begin{cases} 0 & \text{si } X < X_a \\ \min \{ F_{n+c}, 1 \} & \text{si } X_a \leq X. \end{cases}$$

Para la estimación de la media, conviene escribir las funciones de probabilidad correspondientes:

$$f_i(X) = \begin{cases} j/n - c & \text{si } X = X_{i,j}, \\ 1/n & \text{si } X = X_{i,j+1}, \dots, X_{i,n}, \\ c & \text{si } X = X_n \\ 0 & \text{de otra manera,} \end{cases}$$

y

$$f_d(X) = \begin{cases} c & \text{si } X = X_n \\ 1/n & \text{si } X = X_{i,1}, \dots, X_{i,n-j} \\ j/n - c & \text{si } X = X_{i,n-j+1} \\ 0 & \text{de otra manera,} \end{cases}$$

Para la estimación de la media aritmética de la población, se utilizan las medias aritméticas dadas por f_i y f_d . Se obtiene entonces que

$$X_d = \sum_{k=1}^{n-j} X_{i,k} / n + j X_{i,n-j+1} / n - c \{ X_{i,n-j+1} - X_n \}$$

y

$$X_i = \sum_{k=j+1}^n X_{i,k} / n + j X_{i,j} / n + c \{ X_n - X_{i,j} \}.$$

Notese que los intervalos de confianza se construyen utilizando a las cotas supuestas y eliminando hacia cada lado de la banda a los valores extremos observados, tantos como veces quepa $1/n$ en c . Conviene que las bandas de confianza se formen con valores de c que se obtengan de la distribución de probabilidades de las distancias, eligiendo el valor de manera que se tenga una probabilidad dada de

que la muestra elegida acote a la distribución de la población. Por ejemplo, para N=10 y n=7, se puede tomar a la distancia de .2 que garantiza que el 83.3% de las muestras darán una banda que cubre a la población, o .2143 que da la misma garantía para el 93.3% de las muestras. Se puede comenzar con una distancia que tenga sentido para el trabajo, aunque habrá que desarrollar experiencia antes de poder dar valores que tengan sentido práctico para determinado tipo de estudios. Si en el ejemplo se decide que la distancia no exceda .3 se tiene que el 100% de las muestras darán una banda que cubre a la población.

REFERENCIAS

- Casià M. M. A. Y Castillo M. A. (1990). Algunas reflexiones sobre los estudios por muestreo en la actividad agropecuaria. Monografías y Manuales en Estadística y Cómputo, CEC., CP. Vol. 9, Núm. 4.
- Gibbons, J. D. (1971). Non parametric statistical inference. McGraw Hill Book Co, New York. Pag. 75.
- Kruskal, W. and Mosteller, F. (1979). Representative sampling, II: Scientific literature, excluding Statistics. International Statistical Review, 47: 111-127.
- Kruskal, W. and Mosteller, F. (1979). Representative sampling, III: The current Statistical literature. International Statistical Review, 47: 245-265.

VII Foro Nacional de Estadística
7 al 11 de septiembre de 1992
Universidad de las Américas, Puebla

**La aplicación del modelo de alta y baja movilidad
en el estudio de la migración**

Vázquez Benítez, Gabriela*

Tijuana B.C., octubre de 1992

* Investigadora del Departamento de Estudios de Población, El Colegio de la Frontera Norte, Blvd. Abelardo L. Rodríguez no. 21 Zona del Río, Tijuana Baja California.

R e s u m e n

En este trabajo se presenta el modelo de alta y baja movilidad, a fin de evitar el empleo del supuesto de homogeneidad de las cohortes, al utilizar las probabilidades de transición de dos grupos homogéneos hacia el interior y diferenciales entre si, en el estudio de la migración bajo la perspectiva de los modelos multiregionales. La aplicación de este modelo busca reconciliar información referida a diferentes períodos, que presentan diferentes resultados. La explicación de estos resultados no necesariamente se debe a diferentes calidades en la información o patrones, sino que también se debe a deficiencias metodológicas.

Introducción

El estudio de los fenómenos demográficos, requiere del desarrollo de herramientas metodológicas, las cuales no deben permanecer en el plano abstracto, sino que deben ser concebidas en función de su futura aplicación en investigaciones específicas dentro de una perspectiva multidisciplinaria, con base en la interrelación entre los métodos estadísticos, la búsqueda de su adecuación a los marcos teóricos, así como a la información misma.

La migración, por sus propias características de temporalidad, espacialidad, repetición y selectividad, es uno de los fenómenos demográficos que mayor dificultad presenta para su medición y por lo tanto, requiere de metodologías más sofisticadas para abordarlo.

Una de las herramientas que se han utilizado en el quehacer demográfico ha sido la tabla de vida, la cual es un modelo probabilístico, cuyo objetivo es representar el comportamiento de la mortalidad por edad de una población en términos de una función de sobrevivencia.

El estudio de otros fenómenos a partir del uso de esta herramienta, que por sus características de repetición no pueden ser expresados en una tabla unidimensional como la mortalidad, es

posible a través de las tablas de vida por estados múltiples (o tabla de incrementos-decrementos). Ejemplos de ello pueden ser la migración, la nupcialidad, la participación en la actividad económica, fenómenos que están conformados por un sistema de más de un "estado", entre los que la población transita repetidas veces.

Para aplicar este tipo de análisis en el estudio de la población, se requiere establecer algunos supuestos que serán descritos en el siguiente apartado, junto con una descripción general del modelo multiregional; dichos supuestos deben ser planteados atendiendo a la dificultad de su adecuación a la realidad social. Uno de ellos es la homogeneidad entre los miembros de la cohorte respecto a la intensidad con que se presenta el fenómeno, ya sea mortalidad, migración o cualquier otro (i.e. la fuerza de movimiento a la edad x , $\mu(x)$).

En el tercer apartado, se presenta el **modelo de alta y baja movilidad**, como forma de evitar la necesidad del supuesto de homogeneidad de los miembros de una cohorte, al identificar dos grupos homogéneos hacia el interior y diferenciales entre si.

Este modelo busca conciliar información referida a períodos con diferente longitud, con los que se obtienen estimaciones distintas. La explicación de estos resultados no necesariamente se debe a diferencias en la calidad de la información sino a diferentes patrones a lo largo del tiempo y a la forma de abordarlos; de ahí la importancia del avance en la metodología y el acercamiento de ésta al estudio aplicado de los fenómenos sociales, y en especial a la migración.

Modelos multiregionales

La tabla de vida por estados múltiples (modelo multiregional) describe la sobrevivencia y distribución de una o más cohortes a edades sucesivas a través de un sistema hasta la extinción del último miembro. El sistema se conforma por los estados transitorios, excluyentes y exhaustivos, determinados por alguna variable de la población. La base de la tabla de vida por estados múltiples es el supuesto markoviano, que consiste en asumir que la probabilidad de cambiar de un estado a otro sólo depende de la condición presente del individuo y no de su historia anterior¹.

El proceso puede entenderse a partir de la siguiente relación: piénsese en individuos que se encuentran en el estado i a la edad x , nacidos en cualquier estado y en un intervalo de tiempo suficientemente pequeño en el que sólo puedan realizar un movimiento, entonces se les presentan las siguientes opciones: permanecer en el estado i , pasar a otro estado j y fallecer (moverse al estado de absorción δ). Denótese a los que pasan del estado i al j como $d_{ij}(x)$ y a los que fallecen como $d_{i\delta}$, y a la cohorte sobreviviente en el estado i como $l_i(x)$.

^{1.} Un proceso markoviano finito (número finito de estados) y discreto $\{X(t)\}$ es aquél que para cualquier $t_0 < t_1, \dots < t_i < t_{i+1}$ y estados k_0, k_1, \dots, k_{i+1} : $P\{X(t+i)=k_{i+1} | X(t_0)=k_0, \dots, X(t_i)=k_i\} = P\{X(t_{i+1}) | X(t_i)=k_i\}$. Las propiedades de los procesos markovianos implican relaciones interesantes en las probabilidades de transición P_{ij} que pueden ser expresadas por la ecuación de Chapman-Kolmogorov. La ecuación diferencial de Kolmogorov es: $d/dt P_{ij}(t, t+h) = P_{ik}(t, t+h) \mu_{kj}(t+h)$ con la condición inicial de $P_{ij}(t, t)=1$ si $i=j$ y 0 en otro caso; donde μ_{kj} es la tasa instantánea de movimiento del estado k al j .

Entonces, los sobrevivientes a la edad $x+dx$ en el estado i ($l_i(x+dx)$) son los que se encuentran vivos a la edad x menos aquellos que parten a otros estados o fallecen (decrementos) más los que se incorporan (incrementos):

$$l_i(x+dx) = l_i(x) + \sum d_{ji}(x) - \sum d_{ij}(x) - d_{is} \quad (1)$$

Sustituyendo $d_{ij}(x)$ por $\mu_{ij}(x) l_i(x) dx$, en la ecuación 1, obtenemos:

$$l_i(x+dx) = l_i(x) - \sum \mu_{ij}(x) l_i(x) dx + \sum \mu_{ji}(x) l_i(x) dx \quad (2)$$

donde $\mu_{ij}(x)$ es la fuerza instantánea de movimiento, del estado i al j y; cuya expresión en forma matricial es²:

$$l(x+dx) = l(x) - \mu(x) l(x) dx \quad (3)$$

donde $\mu(x)$ es la matriz de movimiento:

$$\mu(x) = \begin{matrix} \sum_{j=1} \mu_{1j}(x) & -\mu_{21}(x) & \dots & \dots \\ -\mu_{12}(x) & \ddots & & \\ \vdots & & \ddots & \dots \\ \dots & \dots & \dots & \sum_{j=s} \mu_{sj}(x) \end{matrix}$$

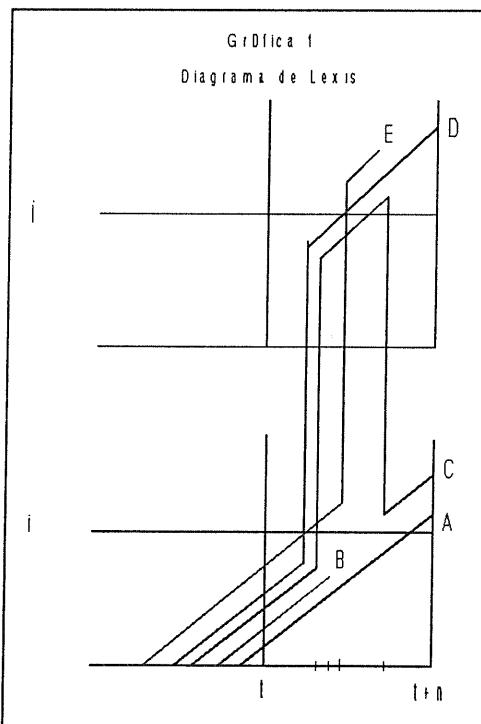
Para la aplicación del modelo antes descrito, existen dos tipos de información: la obtenida con base en observaciones de la población en dos momentos en el tiempo, o a través del registro de sus movimientos en un intervalo de tiempo.

². Esta expresión corresponde a la ecuación de Kolmogorov.

El primer caso se refiere a un enfoque de transición, e identifica como cambio, la presencia de un individuo en dos estados distintos y dos momentos en el tiempo, la medida que se obtiene es la probabilidad perspectiva observada (S_x) y se supone igual a la de la tabla. Este enfoque resulta de una pregunta retrospectiva en censos o encuestas, sobre el estado en la que se encontraba la persona n años antes, generalmente el tiempo de referencia es de 1, 5 ó 10 años. Un ejemplo para la migración es el censo de 1990, en el cual se preguntó sobre el lugar de residencia en 1985.

El segundo corresponde al enfoque de movimiento, que se construye a partir de todos los cambios que efectúan los miembros de una población en un intervalo de tiempo, considerados éstos como eventos y referidos a la población expuesta; lo que se obtiene son las tasas centrales de movimiento observadas (nM_x), las que se suponen iguales a las de tasas de movimiento de la tabla (m_x). Este enfoque corresponde a la forma clásica en que se construye la tabla de mortalidad en las sociedades actuales y generalmente se encuentran referidas

Gráfica 1
Diagrama de Lexis



a un período anual; la información se obtiene a partir de registros, que corresponden a los eventos (estadísticas vitales o registros continuos) y de censos o encuestas, que dan cuenta de la población expuesta al fenómeno (Ledent 1980). Hasta la fecha no se cuenta con registros continuos de migración, sin embargo se contempla la realización de una encuesta continua de migración internacional, de la cual se hará referencia adelante.

La gráfica 1 es un diagrama de lexis para un sistema de dos estados, representados uno arriba del otro, conectados por líneas verticales; en ella se muestra la historia de vida de 5 hipotéticos individuos que transitan, entre dos momentos, de uno a otro estado, o a la muerte. El diagrama de lexis permite identificar la diferencia entre uno y otro enfoques. Por ejemplo, para el de movimiento, la línea C corresponde a dos cambios en el intervalo t a $t+n$, mientras que para el de transición no, al haber una migración de retorno; para el enfoque de transición, sólo la línea D representaría un cambio³.

El modelo de tabla de vida por estados múltiples en su desarrollo continuo requiere del supuesto markoviano; y al pasar al caso discreto, que las tasas de movimiento se supongan constantes⁴; o en su defecto, que la función de sobrevivencia sea lineal en los

^{3.}. Las línea A representa a un individuo que no efectuó ningún movimiento, mientras las líneas B y E se refieren a dos personas que murieron, sin haber realizado ningún cambio, y la otra después de haber migrado.

^{4.}. $P(x,n)=\exp[-n_n m_x]$

intervalos de edad con que se estén trabajando (edades individuales o quinquenales)⁵.

Para la aplicación a un fenómeno concreto deben añadirse los siguientes supuestos sobre la población en estudio:

i) los 'movimientos' entre estados en un intervalo de edad (x a $x+n$) son independientes del origen de la cohorte (este supuesto es necesario en aplicaciones donde existe más de una cohorte inicial al nacimiento como es la migración, sin embargo se puede eliminar si se cuenta con la información de movimientos referidos a la población según su lugar de nacimiento (Ledent 1981));

ii) que la población que realiza los movimientos esté sujeta a la misma fuerza de movimiento ($\mu(x)$) (supuesto de homogeneidad de las cohortes, sobre el cual se discutirá en el siguiente apartado);

iii) que la población no sea perturbada por otros fenómenos demográficos.

Modelo de alta y baja movilidad

Antes de presentar el modelo mencionaremos algunas consideraciones hechas por Rogers (1992):

"Individuos con altas propensiones de muerte en una población dejarán de pertenecer a ella en mayor número que aquellas con tasas menores... El efecto selectivo hace imposible determinar si existe una población heterogénea con diferentes tasas de movimiento entre sus subpoblaciones o una población homogénea con una misma tasa pero decreciente para todos sus miembros."

⁵. $n^P_x = [I + n/2 \ n^m_x]^{-1} [I - n/2 \ n^m_x]$

Al igual que una población como la que se refiere Rogers en un proceso unidimensional, este efecto está presente en las poblaciones analizadas bajo un sistema de estados múltiples.

Bajo esta perspectiva, en el artículo de Kistul y Philipov (1981), se presenta el desarrollo del modelo de alta y baja movilidad (migrantes crónicos y no crónicos). Dicho modelo permite equiparar estimaciones con base a información de un año con respecto a la de intervalos mayores (5 años), ya que la heterogeneidad tiene un efecto en períodos más largos en los que algunos miembros tuvieron la posibilidad de moverse varias veces (crónicos) mientras que otros a lo más lo hicieron una vez (no crónicos); se parte de la identificación de las probabilidades de transición de dos subpoblaciones que al interior de cada una sean homogéneas y diferentes entre las subpoblaciones.

En el caso de la migración, existen pocas fuentes de información en México que permiten la construcción de tablas de origen-destino (modelo multiregional)⁶. Los censos hasta el de 1980, se referían al último cambio de residencia, que bajo algunos supuestos permite calcular la migración interestatal (o intraregional) en el último año aproximándose al enfoque de movimiento. En el más reciente censo (1990), se hizo referencia a

^{6.} En el caso de la migración internacional, en el Colegio de la Frontera Norte se está llevando a cabo una investigación que consiste en una encuesta a migrantes mediante la aplicación de la metodología que han denominado poblaciones móviles y que permitirá obtener estimaciones del flujo de migración a lo largo del año.

la residencia 5 años antes, es decir, utilizando el enfoque de transición.

Debido a la carencia de información sistemática sobre migración es difícil entender las diferencias en cuanto a la estimación entre uno u otro tipo de datos, ya que si bien puede estar presente la distinta calidad de la información, pueden ser resultado de diferentes patrones de migración y la metodología con que se aborda.

Considérese una población multiregional, por ejemplo sujeta a la migración, para la cual se tenga la información necesaria para la construcción de la tabla referida a uno o cinco años ($n=1$ y $n=5$).

Bajo el supuesto de linealidad en el intervalo de la función de sobrevivencia (i.e. los eventos (migraciones) ocurren uniformemente en el intervalo) se obtiene la probabilidad de transición, a partir de las tasas de movimiento, mediante las siguientes ecuaciones:

$$P_x^5 = [I + \frac{5}{2}M_1(x)]^{-1} [I - \frac{5}{2}M_1(x)] \quad (5)$$

en el caso de que se tenga la información de un año ($n=1$), y

$$P_x^5 = [I + \frac{1}{2}M_5(x)]^{-1} [I - \frac{1}{2}M_5(x)] \quad (6)$$

para el caso en que ésta sea por quinquenios ($n=5$).

O bien, bajo el supuesto de la tasa de movimiento constante:

$$P_x^5 = e^{-5M_1(x)} \quad (7)$$

y

$$P_x^5 = e^{-M_5(x)} \quad (7')$$

Kitsul y Philipov (1981) muestran mediante ejemplos que las estimaciones obtenidas de la información de un año ($n=1$) y de un lustro ($n=5$) son inconciliables, aun y cuando el supuesto de tasa de movimiento constante (ecuación 7) presente mejores resultados al considerar los movimientos múltiples (ej. migración de retorno). Véase gráfica 1. Sin embargo el resultado puede ser más preciso al aplicar el modelo de alta y baja movilidad.

Supóngase que la población está compuesta por dos subpoblaciones con diferentes tasas de movimiento por edad, entonces la matriz de transición $P_n(x)$ se puede expresar como una combinación lineal entre las dos matrices de probabilidad asociadas a cada subpoblación π y p :

$$P_n(x) = \alpha(x)\pi_n(x) + [I - \alpha(x)]p_n(x) \quad (8)$$

donde $\alpha(x)$ es un escalar que depende de la edad pero no del estado inicial ni final, $0 < \alpha < 1$; y, π y p representan procesos markovianos (aun y cuando P_n no necesariamente lo sea); π y p se relacionan con sus respectivas tasas instantáneas de movimiento de la siguiente forma:

y

$$\pi_n(x) = e^{n\mu_n(x)} \quad (9)$$

$$\rho_n = e^{n\mu_p(x)} \quad (9')$$

Entonces la ecuación 8, en términos de las tasas instantáneas de movimiento, queda expresada como:

$$P_n(x) = \alpha(x) e^{n\mu_n(x)} + (1-\alpha(x)) e^{n\mu_p(x)} \quad (10)$$

con n igual a 1 ó 5.

Si se supone que la tasa de movimiento asociada a una de las subpoblaciones (digamos μ_p) es proporcional a la otra, la podemos expresar de la siguiente manera:

$$\mu_p(x) = k(x) \mu_n(x) \dots \text{con } k(x) \in (0, 1)$$

entonces, la ecuación 10, omitiendo la edad en α y k queda:

$$P_n(x) = \alpha e^{n\mu_n(x)} + (1-\alpha) e^{nk\mu_n(x)} \quad (12)$$

para n=1 y n=5.

