UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

COLEGIO DE CIENCIAS Y HUMANIDADES

Unidad Académica de los Ciclos Profesional y de Posgrado

1990

M E M O R I A S

D E L

I V F O R O D E E S T A D I S T I C A

MEMORIAS
DEL
IV FORO DE ESTADISTICA

EDITADAS POR:

FERNANDO AVILA
Universidad de Sonora

VICTOR M. GUERRERO
ITAM

VICTOR M. PEREZ-ABREU
CIMAT

JOSE A. VILLASEÑOR
Colegio de Postgraduados-Chapingo

BAJO LOS AUSPICIOS DE:

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

UNIDAD ACADÉMICA DE LOS CICLOS PROFESIONAL Y DE POSGRADO

COLEGIO DE CIENCIAS Y HUMANIDADES

1990

C O N T E N I D O

IV FORO DE ESTADISTICA

PRESENTACION

En este volumen se incluyen los artículos aceptados por el Comité Editorial de las Memorias del IV Foro de Estadística. Por primera vez en la historia de este evento los artículos recibidos para su posible publicación fueron sometidos a un proceso de arbitraje en el que participó la comunidad estadística de México, y cuyo propósito fue depurar y mejorar la calidad de las contribuciones que finalmente aparecen en este volumen.

Debido a la diversidad de los temas tratados en los trabajos aceptados, éstos aparecen en orden alfabético por apellido del primer autor.

El Comité editorial hace patente su reconocimiento a los colegas que anónimamente participaron con entusiasmo en el proceso de arbitraje, lo cual contribuyó a la selección y mejoramiento de los artículos. Asimismo, se agradece a la Unidad Académica de los Ciclos Profesional y de Posgrado del CCH de la UNAM su apoyo para la impresión de este volumen.

El Comité Editorial de las
Memorias del IV Foro de Estadística

Junio de 1990

ANTECEDENTES

Este evento tiene sus orígenes en el Primer Foro de Estadística Aplicada que se realizó los días 24, 25 y 26 de septiembre de 1986 en la Unidad de Seminarios "Dr. Ignacio Chávez" de la Universidad Nacional Autónoma de México. En aquella ocasión, el evento fue organizado por la Coordinación de la Especialización en Estadística Aplicada de la UACPyP del CCH, con el apoyo de la Facultad de Ciencias y el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la UNAM, presentándose 32 ponencias y una mesa redonda; estuvieron representadas, además, 18 instituciones de educación superior y de investigación tanto del sector público como privado. En las memorias del Primer Foro aparecen 14 trabajos publicados que fueron seleccionados en función de su originalidad y del tipo de técnica estadística utilizada. En la mesa redonda "La Estadística en México: sus Problemas y sus Perspectivas", tanto los organizadores como los participantes se propusieron trabajar para realizar un Segundo Foro en el que se abrieran posibilidades de participación a un número mayor de instituciones del sector público y privado y centros de investigación, especialmente de provincia.

El Segundo Foro se realizó en el Auditorio del Jardín Botánico de la Ciudad Universitaria de la UNAM, en el mes de octubre de 1987, e incluyó tres conferencias magnas y 27 ponencias en áreas más diversas que las del Primer Foro. Contó, por primera vez, con la asistencia de investigadores y estudiantes provenientes de universidades y centros de investigación de provincia. En las memorias del Segundo Foro aparecen 10 trabajos que reflejan parcialmente el tipo de investigación básica y aplicada que se realiza en el medio estadístico en México. En esa ocasión, se contó con el apoyo de la Coordinación de la Especialización en Estadística Aplicada de la UACPyP/CCH, el Departamento de Probabilidad y Estadística del IIMAS y el Departamento de Matemáticas de la Facultad de Ciencias de la UNAM, así como con el de la Sociedad Matemática Mexicana.