La obtención de los parámetros α y k y la matriz de movimiento μ_n , a partir de la relación entre las matrices de probabilidad P_1 y P_5 , es mediante la diagonalización de la matriz de movimiento μ_n y por lo tanto de P_1 y $P_5(x)$ ⁷; una vez realizado este procedimiento,

7. Se tienen n^2+2 incógnitas con $2n^2+n$ ecuaciones. Para encontrar una solución que relacione P_1 con P_5 sólo se puede mediante la diagonalización con lo que se reduce a $2n-2$ ecuaciones y $n+1$ incógnitas.

con un algoritmo de optimización no lineal se pueden obtener los parámetros $\alpha(x)$, $k(x)$ ⁸.

Kitsul y Philipov encontraron, en la aplicación del método descrito anteriormente al caso de la migración de Gran Bretaña, que $k(x)$ variaba muy poco respecto a la edad, $\alpha(x)$ y los eigen-valores v_i representaban la curva típica de migración. Por lo que los autores proponen que si se utiliza la información de un año y el modelo de alta y baja movilidad se podría estimar la probabilidad de transición para 5 años, con mejores resultados, no obstante que es necesario conocer la pauta típica de migración⁹.

Consideraciones finales

Si se toman como válidos los resultados obtenidos por dichos autores, es decir que las probabilidades de transición P_1 y P_5 se relacionan mediante el modelo de alta y baja movilidad, al contar con información para un año se podría obtener estimaciones más confiables para períodos más largos. Esto permitiría estimar la migración interestatal o internacional a lo largo de los años que transcurren de un censo a otro y por lo tanto completar los componentes de la dinámica demográfica de forma más precisa.

⁸. El algoritmo consiste en construir una función a partir de las ecuaciones del sistema dependientes de los eigen-valores v_i , los parámetros α y k , para posteriormente minimizarla en base a optimización no lineal.

⁹. Se tendría que suponer α y k constantes por edad e iguales para el total de la población; o bien, k constante e igual a la del total de la población y $\alpha(x)$ como la estructura por edad de la migración que se conozca para dicha población.

Sin embargo queda por comprobar si el modelo es aplicable al caso mexicano, tanto para la migración interna como para la internacional.

Referencias

- Kitsul, P. y Philipov D (1981), "The one-year/five-year migration problem" Advances in multiregional demography, RR-81-6, IIASA Laxenburg, Austria, 5-1981, 1-34.
- Ledent, J. (1981), "Constructing multiregional life tables using place of birth-specific migration data", Advances in multiregional demography, RR-81-6, IIASA, Laxenburg, Austria, 5-1981, pp. 35-50.
- Ledent, J. (1982), "Tablas de vida de estados múltiples: perspectivas de movimiento y transición", Demografía y Economía, 16(3) 51, El Colegio de México, México, pp. 399-438.
- Partida B., V. (1982), "Aplicación del modelo multiregional de población de población al caso de México", Demografía y Economía 16(3) 51, El Colegio de México, México, pp. 449-481.
- Rogers, A. (1975), Introduction to multiregional mathematical demography, John Wiley, N.Y.

**VII FORO NACIONAL DE ESTADISTICA
7 A 11 DE SEPTIEMBRE DE 1992
UNIVERSIDAD DE LAS AMERICAS
PUEBLA**

**El estudio de la mortalidad en el pasado mediante
la aplicación del modelo de poblaciones estables:
Istmo de Tehuantepec, siglo XVIII**

**Gutiérrez Montes, José Rodolfo*
Vázquez Benítez, Gabriela***

* Investigadores del Departamento de Estudios de Población; El Colegio de la Frontera Norte. Blvd. Abelardo L. Rodríguez, No. 21, Zona del Rio, Tijuana, Baja California. C.P. 22300

R E S U M E N

En el presente trabajo, inscrito dentro de una investigación que se lleva a cabo sobre las condiciones de la mortalidad en el Istmo de Tehuantepec durante la segunda mitad del siglo XVIII, se presenta la estimación de diversas tablas de vida, mediante la aplicación del modelo de poblaciones estables, debido a las limitaciones que se tienen al trabajar con información referida a poblaciones del pasado. La perspectiva bajo la cual se ha desarrollado el trabajo, busca establecer una relación entre los fenómenos históricos y su cuantificación estadística, considerando la presencia de fenómenos perturbadores del modelo que deben tomarse en cuenta al momento de enfrentar el análisis, ya que de alguna manera pueden hacer variar los resultados.

Introducción

La escasés de información para el análisis de las poblaciones en el pasado, ha requerido el desarrollo de metodologías para enfrentar dicho problema. Dentro de la Demografía Histórica, como rama de la Demografía, se han realizado algunos avances encaminados a ello; se pueden citar la reconstrucción de familias y la aplicación de los **modelos de poblaciones estables** para la elaboración de las **tablas de vida**.

La tabla de vida, es un modelo probabilístico, que permite estimar las medidas básicas demográficas de una sociedad, tales como la estructura etárea de una población o los niveles de mortalidad y fecundidad. A partir del modelo de poblaciones estables es posible construir una tabla de vida sin necesidad de contar con información completa sobre nacimientos, defunciones y estructura por edad de la población.

Los datos para Tehuantepec con que se cuenta para el desarrollo del trabajo, fueron tomados del Censo de Revillagigedo de 1792-93, y del levantado en 1777, así como de los registros parroquiales

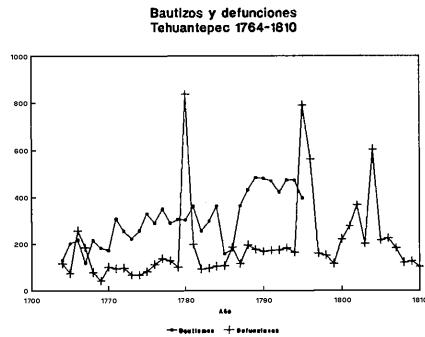
de bautismos y entierros, a lo largo de la segunda mitad del siglo XVIII. Se aborda el problema con base en dos enfoques: uno a partir de la estructura de la población y el otro con base en la de las defunciones. Se realiza también una comparación de los resultados, a fin de identificar su robustez.

Antecedentes

Es necesario, previo a la presentación del desarrollo metodológico, hacer algunas consideraciones de carácter histórico, a fin de no perder la perspectiva de la época.

No se cuenta con registros Gráfica 1 del volumen de población sistemáticos, tal y como ahora se conocen, por ello las posibilidades de describir las características de los habitantes de esa época, depende de la información recabada en los registros parroquiales, que han sido consultados en el ramo de Genealogía del Archivo General de la Nación (AGN-GEN).

Entre 1750 y 1810, se han logrado identificar 6 crisis causadas por epidemias, que fueron de sarampión, viruelas y diarreas, que afectaron sobre todo a la población infantil, representadas



claramente en la gráfica 1, en la que se encuentran las tendencias de mortalidad y natalidad de la época.

La presencia de epidemias en la Villa de Tehuantepec adquiere particular importancia en el concierto nacional, ya que se le ubica como puerto de entrada de algunas, en tanto se manifestaron a lo largo del territorio de la Nueva España, siguiendo las rutas comerciales y de tránsito.

Desarrollo Metodológico

Como antecedente a la presentación del modelo demográfico de poblaciones estables, se hace una breve descripción del proceso en que se basa la construcción de la tabla de vida: Supóngase una cohorte con $N(0)$ personas al nacimiento, sujeta a la fuerza de decremento $\mu(x)$ a la edad x . Sea $N(x)$ el número de sobrevivientes a la edad x , entonces $\mu(x)$ se puede expresar como sigue:

$$\mu(x) = \frac{-dN(x)}{N(x)} = \frac{d\ln N(x)}{dx} \quad (1)$$

que representa el negativo del cambio proporcional de $N(x)$ por unidad de cambio en x .

Al integrar ambos lados de la ecuación, desde la edad 0 a una edad a , y aplicando la exponencial, se tiene:

$$N(a) = N(0) e^{-\int_0^a \mu(x) dx} \quad (2)$$

Si se introduce la variable tiempo en el proceso, $N(x,t)$ representa el número de sobrevivientes de la cohorte a la edad x al tiempo t , sujeta a $\mu(x,t)$. Esta función es, de igual forma, el negativo del cambio proporcional del tamaño de la cohorte por unidad de cambio en la edad (x) al tiempo (t), que corresponde a la diferencial total de la función $\ln N(x,t)$:

$$-\mu(x,t) = \frac{\partial \ln N(x,t)}{\partial t} + \frac{\partial \ln N(x,t)}{\partial x} \quad (3)$$

el primer término de la derecha representa la tasa de crecimiento a la edad x al tiempo t ($r(x,t)$), por lo que la ecuación queda:

$$\frac{\partial \ln N(x,t)}{\partial x} = -\mu(x,t) - r(x,t) \quad (4)$$

Integrando de la edad 0 a la edad a y al aplicar la exponencial, se puede expresar a $N(a,t)$ de la siguiente manera:

$$N(a,t) = N(0,t) e^{-\int_0^a \mu(x,t) + r(x,t) dx} \quad (5)$$

que equivale a la ecuación 2, bajo el impacto que tiene el ritmo de crecimiento por edad en la estructura de la población. Despejando el término $\mu(x,t)$ obtenemos la siguiente relación:

$$p(a,t) = e^{-\int_0^a \mu(x,t) dx} = \frac{N(a,t)}{N(0,t)} e^{-\int_0^a r(x,t) dt} \quad (6)$$

donde $p(a,t)$ representa la probabilidad de sobrevivencia desde la edad 0 a la edad a .

Existen dos formas para evaluar empíricamente esta función. La primera es con las tasas de exposición por edad al momento t , lo que requiere conocer los eventos y la distribución de la población en riesgo de fallecer. La segunda es mediante la distribución de la población en riesgo, que permite estimar el número de sobrevivientes de edad a , al tiempo t ($N(a,t)$) y la tasa de crecimiento en el período ($r(a,t)$), al no contar con la información de las defunciones situadas en el mismo momento que la de población, o debido a la deficiente calidad de los datos.

El modelo de poblaciones estables se basa en modelos macrodemográficos de crecimiento, desarrollados al inicio del presente siglo por Lotka (Lotka, 1973). El supuesto implícito es que la estructura y fuerza de mortalidad y fecundidad permanecen invariantes en el tiempo, en una población cerrada, por lo que su estructura permanece constante.

Supóngase que la población total (P^t), sigue un crecimiento exponencial r ($P_t = P_0 e^{rt}$) y dado que las tasas de natalidad (b) y de mortalidad (d) permanecen constantes a través del tiempo, los nacimientos y las defunciones siguen el mismo comportamiento ($B(t) = B(0)e^{rt}$ y $D(t) = D(0)e^{rt}$).

La estructura de una población estable está dada por:

$$c(a) = b e^{-ra} e^{\int_0^a \mu(x) dx} = b e^{-ra} p(a) \quad (7)$$

Si la tasa de crecimiento y la probabilidad de sobrevivencia son conocidas, se puede estimar la tasa de natalidad, despejando b de la ecuación 7:

$$b = \frac{1}{\int_0^w e^{-rx} p(x) dx} \quad (8)$$

Si se conocen la estructura de la población y las tasas de crecimiento y de natalidad, se obtiene el patrón de mortalidad:

$$p(a) = c(a) \frac{e^{-ra}}{b} \quad (9)$$

Por otro lado, la forma en que se relacionan las defunciones de una población estable ($D(a)$) con las de la tabla de vida, se expresa mediante la siguiente forma:

$$D(a) = N(a) \mu(a) = N(0) e^{-ra} p(a) \mu(a) \quad (10)$$

o bien:

$$D(a) = N(0) e^{-rt} d(a) \quad (10')$$

donde $d(a)$ representa las defunciones a la edad a en la tabla de vida en el tiempo t (con radix igual a 1). Entonces, la distribución de frecuencias a la edad a la muerte es:

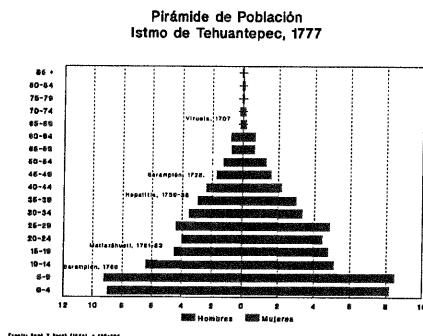
$$\delta(a) = \frac{D(a)}{\int_0^w D(a) da} = \frac{d(a) e^{-ra}}{\int_0^w d(a) e^{-ra} da} \quad (11)$$

Aplicación del modelo

Para la construcción de la tabla se aplicaron dos criterios: el primero con base en la estructura de la población por edad reportada en el censo de 1777; el segundo en la estructura de las defunciones por edad, reconstruida a partir de la información de archivos parroquiales entre 1764 y 1810.

Desarrollo a partir de la estructura de la población: La estructura por edad de la población corresponde a la del censo de 1777 (Cook y Borah, 1974) para la región del Istmo de Tehuantepec (región I). Dicha estructura se encuentra afectada por una deficiente calidad, característica de la información histórica, además de las crisis demográficas que diezmaron la población en la segunda mitad del siglo XVIII (véase gráfica 2).

Una constante en el registro Gráfica 2 de las poblaciones, es la omisión de efectivos entre las edades 0 y 4 años, especialmente en el pasado*. Con el fin de corregir, en la medida de lo posible, el problema señalado, se llevó a cabo una revisión del cociente entre las poblaciones 0-4 y 5-9, en las regiones que manejan Cook y Borah para

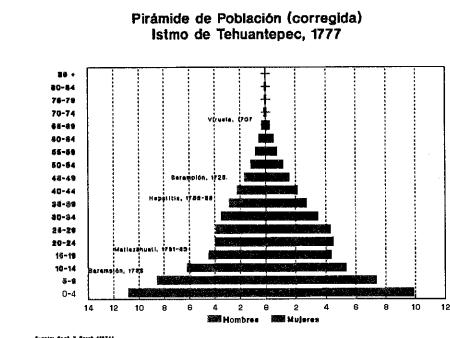
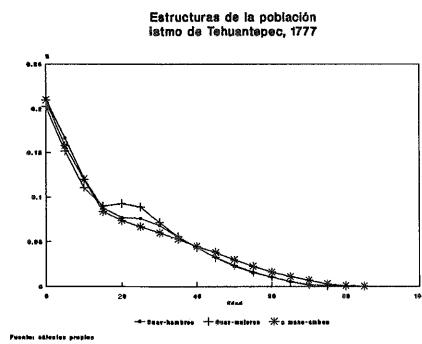


* Se debe esperar que la estructura de la población muestre una pirámide de base amplia, resultado de una alta fecundidad prevaleciente en las sociedades pasadas.

el resto de Oaxaca, observándose una mayor omisión en el caso de la región I. A partir de ello, se corrigió el dato suponiendo un grado de subcobertura del 25%.

En busca de contar con una Gráfica 3 estructura de población que no representara los efectos de las crisis y la mala calidad, se procedió a suavizarla por medio de la técnica llamada de 1/16 (medias móviles de rango 3) con lo que se llegó a una nueva estructura que se muestra en la gráfica 3.

Gráfica 4



El resultado no fue del todo satisfactorio, por lo que se realizó un nuevo ajuste agrupando ambos sexos, esta vez mediante el método gráfico, que se ilustra en la gráfica 4.

Una vez suavizada la estructura de la población, se estimaron los componentes de la

tabla, mediante la ecuación 9. Para ello, se obtuvo la estructura de la población a edad exacta, $c(a)$:

$$c(a) = \frac{C_{a,a+n} + C_{a-n,a}}{2}$$

y aplicando la ecuación 9:

$$p(a) = \frac{c(a) e^{-ra}}{b}$$

se obtiene $p(a)$ (probabilidad de sobrevivencia a la edad a); mientras que las tasas de crecimiento (r) y natalidad (b), necesarias para el cálculo de las probabilidades de sobrevivencia, resultaron del siguiente procedimiento:

Con la información referente a bautizos de Tehuantepec (Gutiérrez, 1992) se estimó el crecimiento de la natalidad entre 1765 y 1794, agrupando la información en lustros: tasa mínima, 0.4%; promedio, 0.63% y máxima 0.85%, como se ve en el cuadro 1. Al incorporar el registro de defunciones, se estimó el crecimiento natural, sin encontrar evidencias de estabilidad debido a la presencia de epidemias, aún así se procedió a realizar el ejercicio. Suponiendo una población cerrada, el ritmo de crecimiento varió desde -0.71% a 0.83%. Se experimentó con tasas de -1%, 0% y 1%, la última de las cuales fue desechada.

La tasa de natalidad asociada a la tabla de mortalidad se obtuvo mediante la relación entre bautismos y volumen de población; con los ritmos de crecimiento señalados se obtuvieron resultados que variaban de 71.6 o/oo a 75.7 o/oo. Debido a que éstos no satisfacieron la construcción de la tabla, se obtuvo como el promedio de la tasa de natalidad del ejercicio basado en la estructura de las defunciones, y los valores presentados anteriormente. El resultado fue de 60o/oo.

Cuadro 1
Estimación de las tasas de crecimiento, Tehuantepec (1765-1794)

Año	Tasas de Crecimiento (%)											
	Eventos			TCB	Tasa de Crec. Nat. (TCN)				Tasa Bruta de Natalidad (o/oo)			
	Baut	Def	SN		r 0.0%	r 0.4%	r 0.63%	r 0.85%	r 0.0%	r 0.4%	r 0.63%	r 0.85%
65-69	933	1,257	-324	0.79	-0.30	-0.33	-0.35	-0.37	42.9	47.4	50.2	53.0
70-74	1,215	799	416	0.78	0.38	0.41	0.43	0.45	55.9	60.6	63.4	66.2
75-79	1,567	969	598	0.40	0.55	0.58	0.60	0.62	72.1	76.6	79.2	81.8
80-84	1,584	2,289	-705	0.46	-0.65	-0.68	-0.69	-0.71	72.9	75.9	77.6	76.3
85-89	1,616	1,258	358	0.85	0.33	0.34	0.34	0.34	74.4	75.9	76.7	77.6
90-94	2,314	1,407	907		0.83	0.83	0.83	0.83	106.5	106.5	106.5	106.5
MG				0.63					68.1	71.6	73.7	75.7

Población estimada a mitad de cada período										
Población en 1792	Períodos									
	65-69	70-74	75-79	80-84	85-89	90-94				
21,731										
Población estimada										
r=0.4%	19,667	20,063	20,468	20,881	21,302	21,731				
r=0.63%	18,591	19,180	19,788	20,416	21,063	21,731				
r=0.85%	17,607	18,363	19,153	19,976	20,835	21,731				

Fuente: Censo de Revillagigedo, 1792-93, AGN-GEN y cálculos propios.

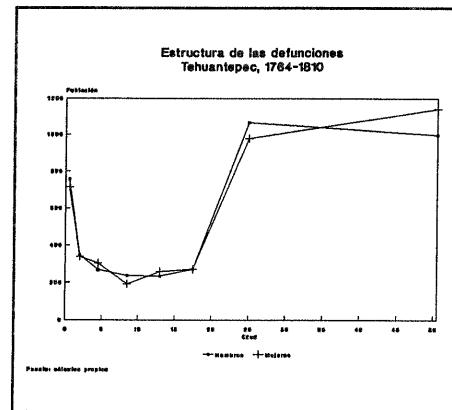
Desarrollo a partir de la estructura de las defunciones: La información de defunciones para Tehuantepec presenta la edad de los difuntos, especialmente en menores. Los grupos de edad se organizaron con el fin de obtener la menor inconsistencia entre las edades que estaban especificadas y las que no, con base en el criterio del estado civil, o bien anotaciones diversas.