Bajo los auspicios del Centro de Investigación en Matemáticas, A. C. y de la Universidad de Guanajuato, en el mes de septiembre de 1988 se realizó en la ciudad de Guanajuato el Tercer Foro de Estadística. El evento tuvo un registro de 47 ponencias repartidas en 15 sesiones sobre diferentes áreas de investigación básica y aplicada de la Estadística en nuestro país. En él participaron profesionales y estudiantes de distintas instituciones, entre las que se encuentran CIDE, CIMAT, UNAM (IIMAS, FC, FQ, UACPyP y ENEP-Acatlán),

Universidad Autónoma del Estado de México, Universidad Veracruzana, Banco de México, Centro de Estadística y Cálculo de Chapingo, IPN, Universidad de Colima, INIFAP-Tabasco, CIFAP-D.F., Universidad Autónoma de Sinaloa, Centro de Investigación en Salud Pública, Secretaría de Salud, ITAM, Universidad de Sonora, Universidad de Guanajuato, Universidad Autónoma Metropolitana, Instituto Tecnológico de Celaya, Instituto Tecnológico y de Estudios Superiores de Monterrey, Universidad Autónoma de Nuevo León, Universidad Autónoma de Aguascalientes, Instituto Mexicano del Petróleo. Las memorias del Tercer Foro contienen 20 artículos.

La sede del IV Foro de Estadística fue otorgada a la Facultad de Ciencias Físico-Matemáticas de la Universidad Autónoma de Nuevo León y se llevó a cabo del 11 al 14 de septiembre de 1989. Su realización fue posible gracias a los esfuerzos y apoyos de diversas instituciones: la Subsecretaría de Educación Superior e Investigación Científica de la SEP, el Consejo Nacional de Ciencia y Tecnología, la Universidad Autónoma de Nuevo León y el Centro de Investigación en Matemáticas. El programa del IV Foro de Estadística presentó algunas diferencias con los programas de los Foros anteriores. Aparte de las contribuciones libres de reportes de investigación básica, aplicada y de tesis (previa selección del comité académico) que han sido la base de estos eventos, el programa del IV Foro de Estadística incluyó cuatro **Conferencias Especiales por Invitación**, impartidas por destacados especialistas mexicanos en el área. También se incluyeron dos Talleres Especiales sobre áreas de aplicación de la Estadística de importancia actual para México: el **Taller de Geoestadística** y el **Taller de Control de Calidad.**

La participación de estudiantes en los Foros de Estadística ha sido numerosa e importante. Con el fin de canalizar esta participación, dentro del programa del IV Foro se incluyeron **cuatro cursos** dirigidos tanto a estudiantes como a investigadores e impartidos por destacados expositores.

El programa contempló también una **Mesa Redonda** sobre **"La Estadística en México"**, así como una Asamblea de la Asociación Mexicana de Estadística. Con este programa se cubrieron los objetivos de difundir la Estadística y sus aplicaciones, promover el intercambio de experiencias y la unión de los profesionales de la Estadística en México. El IV Foro de Estadística contó con la inscripción de 123 participantes de distintas instituciones, tanto nacionales como extranjeras.

Los foros de Estadística han permitido un intercambio de experiencias entre los profesionales del área que teniendo un dominio similar de

conocimientos pertenecen a diferentes instituciones nacionales. También han sido parte complementaria de la formación de los estudiantes que en ellos han participado. Estos eventos se han convertido en un medio natural en el que se han reflejado las diferentes áreas de investigación básica de la estadística en México y la metodología estadística utilizada en nuestro país en áreas y sectores tan diversos como Antropología, Biología, Economía, Ecología, Demografía, Agricultura, Medicina y Salud, Química, Industria, etc. La difusión de la cultura estadística y el promover la unión de los estadísticos del país han sido objetivos alcanzados por estos Foros.

Comité Organizador
IV Foro de Estadística

UNA COMPARACION ENTRE DOS INDICES CLINIMETRICOS PARA CLASIFICAR PACIENTES CON INFARTO AGUDO DEL MIOCARDIO

Javier Alagón Cano[1]
Antonio Villa-Romero[2]
Cecilia García-Sancho[3]

[1]Instituto Tecnológico Autonómo de México, Río Hondo 1, Tizapán San Angel, México D.F., 01000

[2]Instituto Nacional de la Nutrición "Salvador Zubirán", Vasco de Quiroga 15, Tlalpan, México, DF, 14000

[3]Instituto Nacional de Salud Pública, Francisco de P Miranda 177, Merced Gómez, México, DF, 01480

RESUMEN

Este artículo presenta una comparación empírica entre dos índices de clasificación de pacientes con infarto agudo al miocardio, el índice de Norris y la escala de Killip. Los datos fueron recabados en el Instituto Nacional de Cardiología de la Ciudad de México para el período de tiempo 1975 - 1985. El análisis incluye medidas de correlación y validación de ambos índices así como los riesgos de muerte estimados para cada uno. Los dos índices muestran poderes predictivos de muerte similares. Sin embargo, dado el diseño de casos y controles utilizado en este estudio, el poder predictivo del índice de Norris es disminuido. Por tanto, es de esperarse que dicho índice sea mejor en la clasificación de pacientes con infarto agudo al miocardio.

INTRODUCCION

Existe un uso difundido de índices clinimétricos en las diferentes especialidades médicas [1]. En Cardiología, dos de estos índices han sido ampliamente utilizados con la finalidad de evaluar el pronóstico de pacientes con infarto agudo del miocardio (IAM). Dichos índices son conocidos como la **escala de Killip** [2] y el índice de pronóstico coronario o **índice de Norris** [3]. Los dos índices estiman la gravedad del IAM. El trabajo reportado en este artículo sigue la línea establecida por Horwitz *et al* [4], quienes fueron los primeros en hacer una comparación empírica entre los dos índices.

El objetivo principal de este artículo es comparar la eficiencia de ambos índices, al respecto de su poder predictivo entre pacientes con IAM. El artículo proporciona otra perspectiva al problema de selección de escalas de severidad de infartos al miocardio, en el sentido de que se utiliza un grupo de pacientes, en un lugar y en un período de tiempo, distintos a los analizados por Horwitz *et al*. Un objetivo secundario es la estimación de la concordancia estadística entre los dos índices, basados en el uso de métodos estadísticos multivariados. El trabajo de este artículo formó parte de un proyecto de investigación llevado a cabo en el Instituto Nacional de Salud Público, dirigido a evaluar una terapia anticoagulante en pacientes con AMI [5].

El índice de Norris está constituido por un conjunto de seis variables: edad, localización anatómica del infarto, tensión arterial sistólica en el momento de la admisión, tamaño

del corazón, estado de los campos pulmonares y antecedente de infarto o isquemia previos (de tal forma, las seis variables son simbolizadas respectivamente por $X_1, X_2, \ldots, X_6$). Así, una combinación lineal de las variables determina el valor del índice de Norris para cada paciente:

$$N(X) = a_1X_1 + a_2X_2 + \ldots + a_6X_6$$

donde X_1 a X_6 denotan los valores de las seis variables de cada paciente y a_1 a a_6 es el conjunto de pesos que representan la importancia relativa de cada variable. Estos pesos fueron determinados por la aplicación de la técnica estadística conocida como *análisis discriminante* en la muestra total de pacientes.

Por otra parte, la construcción de la escala de Killip es algo menos elaborada. Es más utilizada que el índice de Norris ya que incluye la valoración clínica del paciente mediante signos y síntomas principalmente referidos a corazón izquierdo. La escala de Killip posee una naturaleza ordinal con cuatro categorías:

(I) Sin insuficiencia cardíaca, es decir, ausencia de signos clínicos de descompensación cardíaca,

(II) Insuficiencia cardíaca moderada, definida por la presencia de soplos, galope protodiastólico e hipertensión venosa,

(III) Insuficiencia cardíaca grave, determinada por la presencia de edema agudo pulmonar y,

(IV) Choque cardiogénico, incluyendo hipotensión (tensión

arterial sistólica menor de 90 mmHg) y evidencia de vasoconstricción periférica por oliguria, cianosis o diaforesis.