A partir de la estructura de las defunciones observada, se distribuyeron los datos no especificados, con la ventaja de que se contaba con la diferenciación entre párvulos (menores de 10 años de edad) y adultos. El mayor problema se presentó en adultos, ya que no se especificaba regularmente la edad; para resolver este problema se aplicó un factor proporcional al intervalo (gráfica 5).

Con tasas de crecimiento iguales a 0% y a -1% se realizó el cálculo de las defunciones asociadas a la tabla de vida, mediante la siguiente ecuación, derivada de la 11.

$$d_x^n = N(0) \frac{\delta_x^n e^{r\bar{x}}}{\sum_0^{w-n} \delta_y^n e^{r\bar{y}}}$$

Posteriormente se estimaron Gráfica 5 los diferentes componentes de la tabla de vida, mediante la aplicación de los métodos usuales, con lo que se obtuvo una tabla abreviada con las edades tal y como fueron agrupadas. Con base en la función de sobrevivencia ($N(x)$) obtenida, se realizó un ejercicio de interpolación, mediante la aplicación del Spline Cúbico, obteniendo así la $N(x)$ para cada edad exacta, con un intervalo de 5 años. Finalmente, se construyó la tabla de vida por grupos quinquenales de edad.



Resultados

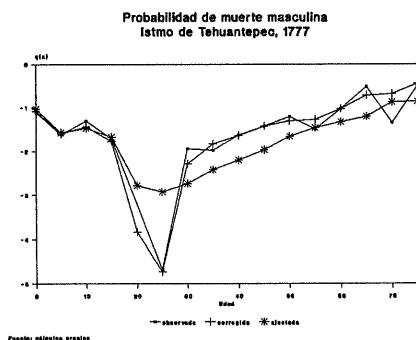
La construcción de la tabla por sexo, a partir de la estructura de la población (criterio a), no resultó satisfactoria ya que el comportamiento de las probabilidades de muerte (q_x), se aleja de

las estructuras observadas en las poblaciones del pasado, como se muestra en las gráficas 6 y 7.

La corrección de la estructura por el método gráfico, permite observar un comportamiento más lógico, a pesar de que se mantienen algunos problemas en los primeros grupos de edad, fuera de lo que se podría esperar.

A partir de la estructura de las defunciones, los resultados

Gráfica 6

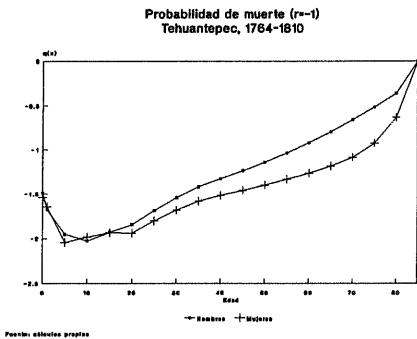


fueron más consistentes, pues se observa una mortalidad infantil alta, que desciende para las edades de 5 a 10 años y se incrementa en las posteriores, con una mortalidad masculina superior a la femenina en los adultos. Véanse gráficas 8 y 9.

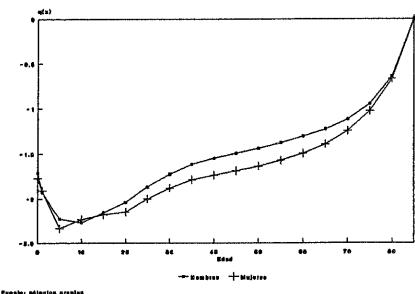
De los elementos de la tabla de vida, resaltan algunos indicadores, tales como la mortalidad infantil (${}_0q_1$), la esperanza de vida al nacimiento (e_0), la estructura de población, asociada a la tabla, vista a través del porcentaje de menores de 15 años, la tasa bruta de natalidad y la tasa de crecimiento.

Al analizar las estimaciones Gráfica 8 obtenidas, bajo ambos criterios, se observan algunos elementos que deben ser anotados. El primero de ellos es la similitud en los niveles de mortalidad infantil (α_5) que se presenta en ambos casos; las esperanzas de vida sí presentan diferencias notables, sobre

Gráfica 9



Probabilidad de muerte (r=0)
Tehuantepec, 1764-1810



todo al considerar que la mortalidad infantil mantiene un comportamiento semejante. Junto a lo anterior, cabe recordar que la Tasa Bruta de Natalidad, para el primer criterio, se fijó en 60 o/oo, ya que cualquier variación en ella ocasionaba cambios significativos en los resultados. Finalmente, es relevante, la variación que se presenta al comparar los porcentajes de efectivos menores de 15 años de edad. Véase cuadro 2.

Debido a las diferencias entre los dos criterios, se llevó a cabo una comparación con los parámetros de la tabla modelo "oeste" (Coale y Demeny, 1982) a fin de identificar la coherencia de los resultados; se observó una mayor consistencia en las estimaciones bajo el criterio b, ya que, según la tabla modelo, a un nivel de

alta fecundidad (TBR=4) y una esperanza de vida entre 20 y 30 años corresponden un porcentaje de población menor a 15 años de entre el 42% y el 46.4% y una tasa de natalidad de 61.3 o/oo a 58.1 o/oo.

Cuadro 2
Estimaciones de los parámetros más importantes de la tabla de vida
Tehuantepec, 1765-1794

a) Estimación a partir de la estructura de la población

TBN=60 o/oo	r=1	q_{0-5}	e_0	Pob<15 años (%)
Hombres				
Observado	0.329		21.37	50.3
Corregido	0.338		21.37	47.8
Mujeres				
Observado	0.357		21.54	46.6
Corregido	0.379		21.54	45.9
Ajustada gráficamente (ambos sexos)	0.355		21.73	48.8

b) Estimación a partir de la estructura de las defunciones

	q_{0-5}	q_{0-1}	e_0	Pob<15 años (%)
r=0%				
Hombres	0.300	0.181	24.09	42.2
Mujeres	0.294	0.170	25.92	39.8
B = 41.51 o/oo				
r=-1%				
Hombres	0.371	0.226	18.99	48.1
Mujeres	0.366	0.215	20.36	45.4
B = 52.66 o/oo				

Fuente: Cálculos propios

Si bien se pueden observar diferencias, sobre todo en los niveles de los distintos indicadores, se puede decir que existe una selectividad de la mortalidad por sexo, al igual que en las sociedades actuales.

Otra característica es la alta mortalidad infantil ya que la presencia de epidemias en el período estudiado afectaba especialmente a los infantes. Por ello, es más lógico el resultado obtenido con una tasa de crecimiento negativo, que presenta tasas de mortalidad infantil más elevadas, que podían llegar a diezmar el volumen de población, dejando secuelas difíciles de recuperar, aun y cuando hubiera un crecimiento en el nivel de la natalidad.

Si bien se puede señalar que las tasas de natalidad, asociadas a la tabla, pueden parecer bajas, esto se explica en virtud del

efecto que tiene en la fecundidad la presencia de epidemias, ya que, como se ha demostrado en diversos trabajos (Malvido, 1973, Carreño, 1978, Cooper, 1965, Gutiérrez, 1992), una crisis demográfica se compone de la existencia de altos niveles de mortalidad, seguidos de un descenso en los niveles de fecundidad.

A manera de conclusión.

Es difícil realizar estudios sobre las poblaciones del pasado desde una perspectiva de análisis estadístico, sobre todo debido a la mala calidad de la información con que se cuenta. Sin embargo, el desarrollo de la Demografía ha posibilitado la aparición de modelos probabilísticos, que permitirán avanzar en ese sentido.

Regularmente se parte del supuesto de que las poblaciones del pasado mantenían un comportamiento estable en su crecimiento, sin considerar la presencia de fenómenos perturbadores tales como las crisis demográficas. En el presente trabajo, se ha logrado identificar cuan sensible resulta la aplicación de técnicas estadísticas a fuertes variaciones en el nivel de la mortalidad.

Al aplicar el modelo de poblaciones estables en el caso de Tehuantepec durante el siglo XVIII, se ha avanzado en el conocimiento de algunos indicadores de las condiciones de vida de aquella época. Es muy probable que la esperanza de vida del conjunto de habitantes no excediera los 26 años, agobiados como estaban por la presencia de severas epidemias.

Cabe señalar que los resultados obtenidos, fueron fruto de ajustes hechos a la información, debido a la mala calidad de ella, pero sin llegar a modificaciones significativas. Por otro lado, una vez obtenida la estructura de la mortalidad, no fue ajustada mediante modelos aplicables a poblaciones actuales.

Referencias

- Camargo, L. y Partida, V (1992), "Algunos aspectos demográficos de cuatro poblaciones prehispánicas de México" en: **El Poblamiento de las Américas**, Vol 1, IUSSP, Veracruz.
- Carreño, G. "Mortalidad en el Obispado de Michoacán a Consecuencia de las Crisis Económicas de 1785-1786", en **Anuario**, No. 3, (Escuela de Historia de la Universidad de Michoacán), Morelia, pp. 187-197.
- Coale, A. (1972), **The Growth and Structure of Human Populations. A Mathematical investigation**, Princeton University Press, California.
- Coale, A., Demeny P y Vaughan B (1982), **Regional Model Life Tables and Stable Populations**, Second edition, Academic Press, N.Y.
- Cook, S.F. y Borah, W. (1978), **Ensayos sobre Historia de la Población: México**, Siglo XXI, México.
- Cooper, D. (1965) **Epidemic Diseases in Mexico City, 1761-1813, an Administrative, Social and Medical Study**, PhD Tesis, University of California Press, Berkeley.
- Gutiérrez, R. (1992), **Crisis Demográficas en el Istmo de Tehuantepec, durante la Segunda Mitad del Siglo XVIII**, Tesis en proceso para obtener el grado de Maestro en Demografía, El Colegio de México.
- Lotka, A. (1973) "La estabilidad de la Distribución Normal por edades" en Lotka, **Demografía Matemática. Selección de Artículos**, CELADE, Santiago de Chile.
- Malvido, E. (1973). "Factores de Despoblación y de Reposición de la Población de Cholula (1614-1810)" en **Hiatoria Mexicana**, XXIII, 89(1), pp. 52-110.
- Naciones Unidas (1970), **El concepto de la población estable. Aplicación al estudio de la población de países que no tienen buenas estadísticas demográficas**, N.Y. (ST/SOA/Serie A/39).
- Preston, S. y Coale, A. (1982), "Age structure, Growth, Attrition, and Accession: a new synthesis", en: **Population Index**, Vol 48 Num. 2, pp. 217-259.

EXTREMOS BIVARIADOS (VERSION PROCESOS PUNTUALES)

Humberto Gutiérrez Pulido¹

RESUMEN

A raíz de la concientización sobre la problemática ecológica mundial se ha intensificado el estudio de la teoría de valores extremos. La versión clásica de esta teoría se identifica con las distribuciones límite de valor extremo (distribuciones de Frechet, Weibull y Gumbel). En los últimos años se ha ido conformando una nueva versión de los valores extremos, que se fundamenta en la teoría de procesos puntuales. En particular el caso bivariado se inicia con el trabajo de Joe, Smith y Weissman(1989).

En este trabajo se describe la teoría bivariada de valores extremos versión procesos puntuales y se señala el potencial que tiene para el análisis de datos ambientales.

INTRODUCCION

La humanidad siempre ha tenido que coexistir con la manifestación extrema de fenómenos naturales, como lo son grandes lluvias, sequías extremas, vientos máximos, olas gigantes, terremotos fuertes, etcétera. En tiempos recientes se han agregado otros fenómenos que preocupa su ocurrencia extrema (máxima o mínima). La teoría de valores extremos estudia la ocurrencia de

¹Facultad de Ingeniería, Universidad de Guadalajara.
Campus Tecnológico, Avenida Revolución, Guadalajara, Jal.

tales tipos de fenómenos desde el punto de vista de probabilidad y estadística.

A la fecha se han desarrollado dos versiones de la teoría de valores extremos. La primera, llamada versión clásica, está basada principalmente en las distribuciones conocidas como de valor extremo (distribuciones Gumbel, Weibull y Frechet). La segunda, llamada versión umbral, se fundamenta en los procesos estocásticos puntuales y en la distribución de los valores que exceden un punto dado (llamado umbral).

La versión umbral de la teoría de extremos multivariados es una aportación reciente de Joe, et al(1989). La cual consiste en ajustar un modelo de procesos puntuales a la distribución conjunta de observaciones bivariadas que exceden algún punto bivariado alejado (umbral). Esta versión tiene algunas ventajas sobre la metodología clásica multivariada, entre las que podemos mencionar:

1.- La metodología clásica tiene el problema de que no hay un orden natural en R^n (el orden con el que se trabaja es tomando los máximos por componente, ver Gutiérrez, 1991). Mientras que el método umbral no tiene tal problema.

2.- La inferencia con la metodología clásica tiene algunas dificultades, por ejemplo falta de regularidad en los estimadores (ver Twan, 1988). La inferencia en la metodología umbral está basada en dos áreas más desarrolladas, como son los procesos de Poisson no homogéneo y la teoría univariada de valores extremos (en sus dos versiones).

3.- Cuando se ajusta un modelo con la metodología clásica,

por lo general, se toma únicamente información de los máximos. En la metodología umbral el proceso puntual se construye con mayor información y no sólo con los máximos, ver Gutiérrez(1991).

LOS EXTREMOS COMO PROCESOS PUNTUALES

El resultado fundamental de la teoría de valores extremos es el siguiente. Sea X_1, X_2, \dots una sucesión de variables aleatorias independientes e idénticamente distribuidas, con función de distribución $F(\cdot)$. Definamos $M_n = \max\{X_1, X_2, \dots, X_n\}$ y Supóngase que existen constantes $a_n > 0, b_n$ tales que

$$\lim_{n \rightarrow \infty} P\left\{ (M_n - b_n) / a_n < z \right\} = F^n(a_n z + b_n) \xrightarrow{D} H(\cdot),$$

entonces la función $H(\cdot)$ es la distribución de valor extremo generalizada:

$$(1) \quad H(x; \mu, \sigma, k) = \exp \left[- \left\{ 1 - k[(x-\mu)/\sigma] \right\}^{1/k} \right].$$

donde x tiene un rango definido por la restricción $\{1-k(x-\mu)/\sigma\} > 0$.

La distribución de valor extremo generalizada toma tres formas particulares, dependiendo del valor de k : distribuciones Weibull Frechet y Gumbel. Así estas son las tres únicas distribuciones límites para extremos (máximos o mínimos).

La relación entre la teoría clásica de valores extremos y los procesos puntuales es dada por el siguiente resultado, ver

Pickands (1971). Sean los puntos $Y_i = (X_i - b_n)/a_n$, $i=1,2,\dots,n$, con P_n se denota el proceso puntual sobre la línea real cuyos puntos son Y_i , entonces el proceso puntual P_n converge débilmente a un proceso de Poisson no homogéneo, cuya medida de intensidad es:

$$(2) \quad \Lambda\{(x, \infty)\} = \left\{1 - k[(x-\mu)/\sigma]\right\}^{1/k},$$

para $1 - k(x-\mu)/\sigma > 0$.

Obsérvese que para definir el proceso puntual P_n se utilizó las constantes normalizadoras $a_n > 0$, b_n de la teoría clásica de valores extremos. Algo destacable es la relación entre la distribución de valor extremo generalizada y la medida de intensidad del proceso límite de Poisson no homogéneo (ver expresiones (1) y (2)).

Para el caso bivariado, sean (X_{11}, X_{21}) , $i=1,2,\dots$ pares de variables aleatorias independientes e idénticamente distribuidas. Supóngase que las variables X_{j1} han sido transformadas a variables Z_{j1} de forma tal que

$$(3) \quad P(Z_{j1} > z) \approx 1/z,$$

cuando $z \rightarrow \infty$, $j=1,2;^2$. Luego si con P_n se denota el proceso

²Se puede ver que una transformación que logra esto es

$$Z_{j1} = 1/\log[n/(R_{j1}-0.5)],$$

donde R_{j1} es el rango de X_{j1} entre $x_{j1}, x_{j2}, \dots, x_{jn}; j=1,2,$

puntual en \mathbb{R}^2 formado por $\left\{ n^{-1}(z_{11}, z_{21}), \dots, n^{-1}(z_{1n}, z_{2n}) \right\}$, entonces P_n converge débilmente a un proceso de Poisson no homogéneo en $[0, \infty) \times [0, \infty) - \{(0,0)\}$, cuya medida de intensidad Λ satisface la siguiente relación

$$(4) \quad \Lambda(B/m) = m\Lambda(B),$$

para todos los conjuntos B Borel medibles que están a distancia positiva de $(0,0)$, y para todo $m > 0$, (ver Resnick, 1987, cap. V).

De esta manera, el conjunto de pares $n^{-1}(z_{11}, z_{21})$ asintóticamente forma un proceso de Poisson no homogéneo en \mathbb{R}^2 , con lo que los extremos bivariados pueden tratarse como tal proceso puntual.

En adelante se fija la atención en la medida de intensidad del proceso Poisson no homogéneo en el conjunto $[(0, z_1) \times (0, z_2)]^c$.

Definamos la medida de intensidad en tal conjunto como sigue

$$(5) \quad \mu(z_1, z_2) = \Lambda\left\{[(0, z_1) \times (0, z_2)]^c\right\}.$$

Por otra parte, la función $\Lambda(B)$ es una medida de intensidad para todo conjunto B que esté alejado de $(0,0)$ y la relación (4) implica que la medida de intensidad puede representarse en forma polar:

$$\Lambda\left\{(dr, dw)\right\} = \frac{dr}{r^2} H(dw),$$

donde la variable ángulo es w . La función $H(dw)$ es una medida en $[0, 1]$. Varias elecciones de r y w son posibles, por ejemplo $r = z_1 + z_2$ y $w = z_1/r$. La relación entre la medida de intensidad μ y la función H está dada por la siguiente expresión ordenando en forma creciente.

$$(6) \quad \begin{aligned} \mu(z_1, z_2) &= \int_0^1 \max\left[wz_1^{-1}, (1-w)z_2^{-1}\right] H(dw) \\ &= (z_1^{-1} + z_2^{-1}) M\left(z_1/(z_1+z_2)\right), \end{aligned}$$

donde

$$M(u) = \int_0^1 \max\left[w(1-u), (1-w)u\right] H(dw) \quad (0 \leq u \leq 1).$$

La función $M(\cdot)$ es convexa en $[0,1]$, acotada superiormente por 1 e inferiormente por $\max\{u, 1-u\}$.

De la expresión (6) se desprende que para modelar la medida de intensidad del proceso Poisson no homogéneo basta modelar la función $M(\cdot)$. Modelos particulares para $M(\cdot)$ son, ver Gutiérrez(1991), el modelo logístico:

$$M(u ; \alpha) = \left(u^{1/\alpha} + (1-u)^{1/\alpha}\right)^{\alpha} \quad \text{para } 0 < \alpha < 1.$$

Con lo que la medida de intensidad del proceso de Poisson no homogéneo es

$$(7) \quad \mu(z_1, z_2; \alpha) = \left(z_1^{-1/\alpha} + z_2^{-1/\alpha}\right)^{\alpha}.$$

Este modelo es simétrico en el sentido de intercambiabilidad, que en ciertos casos reales es una limitante. Otro modelo y que no es simétrico es el biológico:

$$M(u ; \alpha, \beta) = (1-u)v^{1-\alpha} + u(1-v)^{1-\beta} \quad (0 < \alpha, \beta < 1),$$

donde $v=v(u ; \alpha, \beta)$ es la raíz de $(1-\alpha)(1-u)(1-v)^{\beta} - (1-\beta)uv^{\alpha}=0$.

De esta manera la medida de intensidad del proceso formado por los puntos extremos es

$$(8) \quad \mu(z_1, z_2; \alpha, \beta) = z_1^{-1}v^{1-\alpha} + \left(z_1/(z_1+z_2)\right)(1-v)^{1-\beta}$$

De esta manera se tiene completamente modelado el comportamiento de los extremos bivariados, ya que sabemos que asintóticamente se comportan como un proceso de Poisson no

homogéneo con medida de intensidad para la que tenemos dos medelos: las expresiones (7) y (8).

INDEPENDENCIA A PARTIR DE UMBRALES

Los resultados teóricos descritos anteriormente son de utilidad práctica cuando hay dependencia entre los extremos de las dos variables. Se sabe que las variables (X_1, X_2) sean dependientes no implica que los extremos lo sean, por ejemplo si este vector sigue una distribución normal bivariada con $|\rho| \neq 1$, entonces los extremos de ambas variables son independientes. Así, antes de usar la teoría de extremos versión umbral es necesario asegurarse de que haya dependencia entre los extremos (ver Gutiérrez, 1991).

ESTIMACION DE LA MEDIDA DE INTENSIDAD

La estimación de los parámetros de la medida de intensidad para el proceso de Poisson no homogéneo, se hará por el método de máxima verosimilitud.

Sea $X_i = (X_{1i}, X_{2i})$, $i=1, 2, \dots, n$, los datos originales y sea $Z_i = (Z_{1i}, Z_{2i})$ los datos transformados de tal forma que cumplen con la condición (3). Los X_i son realizaciones independientes e idénticamente distribuidas del vector aleatorio $X = (X_1, X_2)$ y los Z_i son tratados como realizaciones independientes e idénticamente distribuidas de un vector aleatorio Z .

Supóngase que los puntos $n^{-1}Z_i$ caen en un conjunto prefijado $B \subset R^2$, la medida de intensidad del proceso es $\Lambda(\cdot; \theta)$; $\lambda(\cdot, \theta)$ es la función de intensidad y θ es un vector de parámetros. La verosimi-

litud para θ está dada por, ver Daley & Vere-Jones(1988),

$$(9) \quad L(\theta) = \exp[-\Lambda(B; \theta)] \left[\prod_{Z_i/n \in B} \lambda(n^{-1}Z_{1i}, n^{-1}Z_{2i}; \theta) \right].$$

Se puede observar de las expresiones (5), (7) y (8) que ya se tienen modelos definidos para las funciones Λ y λ de la función de verosimilitud.

Utilizando los conocimientos sobre las distribuciones marginales, se procede como sigue. Sean η_1, η_2 los vectores de parámetros marginales, los puntos Z_i se pueden obtener con una transformación del tipo

$$(10) \quad Z_i = \left(t_1(x_{1i}; \eta_1), t_2(x_{2i}; \eta_2) \right), \quad i=1, 2, \dots, n;$$

donde t_1 y t_2 son funciones estrictamente crecientes, que tienen como parámetros a η_1 y η_2 respectivamente. De esta manera la función de verosimilitud completa para θ , η_1 y η_2 está dada por

$$(11) \quad L(\theta, \eta_1, \eta_2) = \exp[-\Lambda(B; \theta)]$$

$$\left\{ \prod_{x_i : Z_i/n \in B} \left[\lambda(n^{-1}Z_{1i}, n^{-1}Z_{2i}; \theta) n^{-1} \frac{\partial t_1}{\partial x_1}(x_{1i}; \eta_1) n^{-1} \frac{\partial t_2}{\partial x_2}(x_{2i}; \eta_2) \right] \right\}.$$

Definiendo las distribuciones marginales de la forma siguiente

$$(12) \quad F_j(x_j; \eta_j) = \begin{cases} F_j(x_j) & \text{si } x_j \leq u_j \\ 1 - (1 - F_j(u_j)) \left(1 - k(x_j - u_j)/\sigma_j \right)_+^{1/k} & \text{si } x_j \geq u_j, \end{cases}$$

donde $y_+ = \max\{0, y\}$.

Las transformaciones t_j en (11) pueden ser, ver Gutiérrez(1991), así

$$(13) \quad z_j = t_j(x_j) = \left(-\log F_j(x_j; \eta_j) \right)^{-1}, \quad x_j = F_j^{-1} \left(\exp(-1/z_j; \eta_j) \right)$$

para $x_j \leq u_j$ y

$$(14) \quad z_j = \left[-\log \left\{ 1 - (1 - F_j(u_j)) \left(1 - k_j(x_j - u_j)/\sigma_j \right)_+^{1/k_j} \right\} \right]^{-1}$$

para $x_j \geq u_j$.

La función de verosimilitud (11) puede ser simplificada como

$$(15) \quad L(\theta, \eta_1, \eta_2) = \exp[-\Lambda(B; \theta)]$$

$$\left\{ \prod_{x_i : z_i/n \in B} \lambda(n^{-1}z_{1i}, n^{-1}z_{2i}; \theta) \right\}$$

$$\left\{ \prod_{x_{1i} : z_{1i}/n \in B, x_{1i} \geq u_{1i}} n^{-1} \frac{\partial t_1}{\partial x_{1i}}(x_{1i}; \eta_1) \right\}$$

$$\left\{ \prod_{x_{2i} : z_{2i}/n \in B, x_{2i} \geq u_{2i}} n^{-1} \frac{\partial t_2}{\partial x_{2i}}(x_{2i}; \eta_2) \right\}.$$

De esta manera para obtener los estimadores de máxima verosimilitud de (θ, η_1, η_2) y estimar la covarianza entre ellos, se maximiza el logaritmo de la expresión (15) usando una rutina de tipo Quasi-Newton, por ejemplo. Los valores iniciales para la maximización pueden ser obtenidos separando las verosimilitudes

univariadas; Pareto generalizada para η_1 , η_2 y la verosimilitud para el parámetro θ está dada por

$$L(\theta) = \exp[-\Lambda(B; \theta)] \left(\prod_{x_i : z_i/n \in B} \lambda(n^{-1}\tilde{z}_{1i}, n^{-1}\tilde{z}_{2i}; \theta) \right),$$

donde \tilde{z}_1 y \tilde{z}_2 son las transformaciones usando los estimadores de máxima verosimilitud $\hat{\eta}_1$ y $\hat{\eta}_2$ de las verosimilitudes univariadas.

También se puede estimar la versimilitud de manera no paramétrica, para ver esto y mayores detalles del presente trabajo ver el capítulo 4 de Gutiérrez(1991).

POSIBLES APLICACIONES

Algunas de las aplicaciones de esta teoría al análisis de datos ambientales pueden ser las siguientes:

a. Estudiar la relación de dependencia entre los niveles extremos de dos contaminantes. El que dos contaminantes estén correlacionados, no implica necesariamente que sus niveles extremos mantengan esa relación.

b. Estudiar la dependencia de los niveles máximos de un contaminante en dos lugares distintos. Por ejemplo, ¿existe dependencia entre los niveles máximos de oxono que se registran en dos estaciones de monitoreo?

c. Estimar la probabilidad de que dos contaminantes excedan simultáneamente ciertos niveles extremos.

d. Estimar la curva cuantil bivariada formada por los puntos (x_1, x_2) tal que $P(X_1 > x_1, X_2 > x_2) = p$, para una probabilidad p pequeña.

REFERENCIAS

- Daley, D.J. and Vere-Jones, D.(1988). An Introduction to the Theory of Point Processes. Springer-Verlang, London.
- Gutiérrez P., H.(1991). Extremos Multivariados: Modelos y Estadística. Tesis de Maestría en Estadística. CIMAT-U. DE Gto., Guanajuato.
- Joe, H., Smith, R.L. & Weissman, I.(1989). Bivariate threshold methods for extremes. Preprint, University of Surrey (Guildford, U.K.).
- Pickands, J. (1971). The two-dimensional Poisson process and extremal process. Journal of Applied Probability 8, 745-56.
- Resnick, S.I.(1987). Extreme Values, Regular Variation, and Point Processes. Springer-Verlang, New York.
- Smith, R.L.(1989). Extreme value analysis of environmental time series: an example based on ozone data. Statistical Science 4, 4, 367-93
- Tawn, J.A.(1988). Bivariate extreme value theory: Models and estimation. Biometrika 75, 397-415.

ILUSTRACION DE LA NO ROBUSTEZ DE LA PRUEBA DE t A
DISTRIBUCIONES CON COLAS LARGAS USANDO DATOS REALES

David Sotres Ramos¹

RESUMEN

El propósito principal del presente trabajo es presentar un conjunto de datos (extraido de un experimento real) en donde la prueba de t para la comparación de medias de dos muestras independientes produce resultados incorrectos. En el trabajo se muestra que dicha falla de la prueba de t se debe a que los datos presentan una distribución con una cola muy larga.

En el trabajo también se presentan dos análisis alternativos que sí proporcionan un resultado correcto.

¹ Centro de Estadística y Cálculo del Colegio de Postgraduados, Montecillo, Texcoco, Estado de México. CP. 56230.

INTRODUCCION

El propósito principal del presente trabajo, es presentar un conjunto de datos (extraído de un experimento real) en donde la prueba de t para la comparación de dos medias produce resultados incorrectos.

El ejemplo nos muestra claramente que la prueba de t no es robusta a distribuciones con colas largas. El caso resulta de interés en virtud de las diversas propiedades óptimas que posee la prueba de t.

También se presentan dos análisis estadísticos alternativos que sí proporcionan un resultado correcto.

La prueba de t y sus propiedades óptimas

Considere el problema de contrastar las hipótesis

$$[H_0 : \mu_1 = \mu_2] \quad vs \quad [H_a : \mu_1 \neq \mu_2]$$

en base a dos muestras aleatorias X_1, X_2, \dots, X_m y Y_1, Y_2, \dots, Y_n de dos poblaciones normales $N(\mu_1, \sigma^2)$ y $N(\mu_2, \sigma^2)$ respectivamente.

Se asume también independencia entre las dos muestras

$$\{X_1, \dots, X_m\} \text{ y } \{Y_1, \dots, Y_n\}.$$

La prueba de t rechaza H_0 con nivel de significancia α
si $|T| > t_{\alpha/2}(m + n - 2)$, en donde

$$T = \frac{\bar{Y} - \bar{X}}{\left[S_p^2 (1/m + 1/n) \right]^{1/2}}$$

\bar{X} y \bar{Y} son las medias muestrales; S_1^2 y S_2^2 son las varianzas muestrales; $S_p^2 = [(m-1)S_1^2 + (n-1)S_2^2] / (m+n-2)$; y donde $t_{\alpha/2}(m+n-2)$ es el percentil de nivel $\alpha/2$ de la distribución t de Student con $m+n-2$ grados de libertad. La prueba de t es insesgada y uniformemente más potente, ver por ejemplo Lehmann (1959).

También la prueba de t puede generarse a partir del llamado modelo de aleatorización, ver por ejemplo Kempthorne (1952).

El Origen de los datos

Los datos provienen de un Estudio Clínico Aleatorizado cuyo objetivo es comparar la eficacia y la tolerancia de dos medicamentos (A y B) en el tratamiento de afecciones traumáticas musculoesqueléticas. El estudio se realizó en la Ciudad de México durante los meses de enero y febrero de 1992. Usando aleatorización se formaron dos grupos de pacientes, el grupo 1 con 29 pacientes que se trató durante 10 días con el medicamento A, y el grupo 2 se trató con el medicamento B también durante 10 días.

Cuadro 1. Duración del Dolor (horas) que reportaron 61 pacientes con afecciones traumáticas musculoesqueléticas al inicio del estudio clínico.

Caso	Grupo	Duración del dolor (horas)	Caso	Grupo	Duración del dolor (horas)
2	1	28.0000	9	2	2.5000
4	1	16.0000	10	2	16.2500
5	1	48.0000	11	2	48.0000
6	1	6.0000	12	2	72.0000
7	1	12.0000	18	2	2.0000
8	1	18.0000	20	2	32.0000
13	1	48.0000	21	2	1.3333
14	1	18.0000	22	2	3.2500
15	1	10.0000	23	2	3.2500
16	1	24.0000	24	2	17.0000
17	1	8.0000	25	2	0.5000
19	1	12.0000	27	2	30.0000
26	1	36.0000	29	2	12.0000
28	1	36.0000	32	2	24.5000
30	1	2.0000	33	2	1.0000
31	1	48.0000	37	2	72.0000
34	1	20.0000	40	2	16.0000
35	1	16.0000	42	2	0.7500
36	1	24.0000	43	2	1.5000
38	1	2.0000	44	2	28.0000
39	1	24.0000	45	2	0.7500
41	1	6.0000	46	2	1.1667
53	1	12.0000	47	2	1.2500
54	1	48.0000	48	2	3.3333
55	1	24.0000	49	2	0.7500
56	1	60.0000	50	2	16.0000
57	1	8.0000	51	2	6.0000
58	1	36.0000	52	2	7.5833
59	1	72.0000	60	2	16.0000
1	2	12.0000	61	2	48.0000
3	2	36.0000			

Para comprobar la comparabilidad de los grupos al inicio del estudio se midieron las variables sexo, edad, diagnóstico y la duración del dolor en horas. Y es precisamente esta variable (duración del dolor) la que genera los datos que motivaron el presente trabajo. Los datos se presentan en el Cuadro 1.

Conviene recordar que la variable duración del dolor se midió con el objeto de verificar la comparabilidad de los dos grupos al inicio del estudio, y no para comparar la eficacia de los medicamentos.

Análisis usando la prueba de t

Al aplicar la prueba de t a los datos en el Cuadro 1 se obtienen los siguientes resultados.

Cuadro 2. Aplicación de la prueba de t a los datos sobre duración del dolor en el Cuadro 1.

Grupo	N	Media	Desviación Estándar	Error Estándar	t	Nivel Sig. (*) observado
1	29	24.9	18.2	3.4	1.675	0.0992
2	32	16.6	20.1	3.5		

(*) El nivel de significancia observado = $P\{|T| > 1.675\} = 0.0992$

Usando el valor estándar de $\alpha = 5\%$ nuestra decisión es no rechazar $H_0 : \mu_1 = \mu_2$. Es decir no hay diferencia significativa entre los grupos con respecto a la duración del dolor al inicio del estudio.

Es interesante señalar que la conclusión de la prueba de t va en contra de nuestra intuición, ya que los intervalos de confianza $\left[\bar{x} \pm \frac{s_1}{\sqrt{m}} \right] = (24.9 \pm 3.4)$ y $\left[\bar{y} \pm \frac{s_2}{\sqrt{m}} \right] = (16.6 \pm 3.5)$ para μ_1 y μ_2 respectivamente no se traslapan, lo que nos lleva a pensar que la hipótesis $H_0 : \mu_1 = \mu_2$ es falsa.

Los siguientes resultados sobre la distribución de los datos en el Cuadro 1 desacreditan la conclusión de la prueba de t.

Usando la prueba de normalidad de Shapiro-Wilk (1965) se rechaza la hipótesis de normalidad de la variable duración del dolor con el nivel de significancia observado de $\hat{\alpha} = 0.0001$ para el grupo 1 y con $\alpha = 0.0207$ para el grupo 2. Este análisis estadístico formal confirma la impresión visual que se obtiene al graficar los histogramas de frecuencias para estos datos que resultan ser distribuciones acentuadamente asimétricas con una larga cola derecha para ambos grupos.

Análisis alternativo 1: Usando la prueba No-paramétrica de Mann-Whitney

La prueba no-paramétrica de Mann-Whitney (ver por ejemplo Lehmann (1975)) no requiere de normalidad de las muestras $\{X_1, \dots, X_m\}$ y $\{Y_1, \dots, Y_n\}$. Al aplicar esta prueba a los datos del Cuadro 1 se rechaza la hipótesis H_0 con un nivel de significancia observado de $\hat{\alpha} = 0.0157$,

Análisis Alternativo 2: Usando la prueba de t con los datos transformados

Usando la transformación $y^{(25)}$ del tipo Box-Cox (1964) a los datos del Cuadro 1 se obtiene que los datos transformados son aproximadamente normales en ambos grupos, de acuerdo a la prueba de Shapiro-Wilk para normalidad. La transformación de Box-Cox está definida por la siguiente ecuación:

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & y^{(\lambda-1)} \text{ si } \lambda \neq 0 \\ . & \\ y \ln(y) & \text{si } \lambda = 0 \end{cases}$$

donde $y^* = \ln^{-1} \{(1/n) \sum \ln(y)\}$.

El método para determinar el valor de λ puede verse por ejemplo en Montgomery (1991).

Al aplicar la prueba de t a los datos transformados se obtiene un nivel de significancia observado de $\hat{\alpha} = 0.0089$. Usando $\alpha = 5\%$ la decisión es rechazar [$H_0 : \mu_1 = \mu_2$], o sea que sí existe una diferencia significativa entre los grupos con respecto a la variable duración del dolor al inicio del estudio.

Conclusión

Es claro que los supuestos de los análisis alternativos 1 y 2 se satisfacen aproximadamente, por lo que es razonable afirmar que la conclusión de estas pruebas es la correcta, o sea que la hipótesis [$H_0 : \mu_1 = \mu_2$] es falsa. Esto quiere decir que el resultado obtenido por la prueba de t en el Cuadro 2 es incorrecto. De lo anterior podemos concluir que la prueba de t no es robusta a distribuciones acentuadamente asimétrica con colas muy grandes.

Referencias

- Box, G.E.P. and Cox, D. R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society, B, Vol. 26, pp 211-243.
- Kempthorne, O. (1952). The Design and Analysis of Experiments. Wiley, New York.
- Lehmann, E. L. (1959). Testing Statistical Hypotheses. John Wiley and Sons, New York.

Lehmann, E. L. (1975). Nonparametrics - Statistical Methods based on Ranks. Holden Day, Inc.

Montgomery, D. C. (1991). Diseño y Análisis de Experimentos. Grupo Editorial Iberoamérica.

Shapiro, S. S., and Wilk, M. B. (1965). An analysis of variance test for normality. Biometrika, 52, 3 y 4, pág. 591.

UN MODELO PARA LA ESTIMACION DE PARAMETROS EN GALLINAS DE POSTURA

ESTEBAN BURGUETE HERNANDEZ*, J. FRANCISCO BURGUETE HERNANDEZ**

Y JOSE G. HERRERA HARO***

RESUMEN

Se realizó una investigación para estimar algunos parámetros de la postura en gallinas reproductoras, tales como el número promedio de huevos por gallina la proporción y el total de huevos fértiles, la proporción y el total de huevos incubados con el fin de realizar inferencia estadística. Los resultados obtenidos permiten proponer estimadores con expresiones cerradas y con distribuciones probabilísticas acordes a las variables involucradas.

INTRODUCCION

En la producción comercial de huevo se requiere de una planeación precisa de las parvadas parentales de gallinas, ya que pequeñas fallas pueden ocasionar desabasto o sobreproducción de huevo, con sus consecuencias en la alta o baja de precios. Es por ello que el productor establece indicadores de eficiencia reproductiva en sus parvadas tales como la proporción de huevos fértiles y la proporción de huevos incubados en un tiempo de postura. Estos indicadores pueden ser poco apropiados, bajo la suposición de normalidad de las variables, debido a que el ciclo de postura de la gallina está normado por un ciclo circadiano que restringe la ovulación a un periodo de tiempo (Etches and Schoch,

* Depto de Matemáticas. UDLAP. Cholula Puebla.

** ITESM-CEM. Atizapan, México.

***Centro de Ganadería. CP. Chapingu, México.

1984) además de que las variables establecen una estructura de dependencia entre ellas.

MODELO PROPUESTO

Considere que las variables aleatorias X, Y, Z , denotan el número total de huevos puestos, fértiles e incubados, respectivamente.

Sean X_1, X_2, \dots, X_n los huevos puestos por un periodo de tiempo por n gallinas reproductoras. $Y_1, Z_1, \dots, Y_n, Z_n$ los correspondientes huevos fértiles e incubados.

El número de huevos puestos por una gallina en un intervalo de tiempo cumple los requerimientos de un proceso de Poisson (Villaseñor, 1983), cuya distribución probabilística es la siguiente:

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots \quad (1)$$

$= 0$ de otra manera,

o equivalentemente $X \sim P(\lambda)$.