METODOS

Los datos del estudio fueron recabados en el Instituto Nacional de Cardiología "Ignacio Chávez" de México para el período comprendido entre enero de 1975 y diciembre de 1985. En total fueron incluidos 424 pacientes con diagnóstico de IAM hospitalizados durante el período. El diagnóstico de IAM se basó en:

i) Datos clínicos. Obtenidos mediante la semiología descrita y clasificados como: típicos (dolor retroesternal de tipo opresivo, angustiante, con duración mayor de 20 minutos y con irradación a sitios característicos), atípicos (dolor retroesternal menor de 20 minutos y/o con irradiación a sitios poco comunes como mandíbula, cuello, espalda o brazo derecho), y mal descritos (con semiología característica pero sin duración reportada).

ii) Datos electrocardiográficos. Definidos por presencia de onda Q profunda y/o empastada, inversión de onda T, elevación del segmento ST y/o complejos QS.

iii) Datos de laboratorio. Basados en el aumento de cualquiera de las tres siguientes enzimas séricas: Creatín-fosfoquinasa (CPK), deshidrogenasa láctica (LDH) y transaminasa glutámico-oxalacética (GOT).

Con la finalidad de que un paciente quedara incluido en el estudio, debía cumplir con al menos dos de los tres criterios previos (clínicos, electrocardiográficos o de laboratorio). Se conformaron dos grupos de pacientes: los casos, con 212 pacientes con IAM y que fallecieron durante la hospitalización, y los controles, con 212 pacientes con IAM que sobrevivieron al episodio durante el internamiento. Cada uno de los casos fue pareado con un control de acuerdo con tres variables potencialmente confusoras: edad (dentro de un rango de cinco años), sexo (el mismo sexo), y fecha de hospitalización (dentro de un rango de 13 meses). Esta última variable se refería específicamente a la fecha con la que ingresó un paciente al Instituto referido por el cuadro de IAM.

Para lograr una potencia de 0.90 (es decir, $\beta = 0.10$) con un riesgo relativo mínimo detectable igual a 2.5, se requieren 194 pares de sujetos en un diseño de casos y controles pareado individualmente. Este cálculo puede ser fácilmente estimado mediante la fórmula para tamaños de muestra en estudios de casos y controles pareados individualmente referida por Schlesselman [6] (fórmula 6.20, página 161). De tal forma, el tamaño de muestra anterior fue superado hasta lograr 212 pares de sujetos, dado que la información para 18 casos estaba accesible al momento de llevar a cabo la recolección de los datos. Así, el poder para detectar un riesgo relativo mínimo de 2.5 fue ligeramente incrementado.

Los dos índices clinimétricos fueron evaluados en cada paciente. Con el fin de estandarizar los criterios de

clasificación, dos cardiólogos fueron específicamente adiestrados para el propósito. Los datos utilizados en la construcción de los índices fueron tomados de los expedientes clínicos de los pacientes del Instituto de Cardiología.

Los dos índices fueron correlacionados mediante el coeficiente **gamma,** una medida de concordancia estadística bien conocida [7]. Para lograr tal propósito fue necesaria la categorización del índice de Norris. Con la finalidad de evaluar el poder discriminante para la predicción de la muerte en cada índice, se realizó un análisis discriminante (a través del uso del paquete SPSS-X en una VAX/VMS). Mediante análisis de regresión logística se obtuvieron los riesgos de muerte según los dos índices (usando el paquete EGRET para una IBM-PC).

RESULTADOS

Debe ser señalado en primer lugar que, como resultado del pareamiento, no hubo diferencias estadísticamente significativas entre los dos grupos (casos y controles) con respecto a las variables pareadas. En términos simples, cerca del 60 por ciento de los pacientes en el estudio fueron hombres, sus edades variaron entre 28 y 98 años con una media de 67 años. La distribución de frecuencias entre las diferentes categorías según los dos índices, se muestra en los cuadros I y II. Es evidente que una gran proporción de muertes (casos) ocurrieron entre aquellas categorías de mayor riesgo en los dos índices.