Cuando la gallina ha puesto los huevos, la distribución condicionada del número de huevos fértiles dado el número total de huevos puestos ($Y/X=x$) sigue una distribución binomial con x como el número de eventos y probabilidad de éxito (huevo fertilizado) P_F , es decir $Y/X \sim \text{Bin}(x, P_F)$, con

$$g(y/x) = \binom{x}{y} P_F^y Q_F^{x-y}, \quad y=0, 1, \dots, x, \quad (2)$$

$= 0$ de otra manera

donde $Q_F = 1 - P_F$.

Las ecuaciones 1 y 2 pueden combinarse para obtener la distribución conjunta de huevos puestos y fértiles,

$$h(x, y) = \frac{e^{-\lambda} \lambda^x}{x!} \begin{bmatrix} x \\ y \end{bmatrix} P_F^y Q_F^{x-y}, \quad x = 0, 1, 2, \dots$$

$$y = 0, 1, 2, \dots, x. \quad (3)$$

= 0 de otra manera.

De aquí se obtiene la distribución marginal de los huevos fértiles:

$$g(y) = \sum_{x=0}^{\infty} h(x, y) = 0 \text{ si } x < y, \quad y$$

$$g(y) = \frac{e^{-\lambda} P_F^y}{y! Q_F^y} \sum_{x=y}^{\infty} \frac{\lambda^x Q_F^x}{(x-y)!} \quad \text{si } 0 \leq y \leq x$$

$$= \frac{e^{-\lambda} P_F(\lambda P_F)^y}{y!}, \quad y = 0, 1, 2, \dots;$$

es decir, $Y \sim P(\lambda P_F)$.

Si asumimos que $Z/Y \sim \text{Bin}(y, P_E)$, y siguiendo un razonamiento análogo al anterior puede determinarse que $Z \sim P(P_F P_E)$. Note que $P_F P_E$ es la proporción de huevos puestos que terminan la incubación.

OBTENCION DE LOS ESTIMADORES

Los estimadores de máxima verosimilitud (EMV) son invariantes, consistentes, normalmente asintóticos y mejores asintóticamente normales.

Si x_1, \dots, x_n son los valores observados de X , donde $X \sim P(\lambda)$; entonces el EMV de λ es $\hat{\lambda}$:

$$\hat{\lambda} = \frac{\sum x}{n}$$

Utilizando este resultado y la propiedad de invarianza en $Y \sim P(\lambda P_F)$, se obtiene:

$$\hat{\lambda} \hat{P}_F = \frac{\Sigma y}{n},$$

de donde

$$\hat{P}_F = \frac{\Sigma y}{n\hat{\lambda}} = \frac{\Sigma y}{\Sigma x}$$

Asimismo

$$\hat{P}_E \hat{P}_F \hat{\lambda} = \frac{\Sigma z}{n}$$

de donde

$$\hat{P}_E = \frac{\Sigma z}{\Sigma y}.$$

A continuación se procedera a encontrar las varianzas asintóticas de estos estimadores.

Si x_1, \dots, x_n iid $f(x, \theta)$, $\hat{\theta}$ es el EMV de θ y además

$$I(\theta) = -E\left\{\frac{\partial^2}{\partial \theta^2} \log f(x, \theta)\right\},$$

entonces

$$\sqrt{nI(\theta)} (\hat{\theta} - \theta) \xrightarrow{L} N(0, 1),$$

donde $N(0, 1)$ representa una variable aleatoria con distribución normal estándar (Serfling, 1980).

Entonces el intervalo de confianza asintótico es

$$\hat{\theta} \pm Z_{\frac{\alpha}{2}} (nI(\hat{\theta}))^{-1/2},$$

sustituyendo, como es usual, θ por su EMV $\hat{\theta}$.

(a) Promedio de huevos puestos (λ).

Tomando $h(x, y)$ de la ecuación (3), puede obtenerse

$$\frac{\partial^2}{\partial \lambda^2} \log h(x, y) = \frac{-x}{\lambda^2},$$

de donde

$$I(\lambda) = \frac{1}{\lambda},$$

por lo tanto,

$$nI(\hat{\lambda}) = \frac{n^2}{\sum x} .$$

(b) Promedio de huevos fértiles (λP_F)

Usando los resultados anteriores y sabiendo que la distribución marginal de los huevos fértiles es $P(\lambda P_F)$, se tiene:

$$nI(\hat{\lambda} \hat{P}_F) = \frac{n^2}{\sum y}$$

(c) Promedio de huevos incubados ($\lambda P_E P_F$)

En este caso, la distribución marginal de huevos incubados es $P(\lambda P_E P_F)$. Por lo tanto

$$nI(\lambda P_F P_E) = \frac{n^2}{\sum z}$$

(d) Proporción de huevos fértiles (P_F)

Usando la ecuación (3), se obtiene

$$-\frac{\partial^2 \log h(x,y)}{\partial P_F^2} = \frac{-y}{P_F^2} - \frac{x-y}{(1-P_F)^2} ,$$

de donde

$$I(P_F) = \frac{\lambda}{P_F Q_F} .$$

Por lo tanto

$$nI(\hat{P}_F) = \frac{n \hat{\lambda}}{\hat{P}_F \hat{Q}_F}$$

(e) Proporción de huevos incubados (P_E)

El resultado es similar al obtenido en (d), por lo que se tiene

$$nI(\hat{P}_E) = \frac{n \hat{\lambda} \hat{P}_F}{\hat{P}_E (1-\hat{P}_E)} = \frac{n \hat{\lambda} \hat{P}_F}{\hat{P}_E \hat{Q}_E}$$

Utilizando las propiedades asintóticas de los EMV se pueden construir intervalos de confianza.

Ejemplo. Para estimar la eficiencia reproductiva de una parvada de gallinas se tomó una muestra de animales, registrándose las características de postura total (X_i), número de huevos fértiles (Y_i) y números de huevos incubados (Z_i).

HUEVOS NUM. DE GALLINAS	PARVADA 1				
	1	2	3	4	Σ
Puestos (X_i)	4	3	5	6	18
Fértiles (Y_i)	4	2	3	5	14
Eclosionados (Z_i)	4	1	3	4	12

a) El estimador del promedio del número de huevos puestos por gallina (λ_1), es:

$$\hat{\lambda}_1 = \frac{\sum x_i}{n} = \frac{18}{4} = 4.5 \text{ huevos/gallina}$$

Denótese por E.E. el error estándar estimado de un estimador, así

$$\text{E.E. } (\hat{\lambda}_1) = \frac{\sqrt{\sum x_i^2}}{n} = \frac{\sqrt{18}}{4} = 1.06.$$

por lo que su respectivo intervalo de confianza es:

$$\hat{\lambda}_1 \pm (Z_{\alpha/2}) \text{E.E. } (\hat{\lambda}_1) = 4.5 \pm (Z_{\alpha/2}) 1.06$$

b) Proporción de huevos fértiles (P_F)

$$\hat{P}_{F_1} = \frac{\sum Y_i}{\sum X_i} = \frac{14}{18} = 0.77,$$

$$\text{E.E. } (\hat{P}_{F_1}) = \frac{\sqrt{P_{F_1} Q_{F_1}}}{n_1 \lambda_1} = \frac{\sqrt{0.77 \times 0.23}}{18} = -0.0098$$

El intervalo de confianza es:

$$\hat{P}_{F_1} \pm (Z_{\alpha/2}) \text{E.E. } (\hat{P}_{F_1}) = 0.77 \pm (Z_{\alpha/2}) .0098,$$

c) Proporción de huevos eclosionados (P_{E_1}) y su intervalo de confianza al 95% de seguridad

$$\hat{P}_{E_1} = \frac{\sum Z_1}{\sum Y_1} = \frac{12}{14} = 0.857$$

Su intervalo de confianza es:

$$0.857 \pm 1.96 \sqrt{\frac{0.857(0.143)}{4(\frac{18}{4})(\frac{14}{18})}}$$

o equivalentemente

$$[0.763, 0.950]$$

CONCLUSIONES

En este artículo se han obtenido estimadores de la eficiencia reproductiva de aves de postura. Los estimadores aquí obtenidos tienen expresiones cerradas, lo cual es muy favorable ya que de esa manera no se tienen que utilizar métodos no lineales (Burguete y Burguete, 1991).

REFERENCIAS

- BURGUETE, E. y BURGUETE J.F. 1991. Inferencia en un modelo para calcular riesgos de contraer cáncer. [Agrociencia, serie matemáticas aplicadas, estadística y computación, Centro de Estadística y Cálculo, Colegio de Postgrauados, Chapingo, México.
- ETCHES, R.J. AND J.P. SCHOCH. 1984. A mathematical representation of the ovulatory cycle of the domestic hen. British Poultry Sci. 256: 65-76.
- RAO, C.R. 1973. Linear statistical inference and its applications. Second Edition. Wiley Pub. Company. New York.
- RICE, J.A. 1988. Mathematical statistics and data analysis. Wadsworth Inc. California.
- SAS INSTITUTE. 1985. SAS user's guide: basics. SAS Institute Inc. Cary, N.C.
- SERFLING, R.J. 1980. Approximation theorems for mathematical statistics. Wiley. New York
- THOMPSON, J.R. 1989. Empirical model building. John wiley and Sons. New York.
- VILLASEÑOR, J. A. 1983. Procesos estocásticos: una introducción. Colegio de Postgrauados, Chapingo, México.

ESTANDARIZACION NO LINEAL DE LA LONGITUD DE SENALES QUE TIENEN UN MISMO ORIGEN

V.E. Rohen

Resumen

Es común en Reconocimiento de Patrones el estudio de señales como el resultado de ciertos fenómenos naturales; señales cuyas formas contienen la información relevante para su clasificación. Con frecuencia, en estas señales se presentan problemas de tiempo, longitud variable, ruido, etc. que hacen difícil el proceso de discriminación y clasificación. Concentrándonos en el problema de longitud variable de señales que tienen un mismo origen, y representándolas como vectores multidimensionales, nuestro interés es presentar un método de estandarización no lineal de su longitud basado en la proyección de estos vectores sobre el hiperplano que contiene al templado, sin alterar su forma.

Varios son los pasos usados en este proceso: El primero consiste en desarrollar una medida local de similitud entre la señal \mathbf{p} a ser reconocida y un templado o patrón modelo \mathbf{t} . Esto nos lleva a construir una matriz de similitud s , donde $s(i,j)$ es una correlación cruzada entre los puntos i y j de las dos señales \mathbf{p} y \mathbf{t} respectivamente. A partir de la matriz de similitud, se encuentra un patrón de ajuste continuo con la característica de tener una similitud acumulada máxima a lo largo de todas sus componentes. Es decir, este patrón de ajuste, que será una relación no lineal entre la duración del templado y la señal a ser reconocida, ajustará

la longitud de la señal \mathbf{p} a la del templado \mathbf{t} , de tal manera que en cada componente encontremos el mejor parecido posible entre dichas señales. Este último paso requiere la aplicación de un proceso de optimización local que llevará a una solución global óptima.

Las funciones básicas de un sistema de Reconocimiento de Patrones son las de detectar y extraer características comunes de patrones que describen objetos que pertenecen a una misma clase y la de reconocer un patrón en un medio ambiente nuevo y clasificarlo dentro de una de las clases que se están estudiando. Es común en Reconocimiento de Patrones el estudio de señales como el resultado de ciertos fenómentos, como por ejemplo ecg, ecc, sismogramas, etc. donde la forma de la señal es la importante y la que contiene toda la información. Es frecuente que se presenten problemas de tiempo, longitud, ruido, etc. que hacen difícil el proceso de discriminación y clasificación.

Aquí me voy a concentrar en el problema de longitud variable de señales que tienen un mismo origen. Por ejemplo, en el reconocimiento de palabras habladas hay problemas relacionados con diferencias en longitud. El vector de representación de palabras habladas o frases tienen longitud variable de acuerdo a la velocidad de articulación del que habla. Aún la misma persona bajo circunstancias diferentes puede pronunciar una misma palabra de diferentes maneras, y aún cuando los patrones producidos por esta misma palabra tienen forma similar, no son idénticos y algunos métodos de reconocimiento deben usarse para identificar estas señales como iguales. Otro ejemplo puede ser el ballistocardiograma (bcg): señal

¹ Chartes 79, Villa Verdún, 01810 México, D.F.

que refleja la manera en la cual los ventrículos expulsan la sangre a las arterias. Los balistocardiogramas de dos personas con características de salud similares se supone tienen forma similar, pero es frecuente el caso en que aún teniendo un parecido muy grande, su longitud es diferente, o sea que los latidos de sus corazones tienen diferente duración.

La diferencia en longitud de dos señales similares puede causar su clasificación en dos grupos diferentes, cuando de hecho pertenecen a la misma clase. Por lo tanto es necesario ajustar la longitud de las señales sin alterar su forma para hacer factible su comparación.

El propósito de esta plática es presentar un método de estandarización no lineal de la longitud de señales basado en la proyección de sus vectores de representación sobre el hiperplano que contiene al patrón de referencia o templado. Para ésto, primero se construye una medida de similitud entre los vectores de representación de la señal a ser reconocida y el del templado. A partir de esta medida se encontrará un patrón continuo de ajuste con la característica de tener una similitud acumulada máxima a lo largo de todas sus componentes. Este patrón de ajuste será una relación no lineal entre las escalas de medición de los dos patrones que hará la longitud de la señal a ser reconocida igual a la del templado, manteniendo la forma original o con un mínimo de alteración.

Supongamos que los patrones a ser comparados pueden representarse como vectores que pertenecen a diferentes espacios, i.e; $\mathbf{p}=(p_1, \dots, p_{n_p})$ y $\mathbf{t}=(t_1, \dots, t_{n_t})$ donde $n_p \neq n_t$, entonces es necesario desarrollar una medida de qué tanto se parece una porción del patrón \mathbf{p} con una porción particular del

templado \underline{t} . La medida de similitud más obvia es la distancia Euclíadiana que para el caso de vectores en distintos espacios está definida como (Duda et. al., 1973):

$$\epsilon(k) = \{\sum_j (p_{j-k} - t_j)^2\}^{1/2} \quad (1)$$

donde p es el patrón o señal a ser comparada con el templado \underline{t} que representa una clase específica C y donde $\forall k, j$ es tal que $1 \leq j-k \leq n_p$ (estamos suponiendo que $n_p < n_t$). Nótese que cuando k aumenta, p está siendo recorrido a lo largo del templado \underline{t} . Expandiendo ϵ^2 tenemos

$$\epsilon^2(k) = \sum_j \{p_{j-k}^2 - 2p_{j-k}t_j + t_j^2\} \quad (2)$$

$\sum_j p_{j-k}t_j$ puede reconocerse como la correlación cruzada entre p y \underline{t} , entonces ϵ^2 será pequeña cuando $\sum_j p_{j-k}t_j$ sea muy grande, i.e., cuando los patrones p y \underline{t} sean muy parecidos. Nótese que el término $\sum_j p_{j-k}^2$ permanece constante ya que para cualquier k , $1 \leq j-k \leq n_p$. Por otro lado, como $\sum_j t_j^2$ varía con k ya que j depende de k , y que las diferencias en amplitud de las señales afectan el valor de la correlación cruzada, es necesario normalizar por el factor $(\sum_j t_j^2)^{1/2}$:

$$R_{PT}(k) = \sum_j p_{j-k}t_j / (\sum_j t_j^2)^{1/2} \quad (3)$$

Supongamos ahora que los dos patrones a ser comparados son casi idénticos excepto en una pequeña parte donde difieren sustancialmente, i.e.; existen r y s con $r < s$ tales que

$$\begin{aligned} p_i &\approx t_j && \text{para } i-j = \text{constante, y } 1 \leq i \leq r \text{ ó } s \leq i \leq n_p \\ y \quad p_i &\approx t_j && \text{para } r \leq i \leq s. \end{aligned}$$

La comparación propuesta por $R_{PT}(k)$ en (3) sería susceptible de error ya que las partes que más se parecen de las dos señales enmascarían la diferencia sustancial que existe en otro lado. Sería entonces útil comparar solo pequeñas regiones de los dos patrones para lo cual una ventana W de una longitud predeterminada ℓ sería de utilidad para calcular la correlación cruzada localmente; moviendo la ventana gradualmente a lo largo del templado y transladando el patrón p a lo largo de éste (ver figura 1). Se espera entonces que aquellas partes de los vectores con características similares, tengan correlación más alta que otras no tan similares.

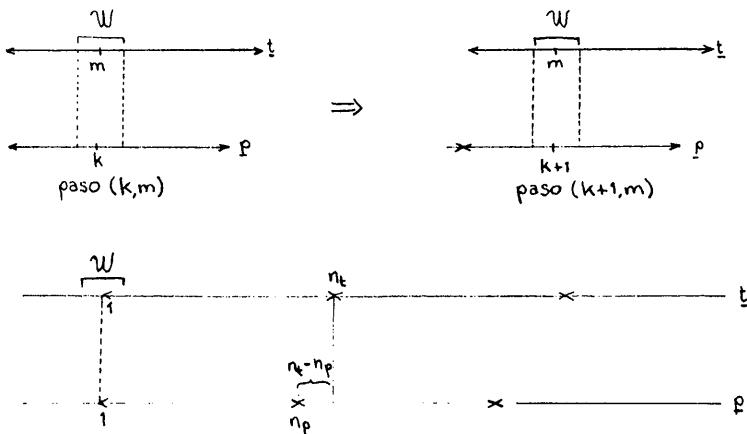


Figura 1

Esto es muy útil en el caso de que la señal que se esté analizando tenga cierta información en una región determinada y otra completamente diferente en otro lado como puede ser el caso del electrocardiograma o del balistocardiograma, donde las primeras oscilaciones en la señal tienen información mucho más relevante que las oscilaciones ocurridas al final

de la señal.

La correlación cruzada local es ahora función de la traslación k del vector \underline{p} y de la posición m de la ventana \mathcal{W} :

$$R_{PT}(k, m) = \sum_j p_{j-k} t_j \mathcal{W}_j / (\sum_j t_j^2 \mathcal{W}_j^2)^{1/2} \quad (4)$$

donde

$$\mathcal{W}_j = \begin{cases} 1 & \text{para } m-l/2 \leq j \leq m+l/2 \\ 0 & \text{en otro caso} \end{cases} \quad (4a)$$

y $l = \sum \mathcal{W}_j$ es la longitud de la ventana con centro en la componente m del templato \underline{t} .

Debemos tomar en cuenta, por un lado que $\sum_j p_{j-k}^2$ ya no es constante, ya que ahora varía con la posición de la ventana y por otro lado la media de las componentes dentro de la ventana varía también con la posición de ésta; entonces $R_{PT}(k)$ tiene la forma

$$R_{PT}(k, m) = \frac{\sum_j (p_{j-k} - \bar{p})(t_j - \bar{t}) \mathcal{W}_j}{(\sum_j (p_{j-k} - \bar{p})^2 \mathcal{W}_j^2)^{1/2} (\sum_j (t_j - \bar{t})^2 \mathcal{W}_j^2)^{1/2}} \quad (5)$$

donde

$$\bar{p} = \sum_{i=1, l} p_i \mathcal{W}_i / \sum_i \mathcal{W}_i \quad \text{y} \quad \bar{t} = \sum_{i=1, l} t_i \mathcal{W}_i / \sum_i \mathcal{W}_i$$

Para cada posición m de la ventana el patrón \underline{p} ha sido trasladado n_p veces (ver figura 1), así pues hemos construido una matriz R de correlación

cruzada local de dimensiones $n_p \times n_t$. Nótese que se está suponiendo que las señales se repiten de una manera continua, siendo así continuo el corrimiento de la ventana.