CUADRO I

Distribución de casos y controles de acuerdo a la escala de Killip

GRADO DE SEVERIDAD	CASOS No.	%	CONTROLES No.	%	TOTAL No.	%
TIPO I	59	27.8	134	63.2	193	45.5
TIPO II	94	44.3	76	35.8	170	40.1
TIPO III	14	6.6	0	0.0	14	3.3
TIPO IV	45	21.2	2	1.0	47	11.1
TOTAL	212	100.0	212	100.0	424	100.0

$\chi^2 = 80.8$, 3 g.l. $p < 0.00001$

CUADRO II

Distribución de casos y controles de acuerdo al índice de Norrris

GRADO DE SEVERIDAD	CASOS No.	%	CONTROLES No.	%	TOTAL No.	%
< 4	1	0.5	9	4.2	10	45.5
4-5.99	21	9.9	33	15.6	54	40.1
6-7.99	47	22.2	78	36.8	125	3.3
8-9.99	58	27.4	65	30.7	123	11.1
10-11.99	30	14.1	23	10.8	53	12.5
> 12	55	25.9	4	1.9	59	13.9
TOTAL	212	100.0	212	100.0	424	100.0

$\chi^2 = 62.2$, 5 g.l. $p < 0.00001$

La concordancia de los dos índices fue calculadacon tres codificaciones distintas del índice de Norris: la primera categorización se hizo con cuatro estratos siguiendo los puntos de corte naturales en la distribución de frecuencias; la segunda categorización se hizo con base en la división por cuartiles del rango y, la tercera categorización se estableció con los cinco puntos de corte originalmente propuestos por Norris y cols. [3]

Sin embargo, con fines de simplificación se presentan sólo los resultados con los puntos de corte naturales y con los puntos de corte originales. En el Cuadro III se presenta un cruce de la distribución de pacientes según la escala de Killip y el índice de Norris y la concordancia numérica estimada. De acuerdo con los valores de la estadística de concordancia empleada (gamma = 0.63 y 0.69) el nivel de concordancia entre los dos índices es alto. Los valores de la chi cuadrada reportados en el Cuadro III, reflejan que hay una fuerte evidencia estadística de asociación entre ambos índices.

El análisis discriminante realizado con cada índice (tomando muerte y sobrevida como las dos categorías de la variable dependiente, y los valores de los dos índices para cada paciente como las variables independientes) mostró que los dos índices tiene proporciones similares de clasificaciones incorrectas. En la tabla IV se presentan los resultados del análisis discriminante en términos de la validez de cada índice (sensibilidad y especificidad) así como el poder predictivo. Hay una cercana similitud entre los resultados de la clasificación para la escala de Killip y el índice de Norris categorizado en cuatro estratos naturales. Sin embargo, como el índice de Norris incluye a la edad como uno de sus predictores y el efecto de esta variable está fuertemente atenuado por el diseño pareado, la capacidad de clasificación del índice de Norris se ve disminuída. Esto significa que el índice de Norris puede tener una mejor capacidad de discriminación que la aquí reportada.

CUADRO IV

Validez y valores predictivos para cada índice obtenidos con análisis discriminante

		KILLIP		NORRIS NATURAL		NORRIS ORIGINAL	
		RESULTADO					
		VIVO	*MUERTO*	*VIVO*	*MUERTO*	*VIVO*	*MUERTO*
PREDICCION	*VIVO*	*134*	*59*	*134*	*58*	*167*	*89*
	MUERTO	*78*	*153*	*78*	*148*	*45*	*117*
Sensibilidad (Se)		63.2 %		63.2 %		78.8 %	
Especificidad (Esp)		72.2 %		71.8 %		56.8 %	
(Se + Esp) / 2		67.7 %		67.5 %		67.8 %	
+ Valor predictivo (Vp+)		69.4 %		69.8 %		65.2 %	
- Valor predictivo (Vp-)		66.2 %		65.5 %		72.2 %	
[(Vp+) + (Vp-)] / 2		67.8 %		67.7 %		68.7 %	