Una vez obtenida la matriz de similitud entre las dos señales, procedemos a aplicar una función de decisión local que nos ayude a estirar o contraer el patrón p de tal manera que tenga la misma longitud que el templado. Por supuesto nos interesa comparar las regiones que contienen la misma información y se espera que si hay gran parecido entre las dos señales, sea precisamente en regiones equivalentes.

Podemos describir el algoritmo de la siguiente manera: Estando en la entrada (k,m) que suponemos tiene un valor cercano a 1.0, nos fijamos en las tres entradas adyacentes a ésta, en la siguiente columna: $(k-1,m+1)$, $(k,m+1)$, $(k+1,m+1)$ y tomamos aquella entrada con mayor correlación y se guarda la posición en un arreglo p :

$p_i(m)=k$, donde k es tal que

$$R(k,m+1) = \max \{R(k+1,m+1), R(k,m+1), R(k-1,m+1)\} \quad (6)$$

Si proyectamos los valores de R en un plano cuyos ejes coordenados representen la posición de la ventana y la traslación del patrón p respectivamente obtendremos un mapa de contornos de las correlaciones cruzadas locales de los dos patrones (ver figura 2). Si sobreponemos a este mapa las curvas determinadas por los arreglos p_i que encontramos con valores máximos de correlación cruzada, podemos ver que en general están sobre los niveles mas altos del mapa de contornos. Y es de suponerse que solo uno de estos arreglos es óptimo en el sentido de que encuentra los mejores parecidos por regiones. La manera de encontrar tal

arreglo es la siguiente: Moviendo el patrón \mathbf{p} iterativamente a lo largo del templado \mathbf{t} es equivalente a tener las dos señales repetidas infinitamente como procesos periódicos; entonces el mapa de contornos de las correlaciones cruzadas locales se repite bajo la traslación $(m,k) \rightarrow (m+n_t, k+n_t - n_p)$ como se muestra en la figura 3. R también se repite similarmente, y para efectos de cómputo y para tener una visión más clara del comportamiento del arreglo óptimo que estamos buscando, la repetición se hará en la dirección de k . Por otro lado, como el arreglo se encuentra considerando entradas adyacentes de acuerdo al algoritmo (6), el arreglo está delimitado por el paralelogramo formado por las rectas:

$$m=k, \quad m=-k, \quad k=2n_t-n_p-m-1 \quad \text{y} \quad k=-n_p+m+1. \quad (\text{Ver figura 4})$$

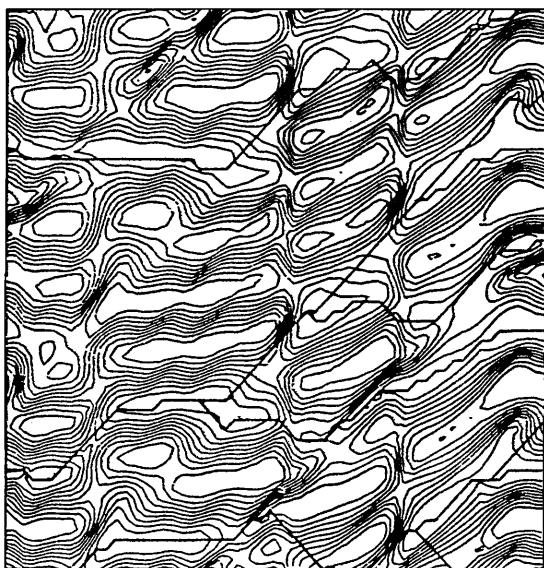


Figura 2

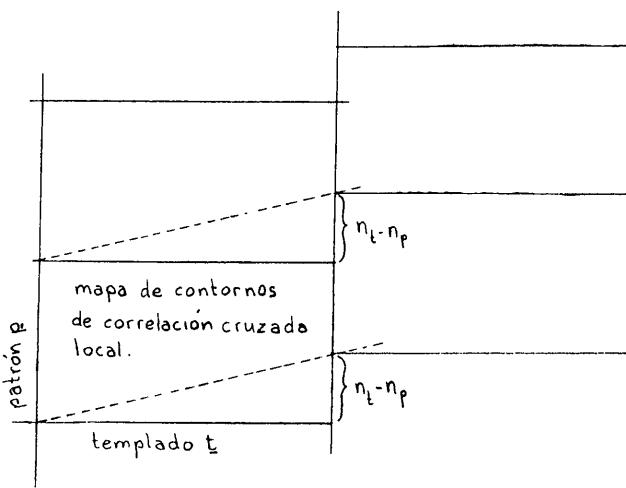


Figura 3

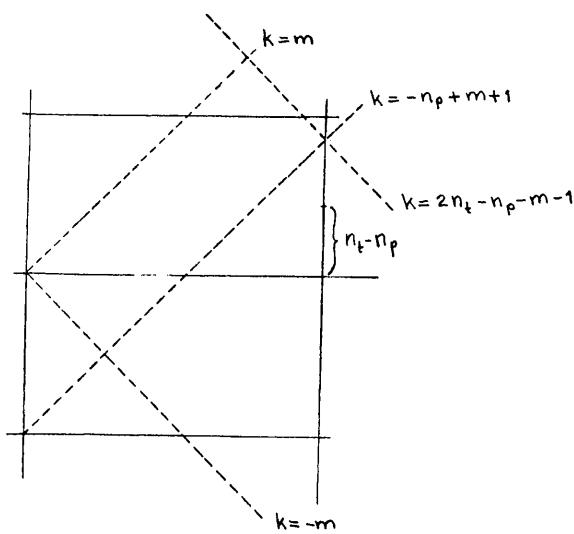


Figura 4

Ahora bien, para asegurar que el arreglo contendrá valores altos de correlaciones cruzadas, es de utilidad calcular la matriz R^* de dimensiones $(2n_p \times n_t)$ de correlaciones cruzadas acumuladas de acuerdo al algoritmo:

$$R^*(k,m) = R(k,m) + \max \{R(k-1,m+1), R(k,m+1), R(k+1,m+1)\}$$

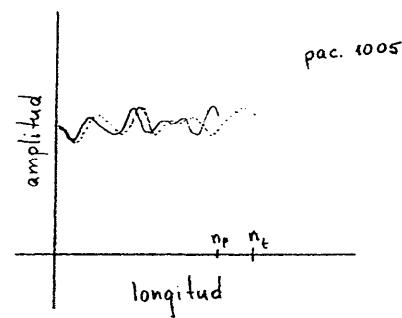
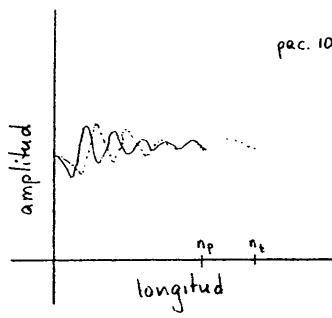
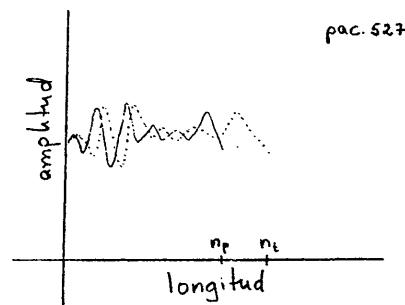
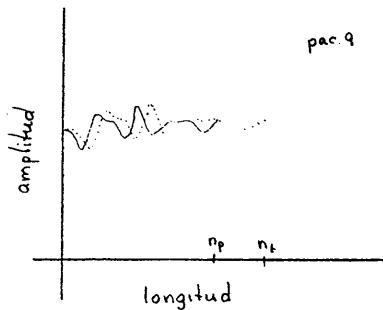
donde $-n_p+1 \leq k \leq n_p-2$ y $0 \leq m \leq n_t-2$.

Para calcular el arreglo óptimo procedemos similarmente a lo hicimos con anterioridad. La expresión recursiva para este proceso está dada por

$p(m) = k$ donde k toma cualquiera de los valores $k-1$, k , ó $k+1$, dependiendo de cual $R^*(k-1,m)$, $R^*(k,m)$, ó $R^*(k+1,m)$ es máximo, y donde k es tal que $p(m-1)=k$.

$$p(0) = 0 \text{ and } p(n_t-1) = n_t - n_p.$$

Para visualizar el significado del arreglo óptimo referimos la curva descrita por éste a la recta de identidad, entonces para cada posición m de la ventana en el templado, $m-p(m)$ dará el punto muestral del patrón p , donde la ventana fué centrada y donde se encontró correlación cruzada local máxima. Este método de estandarización de longitudes fué aplicado a una serie de balistocardiogramas de diferentes personas cuyas longitudes variaban ampliamente. La figura 5 muestra varias de estas señales antes y después de la estandarización y puede apreciarse que en todas ellas se mantiene la información contenida en la forma original. Gracias a este método de estandarización de longitud se han podido clasificar correctamente los balistocardiogramas de pacientes con deficiencia cardiaca, o gente sana, etc, en los grupos adecuados y darles un tratamiento que permita disminuir el riesgo de evento cardiaco.



Referencias:

Duda R.O. and Hart P.E, (1973) "Pattern Classification and Scene Analysis",
John Wiley and Sons, N. Y.

CALCULO DE RIESGOS DE CONTRAER CANCER

J. F. Burguete* y E. Burguete.**

Resumen

En el presente trabajo se comparan algunos modelos paramétricos y no paramétricos cuya finalidad es el cálculo de riesgos de contraer cáncer. Un modelo propuesto mejora el ajuste obtenido por el modelo de Hartley-Sielken (Hartley y Sielken(1977)) y además mejora el ajuste de un modelo no paramétrico.

Se propone un nuevo método para inferencia en caracterizaciones del riesgo. Este método puede disminuir en forma dramática el esfuerzo computacional involucrado en el cálculo de intervalos de confianza ó pruebas de hipótesis.

Ademas se realizó un estudio de simulación para comparar los modelos con respecto al pronóstico de las caracterizaciones de riesgo de contraer cáncer.

Antecedentes

Hartley y Sielken (1977) proponen un modelo para realizar el cálculo de riesgos de contraer cáncer. Básicamente fué obtenido a través de modelación estocástica del proceso biológico y modelos compartamentales. El modelo ha demostrado ser muy versátil y ajustar satisfactoriamente un gran número de conjuntos de datos.

Brown y Hoel (1983) comparan varios modelos y concluyen que el modelo de Hartley-Sielken puede no ser el adecuado para modelar cierto tipo de datos. Los datos utilizados por estos investigadores corresponden a la presencia de cáncer en el hígado en ratones sacrificados en el estudio ED₀₁ (Ver Staffa y Mehlman (1979)).

El Presente trabajo tiene la finalidad de mejorar el ajuste del modelo de Hartley-Sielken a través de una generalización del mismo. Se espera que al mejorar el ajuste se mejore el cálculo de las caracterizaciones del riesgo de contraer cáncer (CRC). Varios modelos son propuestos encontrándose el valor de su verosimilitud en algunos casos. Tambien se propone un nuevo método para encontrar intervalos de confianza para las CRC. Finalmente se realiza un experimento de simulación para observar el comportamiento del modelo con respecto a las CRC.

* Centro de Calidad y Productividad, DGI, ITESM, Campus Estado de México.

** Universidad de las Américas, Puebla.

Modelos para cálculo de riesgos.

En varios artículos se presentan modelos para calcular el riesgo de contraer cáncer. Uno de los pioneros en esta área de investigación es el modelo de Hartley-Sielken. Una generalización parcial puede verse en Burguete, *et al.*(1991). La formulación culmina en la escritura de la función de distribución acumulativa $P(t;d)$, donde t denota un valor particular de la variable aleatoria del tiempo de respuesta, y d la dosis de la sustancia carcinógena. $P(t;d)$ tiene la forma:

$$P(t;d) = 1.0 - \exp(-\Delta(t,d)).$$

donde $\Delta(t, d)$ es una función creciente. La forma de $\Delta(t,d)$ determina varios modelos. Por ejemplo, si:

$$\Delta(t,d) = \sum_{r=1}^R K_r^{(1)} t^r \sum_{s=0}^S K_s^{(2)} d^s, \quad t \geq 0,$$

se le conoce como Modelo Multiplicativo, el cual fué desarrollado por Hartley y Sielken (1977). El Modelo No Multiplicativo, desarrollado por Burguete *et al.*(1991) se obtiene considerando:

$$\Delta(t, d) = \sum_{r=1}^R \sum_{s=0}^S K_{rs} t^r d^s, \quad t \geq 0,$$

en el que puede incorporarse el periodo de latencia (LP) con lo que la expresión de $\Delta(t, d)$ es:

$$\Delta(t, d) = \sum_{r=1}^R \sum_{s=0}^S K_{rs} (t - LP)^r d^s, \quad t \geq LP.$$

Extensiones naturales a esta formulación son la inclusión de un período de latencia variable si

$$\Delta(t, d) = \sum_{r=1}^R \sum_{s=0}^S K_{rs} (t - (a + bd))^r d^s, \quad t \geq a + bd,$$

o bien, puede extenderse para incluir efectos sinergéticos entre sustancias cancerígenas, si

$$\Lambda(t, d_1, d_2) = \sum_{r=1}^R \sum_{s=0}^S \sum_{u=0}^U K_{rsu} t^r d_1^s d_2^u, \quad t \geq 0,$$

donde d_1 y d_2 son las dosis del carcinógeno 1 y 2 respectivamente.

Los modelos arriba mencionados pertenecen a una categoría general conocida como Modelos Paramétricos porque asumen una relación funcional entre dosis y tiempo. Cuando no se asume tal relación el modelo se conoce como Modelo No Paramétrico.

En particular existen dos casos de modelos no paramétricos de suma importancia. El modelo No Paramétrico No Multiplicativo (NPNM) es aquél que estima la $P(T;d) = \mu_{t,d}$, una función creciente. Para este modelo el valor de su verosimilitud es la cota superior para la verosimilitud de cualquier modelo. El modelo No Paramétrico Multiplicativo (NPM) es de la forma $P(T;d) = \mu_t \mu_d$, con μ_t y μ_d crecientes.

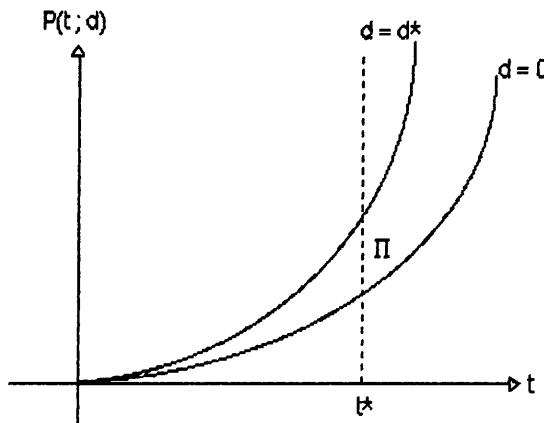
Caracterizaciones del riesgo.

Como caracterizaciones del riesgo de contraer cáncer se entienden aquellas dosis cuyos niveles no causan un aumento alarmante en la proporción de presencia de cáncer, o que la reducción en el tiempo esperado de vida de una persona no es significativa. Dos de estas medidas se mencionan a continuación.

Definición 1. La Dosis Virtualmente Segura (DVS) es la dosis d^* tal que

$$\Pi = P(t^*; d^*) - P(t^*; 0),$$

con Π cercano a cero y t^* fijada como una longitud de vida promedio o la duración del bioensayo. La DVS es la dosis correspondiente a un incremento permisible sobre la proporción espontánea de cáncer. Gráficamente:

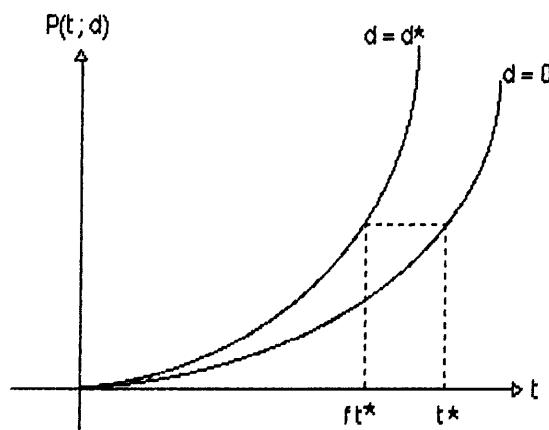


Otra CRC ampliamente usada es la dosis de riesgo tardía que es definida como la dosis correspondiente a un incremento aceptable de riesgo tardío.

Definición 2. La Dosis de Riesgo Tardía (DRT) es la dosis d^* tal que:

$$P(t^*; 0) = P(f t^*; d^*),$$

con f cercano a uno y t^* especificado. Gráficamente:



Procedimiento de Estimación.

Defina θ a ser el vector de parámetros. Entonces bajo el protocolo experimental adecuado la función de verosimilitud resulta ser:

$$L(\theta) = \prod_{i=1}^D \left\{ \left(\prod_{j=1}^{n_{i1}} P(t_{ij}; d_i) \right) \prod_{k=1}^{n_{i2}} P(t_{ik}; d_i) \prod_{m=1}^{n_{i3}} Q(t_{im}; d_i) \right\},$$

donde: D = número total de dosis,

$$p(t; d) = P(t; d),$$

$$Q(t; d) = 1 - P(t; d),$$

n_{i1} = número de animales con observación directa de cáncer,

n_{i2} = número de animales con necropsia positiva, y

n_{i3} = número de animales con necropsia negativa.

Para encontrar los estimadores de los parámetros se maximiza la función de verosimilitud con respecto a los parámetros. Los estimadores así obtenidos se llaman Estimadores de Máxima Verosimilitud (EMV). Es de notarse que el número de parámetros es grande (entre 20 y 40). Por lo que el procedimiento adecuado para optimización de la verosimilitud es el propuesto por Torckson y Dennis(1989). Este procedimiento mostró una superioridad muy clara sobre el de Nelder - Mead , descrito en Thompson(1989).

Por la misma naturaleza del problema de optimización se estudiaron dos subproblemas en el modelo paramétrico no multiplicativo con período de latencia fija. El primero es

$$\text{Max } L(\theta)$$

sujeto a: $\theta \geq 0$,

al que llamaremos Modelo Paramétrico No Multiplicativo Uno (PNM1).

El segundo que se consideró es el proceso Paramétrico No Multiplicativo Dos (PNM2) que consiste en

$$\text{Max } L(\theta)$$

sujeto a: $\Lambda(t; d)$ creciente.

Estos modelos y el de Hartley-Sielken (PM) fueron ajustados a los datos usados por Brown y Hoel (1983). Encontrándose los siguientes resultados de la función de verosimilitud.

PROCESO	NPNM	PNM2	NPM	PNM1	PM
VALOR DE -LOG(L(θ))	1818.0	1837.6	1839.2	1844.5	1849.8

De estos resultados se observa que el mejor ajuste se obtuvo en el modelo PNM2.

Para calcular la DVS y la DRT normalmente se usa bootstrap. Esto implica un esfuerzo computacional muy grande. Se presenta a continuación un teorema para realizar intervalos de confianza asintóticos. Cuya finalidad es aliviar el esfuerzo computacional asociado al cálculo de las CRC. Ver Burguete *et al.* (1991).

Teo. Sea $f(\theta, d) = 0$ una función implícita de $\theta \in \Omega \subseteq \mathbb{R}^k$, donde $d=g(\theta)$. $f(\theta, d)$ continua y $\hat{\theta}$ el EMV DE θ . Si $0 < \left| \frac{\partial f(\theta, d)}{\partial d} \right| \leq M$, en una vecindad de $f(\theta, d) = 0$ que no incluye a 0, entonces $\hat{\theta}$ es AN $(d, v(\theta))$, donde $v(\theta) = u'(l(\theta))^{-1}u$,

$$u'(\theta) = \frac{1}{\partial f(\theta, d)/\partial d} \left(\frac{\partial f(\theta, d)}{\partial \theta_1} \dots \frac{\partial f(\theta, d)}{\partial \theta_k} \right).$$

Resultados y Conclusiones.

Para estudiar el comportamiento empírico del modelo PNM2 cuando los datos provienen de un modelo paramétrico multiplicativo (PM) y viceversa, se realizó un experimento de simulación, con características similares a los experimentos que se llevan a cabo en la actualidad. La medida usada para comparación entre modelos fué la suma de las desviaciones absolutas. Se simularon 20 conjuntos de datos de cada modelo y se ajustó cada conjunto de datos con ambos modelos. Se consideraron períodos de latencia al simular y al ajustar.

Entre los conjuntos de datos que fueron simulados con PNM2, este modelo ajustó mejor en 19 de ellos. Mientras que entre los datos que fueron simulados con PM, el modelo PNM2 ajustó mejor en 17 de ellos.

Las CRC se mostraron favorables al modelo PNM2 aunque no de manera tan dramática. Cuando los datos provenían del modelo PNM2, este modelo estaba más cerca de la CRC verdadera en 16 de los 20 conjuntos de datos. En el caso de los datos simulados con PM, el modelo PNM2 fué mejor en 15 ocasiones.

Los resultados y sugerencias más relevantes son:

El modelo multiplicativo es superado por el modelo no multiplicativo. Este modelo incluso supera a la verosimilitud de un modelo no paramétrico multiplicativo. Esto es importante si se considera que el modelo paramétrico además es útil para llevar a cabo CRC, lo cual no puede ser realizado con el modelo no paramétrico.

Las CRC son mejor estimadas con el modelo PNM2 aunque no de manera uniforme. Nótese que el cálculo de las CRC es la meta de los modelos aquí expuestos. Una mayor investigación en este sentido es requerida.

Para el ajuste de los modelos con un gran número de parámetros se recomienda el uso del algoritmo de búsqueda multidireccional desarrollado por Torckson y Dennis(1989). En todos los ajustes del presente trabajo este algoritmo mostró ser superior al de Nelder - Mead .

El teorema expuesto en la sección anterior puede ser útil para calcular intervalos de confianza y pruebas de hipótesis sobre las CRC. En este sentido es necesario efectuar una comparación entre el método de bootstrap utilizado tradicionalmente y la proposición del presente trabajo.

Referencias Bibliográficas.

- Brown, K.G. y Hoel, D.G. (1983). Modelling time-to-tumor Data: Analysis of the ED₀₁ Study. Fundamental and Applied Toxicology 3, 458-469.
- Burguete, E., Burguete, F. y Sielken Jr. R. L. (1991). Generalizacion de un Modelo de Contraer Cáncer. Agrociencia, Serie MAEC 2 (2):95-114
- Hartley, H.O. y Sielken Jr. R. L. (1977). Estimation of "Safe Doses" in Carcinogenic Experiments. Biometrics 33, 1-30.
- Staffa, J. A. y Mehlman. M. A. (1979). Innovations in Cancer Risk Assessment. Pathotox Publishers. Park Forest South, IL.
- Thompson, J. R. (1989). Empirical Model Building. John Wiley and Sons. New York. USA.
- Torckson, V. A. y Dennis, John. (1989). The Multidirectional Search Algorithm Technical Report 90-7. Department of Mathematical Sciences. Rice University. Houston, Texas.

CAPACITACIÓN EN CALIDAD: ADMINISTRACIÓN VS. ESTADÍSTICA

Humberto Gutiérrez Pulido¹

Román de la Vara Salazar¹

RESUMEN

En la actualidad existe una demanda creciente de capacitación en control de calidad por parte de las empresas. La respuesta de las universidades y de asesores privados ha sido ofrecer diversos cursos y/o diplomados sobre el tema. En muchos casos la incidencia real de estos cursos sobre el nivel de calidad de las empresas ha sido bastante pobre, debido entre otras razones a que el asistente no modifica de fondo su forma de percibir el problema de la calidad y a que no queda realmente capacitado en las herramientas que permiten mejorarla.

En este trabajo se analizan y discuten algunos elementos de la problemática del control total de calidad (CTC); para concluir con el esbozo de una propuesta de qué es lo que se debe enseñar para mejorar la calidad, señalando algunos aspectos esenciales de esta enseñanza.

EL PROBLEMA DE LA CALIDAD

La apertura comercial del país ha provocado que la invasión de productos extranjeros sea prácticamente generalizada, lo que ha llevado a una situación desesperante a muchas empresas y ramas industriales del país. Un dato de esta situación es el déficit comercial con el exterior: en los primeros seis meses de 1992 fue de

¹Facultad de Ingeniería, Universidad de Guadalajara.
Campus Tecnológico, Avenida Revolución, Guadalajara, Jal.

10400 millones de dólares incluyendo el petróleo. El aumento del déficit respecto al año anterior es del 105%. Esto se refleja en que muchas empresas nacionales estén perdiendo mercado y un número importante de ellas se encuentren en un dilema: cerrar y dedicarse a la especulación financiera o a la importación (como ya lo han hecho otras), o arriesgar aún más el patrimonio que durante años lograron construir. Una de las industrias que mejor ilustra esta situación es la zapatera: antes de la apertura comercial era una industria netamente exportadora, ahora es una industria deficitaria.

Nuestra actividad de capacitación en control de calidad nos ha permitido conocer de cerca el problema de la calidad. A continuación señalaremos algunos elementos de tal problemática.

a. Es común encontrarse empresas que sus ventas han venido decreciendo sistemáticamente, generándose hacia el interior una presión sin precedentes, manifestándose ésta en despidos y exigencias como: reducir costos, mejorar la calidad de los productos, desarrollar nuevos productos que compitan con los importados, mejorar continuamente, atender al cliente, etcétera; la reacción de los obreros y mandos medios es de desconcierto, pues no entienden qué es mejorar calidad, sin gastar más, no saben cómo pueden hacerlo y desconocen el con qué. En suma, en muchas empresas hay exigencias y presiones para ser más competitivo, pero no se conoce con profundidad el qué, el cómo y el con qué de la mejora continua.

b. Hemos encontrado empresas que tienen varios años que han iniciado programas de CTC, pero que sus problemas en términos generales siguen siendo los mismos.

c. Existe personal técnico en las empresas que han recibido

varios cursos y/o diplomados de capacitación en CTC, pero que los problemas de calidad los siguen enfrentando igual que siempre.

d. Existen empresas que a pesar de que sus ventas han ido disminuyendo, que han hecho recortes de personal, que su panorama hacia el futuro es gris, no se han decidido actuar y atribuyen la situación a la pérdida de poder adquisitivo, a la competencia desleal; es decir, atribuyen sus problemas a causas que están fuera de su alcance y por lo tanto hay que resignarse, aguantar y, si las cosas se complican más, vender o cerrar.

A continuación señalaremos algunas de las causas de los problemas que hemos señalado antes.

UNA PROBLEMATICA COMPLEJA (ALGUNAS CAUSAS)

1. PENSAR QUE LA MEJORA DE LA CALIDAD ES SÓLO UN PROBLEMA ADMINISTRATIVO. Muchos directivos y consultores piensan que la mejora continua es sólo un problema administrativo, en que hay que exigir calidad, reasignar funciones, renovar estrategias, declarar y exigir la participación de todos, responsabilizar a un departamento de la calidad, etcétera. Muchos de los pequeños cursos o seminarios de calidad que se ofrecen en el mercado están impregnados de este concepto; en ellos se dice que un programa de CTC es concientizar a todos de la importancia de la calidad para que así hagan bien todo "a la primera vez" y cometan "cero defectos", olvidándose que más del 80% de los problemas de calidad en una empresa se debe al sistema y no a la gente. Las propuestas de estos cursos están impregnados de activismos y buenos propósitos, y a mediano plazo no tienen una incidencia real en la calidad y conducen a falsos comienzos,

estancamiento e incertidumbre. En México esta concepción está muy difundida, desafortunadamente.

2. CERRAZÓN EN LAS EMPRESAS. La alta dirección de muchas empresas se encuentra cerrada a todo cambio, a toda innovación y se aferra a seguir aplicando las medidas que siempre les dieron resultado. No promueven la capacitación de su gente, más bien la obstaculizan, no creen en la asesoría externa, no permiten que los mandos medios ensayan nuevas soluciones. En algunas de estas empresas es frecuente que el director general sea el dueño y fundador o un descendiente de éste, por lo que es muy difícil promover un cambio y la respuesta típica es "para que cambiar los métodos de trabajo, si éstos permitieron transformar el taller de hace 30 años a esta empresa que ves ahora". Lamentablemente esa es la situación de muchas empresas del país. Los mercados globalizados exigen un nivel de competitividad sin precedentes, por lo que las empresas cerradas cambian o desaparecen. Mercados abiertos, mentes abiertas e innovadoras.

3. "EXPERTOS" EN CTC CON ESCASA FORMACIÓN ESTADÍSTICA. Muchos de los consultores o "expertos" en calidad son profesionistas de distintas disciplinas (con predominio de las administrativas) que adquirieron sus conocimientos en la práctica, a través de la lectura de algunos libros de "filosofía de la calidad" o por medio de cursos intensivos, las más de las veces. Muchas de estas personas tienen una visión muy particular de la calidad, difícilmente innovadora. Es típico que los "expertos" en CTC se hayan formado en la práctica y que sus lecturas iniciales fueran las obras de P. Crosby, cuya visión de la calidad está impregnada del enfoque descrito en el punto (a), ver Crosby(1984). Los más de los actuales "expertos" mexicanos en CTC

tienen una capacitación deficiente sobre herramientas y técnicas estadísticas. De aquí que sus propuestas y/o cursos de CTC estén enfocados fundamentalmente hacia aspectos administrativos.

4. CAPACITACIÓN INTENSIVA Y NO COMPROMETIDA CON EL CAMBIO.

Muchos de los cursos de capacitación sobre CTC que se ofrecen en el mercado son cortos e intensivos, por ejemplo ocho horas diarias durante una semana o menos. El asistente a estos cursos ve un cúmulo de nuevas ideas y conceptos, todo ocurre tan rápido y artificial que a la siguiente semana cuando llega a su trabajo, lo visto en el curso se le hace como un sueño. En otras palabras, estos cursos no logran modificar de fondo la percepción del asistente y no le proporcionan los conocimientos necesarios para tomar las decisiones adecuadas en los problemas que cotidianamente enfrenta. Otra característica de este tipo de "capacitación" es que los cursos están ya preparados de tal forma que no permiten analizar o abordar problemas propios de la empresa (o de los asistentes).

5. LA CALIDAD COMO NEGOCIO Y COMO MODA. Existen consultores e instituciones que antes de ofrecer y hacer un buen trabajo, ven el signo de pesos en el CTC, como si el movimiento mundial por la calidad fuera una moda a la que hay que sacarle provecho económico.

6. LA REACCIÓN OFICIAL ES TÍMIDA Y TARDÍA. La actual política económica de apertura comercial inicio desde 1982, y debió acompañarse con un programa gubernamental de apoyo a la competitividad de las empresas. Es hasta 1989 cuando el gobierno federal declara su intención de lograr un acuerdo nacional para incrementar la productividad industrial. Los sindicatos se opusieron (¿desde cuando tal oposición ha sido un obstáculo en los planes oficiales?) argumentando

que primero se debería convenir de que manera se iban a repartir los beneficios del supuesto incremento de la productividad. ¡Es hasta 1992! cuando finalmente se logra concretar el "Acuerdo Nacional para la Elevación de la Productividad y la Calidad", que por cierto uno de sus puntos es que en cada empresa, patrones y sindicato, se pongan de acuerdo en cómo se van a repartir los beneficios del aumento de la productividad (tardaron más de tres años para llegar a tal conclusión). Esto y que diez años después de iniciada la apertura comercial se haya concretado tal acuerdo, muestra la disposición gubernamental hacia el problema de la competitividad industrial, nosotros decimos que es tímida y tardía. Si la política oficial hacia la calidad sigue igual es seguro que el citado convenio incremente las ya abultadas filas de los discursos, declaraciones y buenas intenciones.

7. PLANES Y PROGRAMAS DE ESTUDIOS OBSOLETOS. El CTC ha traído con sigo una nueva concepción de la empresa, las formas organizativas, los métodos de trabajo, las relaciones laborales y de la administración misma. Además han aparecido nuevas técnicas y herramientas, o bien nuevos enfoques de técnicas ya existentes. Sin embargo, los más de los planes y programas de estudio en los centros educativos no han registrado tales cambios. Varios programas educativos de todos los niveles siguen con el mismo enfoque de hace décadas, se siguen enseñando los métodos y técnicas que se aplicaron hace varios años.

Los más de los ingenieros egresan sin saber CTC, sin contar con los instrumentos para tomar decisiones. Los egresados de las carreras administrativas se les sigue enseñando como mantener la riqueza, y no cómo generarla, sus conocimientos son sumamente deficientes en cuanto

a instrumentos analíticos, por lo que es común que las decisiones las tomen apoyándose en coronadas y en la experiencia (costumbres).

8. LOS GREMIOS EMPRESARIALES NO SE HAN COMPROMETIDO DE MANERA DECIDIDA CON EL PROBLEMA DE LA COMPETITIVIDAD. Los dirigentes empresariales en lugar de comprometerse y trabajar en el problema de la competitividad de sus dirigidos, siguen en la tónica de realizar actos y declaraciones de "relumbrón" que les haga eco la prensa. Al menos en Guadalajara, las actividades de las cámaras empresariales entorno a la calidad ha sido poca y deficiente.

CONCLUSIONES

Cuando se ofrece capacitación en calidad, hay que hacerlo con calidad, y eso implica enseñar lo adecuado y enseñarlo bien.

Hemos visto una serie de causas debido a las cuales el problema de la competitividad de las empresas nacionales es cada día más grave. Estas causas están interrelacionadas, de tal forma que unas agravan las otras. Por ejemplo, existen casos en los que la oposición de los directivos a todo cambio en las empresas ha sido agravada por intentos fallidos en la implementación del CTC. La pérdida de mercados y la falta de apoyo gubernamental provocan que las empresas difícilmente estén dispuestas a hacer gastos adicionales con la finalidad de iniciar un programa que conduzca al CTC.

Nuestra primera conclusión es que los involucrados en el problema del CTC deben estar conscientes de que la transformación hacia la calidad empieza con educación y continua con educación. Los empresarios y directivos deben estar convencidos que es necesario que se capaciten para iniciar un programa de CTC, que conozcan qué es y

cómo se logra el CTC, que dominen las siete herramientas básicas para el CTC como instrumentos indispensables para tomar decisiones y para interactuar en un proceso de mejora. No es posible que los empresarios y altos directivos no estén dispuestos a dedicarle 30 ó 40 horas a un curso de capacitación.

Los consultores y expertos en calidad deben mejorar su preparación en métodos estadísticos y en el qué y cómo de la mejora continua. Deben ofrecer un mejor trabajo.

Quien mejor ha definido el qué del CTC es W.E. Deming (ver Deming, 1989). Algunos de los que más han aportado al cómo del CTC es J. Juran (ver Juran, 1990, por ejemplo) y K. Ishikawa (ver Ishikawa, 1986). El con qué del CTC son esencialmente métodos y técnicas estadísticas, ver por ejemplo Kane(1987) y Montgomery(1985 y 1987). Una síntesis de los aspectos básicos del qué, el cómo y el con qué se puede consultar en Gutiérrez(1992).

Una conclusión mas es que las instituciones (gobierno, centros de enseñanza, cámaras empresariales, sindicatos) asuman su papel en la calidad con pleno conocimiento de causa. El gobierno elaborando un programa sólido de apoyo a la competitividad de las empresas. Los centros de enseñanza formando y capacitando el recurso humano. Las cámaras empresariales estimulando a sus agremiados con apoyo en CTC. Los medios de comunicación pueden desempeñar un papel sumamente importante. Japón es un ejemplo ilustrativo de cuál es el papel de las instituciones en el CTC.

Otra conclusión es que más estadísticos se involucren en el CTC. No es aceptable que varios de los grupos importantes de estadísticos del país no estén aportando sus conocimientos para apoyar el CTC. No

estamos proponiendo que todos los estadísticos se dediquen al CTC, pero sí que en todo grupo de estadísticos haya una parte de éste que se involucre en el CTC. Para ilustrar la importancia de esto, citamos una afirmación de E.W. Deming: "El conocimiento estadístico es la tecnología más escasa en las empresas, y es la tecnología más útil para mejorar la calidad". Apoyándonos en esta frase y en la historia del CTC, afirmamos que el CTC no es sólo un problema administrativo, es también un problema de índole técnico (estadístico).

Finalmente proponemos que la capacitación en CTC tenga las siguientes características.

i) Si la capacitación se va a dar a una empresa en especial, ésta debe estructurarse a partir de un diagnóstico de la problemática de la compañía, de tal forma que el curso esté enfocado a los problemas actuales y que la capacitación se retroalimente de los problemas vigentes. Por ejemplo, podría iniciarse presentando los conceptos y herramientas generales, y a medida que avanza el curso se podrían analizar algunos problemas de la empresa.

ii) La capacitación dirigida a altos directivos debe lograr que estos conozcan lo suficiente del qué, el cómo y el con qué del CTC.

iii) La capacitación que esté dirigida a personal técnico y directivo que estén involucrados de manera importante en el CTC, debe cumplir el objetivo de que además de que conozcan los aspectos básicos del CTC, deben dominar bien la aplicación de los diseños de experimentos. Nuestra propuesta concreta de capacitación en este caso la sintetiza el programa del "Diplomado en Control Total de Calidad" que coordinamos y ofrecemos en la U. de Guadalajara. El temario de tal diplomado es el siguiente. 1. Calidad, productividad y com-

petitividad. 2. Filosofía Deming. 3. Paquete computacional Statgraphics. 4. Las herramientas básicas. 5. Fundamento y aplicación de las Cartas de control. 6. Capacidad de procesos. 7. Estrategias para mejorar la calidad. 8. Círculos de calidad. 9. Metodología para la solución de problemas. 10. Muestreo de aceptación. 11. Normas ISO-9000. 12. Análisis de varianza y regresión. 13. El diseño de experimentos en el CTC. 14. Diseños factoriales y f. fraccionados. 16. Optimización de procesos industriales. 17. Justo a tiempo. 18. Teoría de restricciones 19. QFD. La duración del diplomado es de 160 horas, repartidas en dos sesiones semanales de 4 horas cada una.

REFERENCIAS

- Crosby, P.(1984). La Calidad no Cuesta. CECSA, México.
- Deming, W.E.(1989). Calidad, Productividad y Competitividad. Díaz de Santos, Madrid.
- Gutiérrez P.,H.(1992). Control Total de Calidad. Editorial Universidad de Guadalajara, Guadalajara.
- Ishikawa, K.(1986). ¿Qué es el Control Total de Calidad?. Norma, Bogota.
- Juran, J.A.(1990). Juran y el Liderazgo para la calidad. Díaz de Santos, Madrid.
- Kane, V.(1987). Defect Prevention: Integrating Process Control and Problem Solving Tools. Marcel Decker, Nueva York.
- Montgomery, D.C.(1985). Introduction to Statistical Quality Control. Wiley, Singapur.
- Montgomery, D.C.(1987). Design and Analysis of Experiments. Wiley, Nueva York.

DISEÑO DE EXPERIMENTOS EN EL MODELO DE MICHAELIS-MENTEN

Dr. Juan Gaytán Iniestra*, Erendira Rendón Lara[†]

En este trabajo se estiman los parámetros del modelo Michaelis-Menten utilizado en la Cinética Enzimática con el apoyo de un diseño experimental basado en el criterio de Box-Lucas.

La estimación de parámetros en modelos de regresión no lineales es una tarea complicada, y por lo general se recurren a técnicas iterativas. Uno de los criterios más utilizados es el de Box-Lucas, el cual consiste en la aplicación iterativa de estimación de parámetros donde se han agregado en cada iteración nuevos puntos de diseño sobre los cuales se observa la respuesta.

Este trabajo propone una función que relaciona los puntos de diseño que se agregan en cada iteración con los puntos de diseño anteriores, esto tiene la ventaja de ir efectuando un nuevo punto a la vez.

La aplicación de la técnica propuesta a un problema tomando de la literatura sugiere que la estimación de los parámetros del modelo puede ser hecha con un número menor de puntos experimentales.

* ITESM - Campus Toluca.
Ex hacienda La Pila
Toluca, Edo, de México.

[†]Facultad de Ingeniería de la UAEM
Unidad Cerror de Coatepec, C.U.
Toluca, Méx. 50130

1.- Concepto de cinética Enzimática.

La cinética enzimática es un campo de la bioquímica que se encarga del estudio de las reacciones enzimáticas y de los factores que afectan la velocidad de una reacción catalizada por enzimas.

Estudios sobre el análisis de datos en cinética enzimática, asumen que la concentración de un sustrato $[S]$ influye sobre la velocidad de reacción v , cuando la enzima se agrega se forma el complejo enzima sustrato $[E_0]$, la velocidad de formación de ese complejo se puede describir en base a las siguientes ecuaciones:

$$v / [E_0] = \sum_{i=1}^n \alpha_i [S]^i / \sum_{i=1}^n \beta_i [S]^i$$

o bien

$$v / [E_0] = \sum_{i=1}^n a_i [S]^i / \left(1 + \sum_{i=1}^n b_i [S]^i \right)$$

La teoría Michaeliana (Michaelis y Menten, 1913) indica que la velocidad v de una reacción enzimática es una función hiperbólica de la concentración del sustrato $[S]$, con la estructura siguiente:

(Modelo 1:1) $v = \frac{V_M [S]}{K_M + [S]}$

Donde los parámetros V_M y K_M representan la máxima velocidad de la reacción y la constante de Michaelis, respectivamente.

La gráfica del modelo anterior es la figura 1.

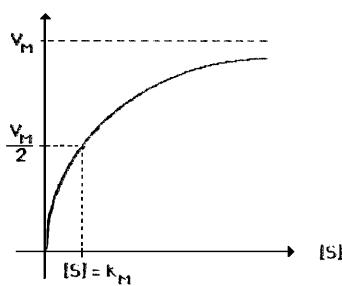


Figura 1

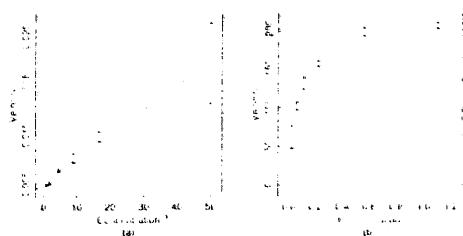


Figura 2

Otras teorías No Michaelianas

En los últimos años se ha cuestionado la validez de la teoría Michaeliana y se han propuesto otros modelos para representar la velocidad en función de la concentración del sustrato, algunos de ellos son:

(Modelo 2:2) $v / [E_0] = \frac{\alpha_1[S] + \alpha_2[S]^2}{\beta_0[S] + \beta_2[S]^2}$ (3)

(Modelo 3:3) $v / [E_0] = \frac{\alpha_1[S] + \alpha_2[S]^2 + \alpha_3[S]^3}{\beta_0[S] + \beta_1[S]^2 + \beta_2[S]^3}$ (4)

(Modelo 4:4) $v / [E_0] = \frac{\alpha_1[S] + \alpha_2[S]^2 + \alpha_3[S]^3 + \alpha_4[S]^4}{\beta_0[S] + \beta_1[S]^2 + \beta_2[S]^3 + \beta_3[S]^4}$ (5)

siendo los parámetros de esos modelos las α_i 's y β_i 's

En este trabajo nos concentraremos en la estimación de los parámetros del modelo Michaeliano (1:1).

2.- Estimación tradicional de los parámetros del modelo Michaeliano.

La estimación de los parámetros del modelo es hecha tradicionalmente utilizando los dos diseños experimentales siguientes:

- **Espaciamiento lineal.** Se divide en partes igualmente espaciadas el intervalo de valores de concentración del sustrato [S]. Es decir, si S_{\min} y S_{\max} son los valores mínimo y máximo de [S], entonces se experimenta en

$$[S]_i = [S]_{i-1} + \Delta, i = 1, \dots, N \quad (6)$$

donde

$$\Delta = \frac{S_{\max} - S_{\min}}{N}$$

$$S_0 = S_{\min}$$

- **Espaciamiento logarítmico.** Se subdivide el intervalo $[S_{\min}, S_{\max}]$ en puntos S_i definidos como sigue:

$$[S]_i = [S]_{i-1} * \Delta, i = 1, \dots, N \quad (7)$$

donde

$$\Delta = [S_{\max} / S_{\min}]^{(i-1)/(N-1)}, i = 1, \dots, N$$

$$S_0 = S_{\min}$$

Los valores S_{\max} y S_{\min} se definen iguales al caso de espaciamiento lineal.

Una vez conocidos los puntos de diseño, se procede a realizar la experimentación, para obtener las parejas $([S]_i, v_i)$, $i = 1, \dots, N$.

Finalmente, la estimación tradicional de los parámetros del modelo es hecha linealizando el modelo; es decir, si observamos que el recíproco de la velocidad v es lineal en el recíproco de la concentración del sustrato $[S]$, entonces se puede utilizar regresión lineal simple ajustando los datos $(1/[S]_i, 1/v_i)$ a un modelo lineal simple. En general esa técnica es satisfactoria, sin embargo no es adecuada en general, sobre todo cuando los datos violan el supuesto de regresión de varianza constante. Por ejemplo considere los datos de un experimento realizado por Treloar (1974) y cuyos datos se encuentran en la tabla 1:

1	0.02	0.02	0.06	0.06	0.11	0.11	0.22	0.22	0.56	0.56	1.10	1.10
2	76	47	97	107	123	139	159	152	191	201	207	200

(1) Concentración del sustrato (ppm)

(2) Velocidad (cuentas/min)²

Tabla 1

La gráfica de los datos transformados se muestra en la figura 2.

3.- Criterio de Box-Lucas (Criterio D-óptimo).

Es un criterio basado en la minimización del determinante de la matriz de varianza-covarianza del diseño respecto a los puntos de diseño. Para modelos no lineales de la forma

$$Y = f(X, \theta) + \varepsilon \quad (8)$$

cuya i -ésima componente de $f(x, \theta)$ es $f(x_i, \theta)$, Box-Lucas proponen utilizar la matriz de derivadas parciales $F(\theta_0)$ (la matriz Jacobiana del modelo) evaluada en un punto inicial θ_0 y resolver el

problema de optimización siguiente:

$$\text{Max } |\mathbf{F}(\theta_0)^T \mathbf{F}(\theta_0)| \quad (9)$$

donde la función $\mathbf{F}(\theta_0)$ es la matriz $N \times p$ (N es el número de observaciones y p es el número de parámetros) cuyo elemento (i,j) es del forma

$$\frac{\partial f(x_i, \theta)}{\partial \theta_j} \Big|_{\theta = \theta_0}$$

Note que esto es equivalente a aproximar en serie de Taylor hasta términos de primer orden a la función:

$$f(x_i, \theta) \approx f(x_i, \theta_0) + \sum_{j=1}^p (\theta_j - \theta_{j0}) \frac{\partial f(x_i, \theta)}{\partial \theta_j} \Big|_{\theta = \theta_0}$$

para $i = 1, 2, \dots, N$.

Es decir, cuando θ es cercano a θ_0 , se tiene aproximadamente

$$\mathbf{Z} = \mathbf{F}(\theta_0) \psi + \mathbf{E} \quad (10)$$

donde

$$\mathbf{Z} = (z_1, z_2, \dots, z_N) \quad \text{con } z_i = y_i - f(x_i, \theta_0), \quad (i = 1, 2, \dots, N),$$

$$\psi = \theta - \theta_0$$

$\mathbf{F}(\theta_0)$ como se definió antes, y

$$\mathbf{E} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N)$$

Así, Box-Lucas (1959) proponen como diseño aquellos puntos que minimicen la varianza del estimador de mínimos cuadrados $\hat{\psi}$ de ψ en el modelo lineal (10). Esto se logra minimizando el determinante de la matriz de varianza-covarianza $\mathbf{M}^{-1}(\mathbf{D}, \theta) \cdot \sigma^2$, donde

$$\mathbf{M}(\mathbf{D}, \theta_0) = \mathbf{F}(\theta_0)^T \mathbf{F}(\theta_0)$$

y donde \mathbf{D} denota la matriz de diseño $N \times K$ con elementos (x_{ik}) , $i = 1, \dots, N$, $k = 1, \dots, K$.

Es decir, el criterio de Box-Lucas elige los puntos de diseño que maximizan el determinante de $\mathbf{M}(\mathbf{D}, \theta_0)$.

Problemática:

En la presentación anterior se supuso que θ_0 es conocido, sin

embargo para conocerlo se requiere haber propuesto un diseño preliminar, pero esto es lo que andamos buscando. Esta dependencia es, sin embargo, una característica no deseable de los modelos no lineales. Para resolver esto se propone una estrategia secuencial donde inicialmente se propone un valor inicial de θ , θ_0 , en base a él se determina un nuevo punto de diseño en base al criterio Box-Lucas (9). Con ese nuevo punto de diseño y los anteriores se encuentra una mejor estimación de θ , digamos θ_1 , y el proceso se repite. El procedimiento termina cuando " θ_i es cercano a θ_{i-1} ".

4.- El Método Secuencial Aplicado al Modelo Michaelis-Menten.

Consideremos el modelo de Michaelis-Menten siguiente:

$$v_i = \frac{\theta_1 x_i}{\theta_2 + x_i} + \epsilon_i \quad (11)$$

donde θ_1 y θ_2 son los parámetros a ser estimados. Note que hemos usado la variable x en lugar de $[S]$.

Supongamos que $x_{\min} \leq x \leq x_{\max}$, con $0 \leq x_{\min} < x_{\max}$ valores mínimo y máximo de la concentración del sustrato.

Supongamos también que $\theta_0 = (\theta_{10}, \theta_{20})$ es el vector que estima inicialmente el vector de parámetros θ .

El criterio Box-Lucas encuentra los dos primeros puntos de diseño resolviendo para x_1 y x_2 con $x_{\min} \leq x_1, x_2 \leq x_{\max}$ el problema: $\{x: \max |F(\theta_0)^T F(\theta_0)|\} = \{x: \max |F(\theta_0)|\}$, (12)

donde

$$F(\theta_0) = \begin{vmatrix} \frac{x_1}{\theta_{20} + x_1} & \frac{-\theta_{10} x_1}{(\theta_{20} + x_1)^2} \\ \frac{x_2}{\theta_{20} + x_2} & \frac{-\theta_{10} x_2}{(\theta_{20} + x_2)^2} \end{vmatrix}$$

El máximo ocurre en: $x_1 = x_{\max}$ y $x_2 = \frac{\theta_{20}^2}{1 + 2(\theta_{20}/x_{\max})}$ (13)

Los puntos anteriores son reportados en la literatura (Bates y Watts, 1988). En este trabajo extendemos ese resultado a un número

arbitrario de puntos de diseño y los puntos obtenidos los comparamos contra los obtenidos con otras estrategias de estimación de los parámetros.

Una vez conocidos x_1 y x_2 , se realiza la experimentación en esos puntos de diseño para poder obtener una nueva estimación $\theta_1 = (\theta_{11}, \theta_{21})$ del vector de parámetros θ utilizando algún método de estimación de regresión no lineal. El nuevo punto de diseño x_3 se obtiene aplicando nuevamente el criterio Box-Lucas como sigue:

$$\text{Si } F(\theta_0) = \begin{vmatrix} \frac{x_1}{\theta_{21} + x_1} & \frac{-\theta_{11} x_1}{(\theta_{21} + x_1)^2} \\ \frac{x_2}{\theta_{21} + x_2} & \frac{-\theta_{11} x_2}{(\theta_{21} + x_2)^2} \\ \frac{x_3}{\theta_{21} + x_3} & \frac{-\theta_{11} x_3}{(\theta_{21} + x_3)^2} \end{vmatrix}$$

$$\text{entonces, } \max_{x_3} |F(\theta_1)^T F(\theta_1)| = \max_{x_3} \sum_{i=1}^3 \left(\frac{x_i}{\theta_{21} + x_i} \right)^2 - \sum_{i=1}^3 \left(\frac{\theta_{11} x_i^2}{(\theta_{21} + x_i)^3} \right)^2$$

de donde, derivando respecto a x_3 e igualando a cero se obtiene la siguiente condición:

$$Ax_3^2 + (2\theta_{21}A - 1)x_3 + A\theta_{21}^2 + \theta_{21} = 0, \quad (15)$$

siendo $A = \frac{-k}{\sum_{i=1}^2 \frac{x_i^2}{(\theta_{21} + x_i)^2}}$ $K = -\sum_{i=1}^2 \frac{\theta_{21} x_i^2}{(\theta_{21} + x_i)^4}$

La solución de la ecuación (15) es la siguiente:

$$x_3 = \frac{2A\theta_{21} + 1 \pm \sqrt{8A\theta_{21} + 1}}{2A}$$

Note que la constante K es negativa, por lo que A es positiva, lo que implica que las raíces de la ecuación (15) son reales puesto que θ_{21} es positivo. El signo de la raíz se elige evaluando la

función $|F(\theta_1)^T F(\theta_1)|$ en x_1, x_2 y x_3 y seleccionando aquel valor de x_3 que la hace.

Generalización del resultado anterior.

Este resultado se generaliza como sigue. Si x_1, \dots, x_{n-1} son puntos de diseño conocidos, y θ_{n-1} es el último vector obtenido con el método, el nuevo punto de diseño x_n se obtiene maximizando la función siguiente sobre x_n :

$$|F(\theta_{n-2})^T F(\theta_{n-2})| \quad (16)$$

donde la matriz $F(\theta_{n-2})$ está dada por:

$$F(\theta_{n-2}) = \begin{vmatrix} 1 & -\theta_{1,n-2} x_1 \\ \frac{x_1}{\theta_{2,n-2} + x_1} & \frac{-\theta_{1,n-2} x_1}{(\theta_{2,n-2} + x_1)^2} \\ \frac{x_2}{\theta_{2,n-2} + x_2} & \frac{-\theta_{1,n-2} x_2}{(\theta_{2,n-2} + x_2)^2} \\ \vdots & \vdots \\ \frac{x_n}{\theta_{2,n-2} + x_n} & \frac{-\theta_{1,n-2} x_n}{(\theta_{2,n-2} + x_n)^2} \end{vmatrix}$$

Haciendo el producto matricial y evaluando el determinante indicado, se obtiene la función que depende de x_n siguiente:

$$\begin{aligned} \text{Max } |F(\theta_{n-2})^T F(\theta_{n-2})| &= \sum_{i=1}^n \left(\frac{x_i}{\theta_{2,n-2} + x_i} \right)^2 \sum_{i=1}^n \left(\frac{\theta_{1,n-2}^2 x_i^2}{(\theta_{2,n-2} + x_i)^4} \right) \\ &\quad - \sum_{i=1}^n \left(\frac{\theta_{1,n-2} x_i^2}{(\theta_{2,n-2} + x_i)^3} \right)^2 \end{aligned}$$

Derivando respecto a x_n , igualando a cero y despejando a x_n se obtiene despues de un buen ejercicio algebráico la siguiente condición:

$$Ax_n^2 + (2\theta_{2,n-2} A - 1)x_n + A\theta_{2,n-2}^2 + \theta_{2,n-2} = 0, \quad (17)$$

siendo $A = \frac{-k}{\sum_{i=1}^{n-1} \frac{x_i^2}{(\theta_{2,n-2} + x_i)^2}}$ $K = -\sum_{i=1}^{n-1} \frac{\theta_{2,n-2} x_i^2}{(\theta_{2,n-2} + x_i)^4}$

Observe que tiene la misma estructura de la ecuación (15). El valor del discriminante de esa ecuación es

$$8A^2\theta_{2,n-2} + 1,$$

el cual es positivo, por lo que las raíces son reales. Eligiendo de ellos el valor de x_n que maximiza la función (16) se obtiene el punto de diseño buscado. En ese punto se realiza un nuevo experimento para obtener la pareja (x_n, v_n) y que junto con las $n-1$ parejas anteriores se realiza la nueva estimación de θ , digamos θ_{n-1} . El procedimiento termina cuando $|\theta_i - \theta_{i-1}| < \delta$, donde $\delta > 0$ es una tolerancia pre-especificada.

5.- Comparación con otras estrategias.

Para realizar la comparación con otros procedimientos se utilizaron los datos del experimento antes indicado (Tabla 1). Se utilizó como punto inicial de búsqueda el punto $\theta_0 = (205, 0.08)$ y la experimentación se simuló considerando que los errores del modelo de Michelis-Menten son Normales con media cero y desviación estandar $\sigma = 0.01$. Se experimentó dos veces en cada punto de diseño y la estimación del modelo se realizó utilizando rutinas numéricas desarrolladas exprofeso para ese tipo de modelos (Burguillo, Velasco y Gaytán, 1989).

La tabla 2 muestra los valores obtenidos por el método propuesto.

Iteración i	No. puntos de diseño n	Puntos de diseño x_n	$\theta_{1,i}$	$\theta_{2,i}$	$ F(\theta_i)^T F(\theta_i) $
0	2	1.1 0.0698	205.000 212.690	0.08 0.064108	622032.81
1	3	0.0830	212.6784	0.064096	1623217.72
2	4	0.098989	212.6668	0.064008	3069999.09
3	5	0.111207	212.66394	0.064078	4968131.66
4	6	0.118648	212.64227	0.064070	7352527.48

Tabla 2

La tabla 3 presenta un resumen de los resultados obtenidos al aplicar los otros métodos reportados en la literatura al ejemplo discutido antes y el método propuesto.

Método	Punto Inicial θ_0	No. punto del diseño	Estimación Final θ	No. de repeticiones	No. total de ensayos
B-W	no dispon.	2	(212.7, 0.0641)	6	12
Treloar	no dispon.	6	(212.7, 0.068)	2	12
Propuesto	(205, 0.08)	3	(212.678, 0.0641)	2	6
Lineal	(205, 0.08)	6	(212.246, 0.0665)	2	2
Logarítmico	(205, 0.08)	6	(212.791, 0.641)	2	12

Tabla 3

Conclusiones.

1.- Resultados obtenidos sugieren que los parámetros se estiman con un número menor de experimentos y lo cual claramente es una ventaja si el costo del experimento es alto.

2.- Se desarrolló una forma general para obtener puntos de diseño secuencial mediante el criterio D-OPTIMO lo cual aparentemente no ha sido reportado en la literatura.

3.- El nuevo método cuando se aplicó a un ejemplo mostró ser tan bueno como el mejor disponible, aunque con un número menor total de ensayos.

4.- Una actividad futura deberá extender las ideas de este trabajo a modelos no Michaelianos como (2:2), (3:3), (4:4), (1:2) y (2,3).

Bibliografía.

Bates, D., 1983, "The derivative of $|X^T X|$ and its uses" Technometrics, 25, 373.

Bates,D. and Watts,D.G., 1988, "Nonlinear regression analysis and its applications", Wiley,.

Box, G.E.P., Hunter, W.G., Hunter, J.S., 1978, "Statistics for experimenters", Wiley.

Velasco F.G., y Gaytan J., 1989, "Non Michaelian enzyme kinetics microcomputer", Symposium Compute Assisted Learning 1989, University of Surrey Guildford, Inglaterra, 1989.

Memoria del VII Foro Nacional de Estadística, edición financiada parcialmente por el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la UNAM. Se terminó de imprimir el mes de septiembre de 1993, en los talleres de Proyección Creativa Papel, S.A. de C.V., México, D.F., con un tiraje de 500 ejemplares.