Con la intención de incorporar el efecto del pareamiento en el análisis, se llevó a cabo un análisis de regresión logística condicional[1]. En términos de la predicción de muerte, los resultados son casi los mismos que aquellos obtenidos por medio del análisis discriminante: ambos índices tienen aproximadamente los mismos errores de clasificaciones incorrectas que aquellos obtenidos con el análisis

[1]*El modelo empleado en el análisis de regresión logística fue condicional con una razón de pareamiento uno a uno entre casos y controles. Las categorías de los índices fueron incluídas en forma ordinal para el modelo, asumiendo que los riesgos de muerte tienen incrementos constantes entre dos categorías subsecuentes. Así, cuando la escala de Killip se usó como variable independiente, los valores para las cuatro categorías fueron 1,2,3 y 4; cuando el índice de Norris fue recodificado en cuatro o seis categorías, los valores fueron 1,2,3,4 y 1,2,3,4,5,6 respectivamente.*

discriminante[2]. A pesar del hecho de que las variables independientes del modelo logístico no son normales, la regresión logística no mejora los resultados del análisis discriminante como podría esperarse (la superioridad de la regresión logística sobre el análisis discriminante cuando las variables independientes no son normales con iguales matrices de covarianza como procedimiento de clasificación ha sido ilustrada por Press y Wilson [8]; y Fienberg [9], sección 6.5).

Los riesgos estimados de muerte por medio del modelo logístico según los dos índices, se presentan en el Cuadro V. Asimismo, se muestran los riesgos ajustados en el mismo cuadro. (el ajuste se hizo de un índice por el otro). Los valores de los riesgos de muerte indican, por ejemplo, que para la escala de Killip un cambio de una categoría (de I a II o de III a IV) aumenta el riesgo de muerte por casi 3 y media veces, mientras que en el índice de Norris (original) este aumento es menor de dos veces.

DISCUSION

Si bien, el nivel de concordancia entre los dos índices es alto, los índices no pueden ser utilizados indistintamente. Los poderes predictivos para los dos índices fueron similares (cerca del 68 por ciento). Sin embargo, el diseño de casos y controles pareado por edad (así como sexo y fecha de internamiento) necesariamente disminuye el poder predictivo del

[2] *El procedimiento de clasificacion basado en el modelo logistico ajustado se hizo de acuerdo con la probabilidad de que un paciente fuera caso (ver Schlesselman [6] formula 8.32)*

CUADRO V

Riesgos relativos estimados (Exp Beta) para los dos índices clinimétricos

	Exp(Beta)	*valor p*
Escala de Killip	3.53	< 0.001
Indice de Norris (natural)	2.50	< 0.001
Indice de Norris (original)	1.81	< 0.001
Killip controlado por Norris (natural)	3.13	< 0.001
Norris (natural) controlado por Killip	1.36	0.110
Killip controlado por Norris (original)	2.94	< 0.001
Norris (original) controlado por Killip	1.31	0.008

índice de Norris dado que la variable edad está especificamente incluída en la construcción de dicho índice. Esto lleva a pensar en una mejor capacidad del índice de Norris para la clasificación de pacientes con IAM. Los estudios que se lleven a cabo en un futuro sobre evaluación de pacientes con IAM deberán intentar el incluir el efecto de la variable edad en la clasificación del índice de Norris.

Los valores de los riesgos de muerte para cada índice (ajustados uno por otro) son 2.94 para la escala de Killip y de 1.31 para el índice de Norris (original). Esto revela un mayor efecto entre los cambios de categorías para la escala de Killip

(derivado de los diferentes puntos de corte para los dos índices) aún cuando los resultados de la clasificación global son casi los mismos para los dos índices.

Los resultados en este trabajo pueden ser contrastados con los obtenidos por Horwitz y cols. [2]. Ellos estimaron un valor de 0.67 para el coeficiente de correlación de Jaspen entre los dos índices. Esta estadística asume una variable continua (Norris) y una variable ordinal (Killip). Este valor tiene aproximadamente la misma magnitud que la obtenida por nosotros mediante el coeficiente gamma (0.63 y 0.69). Ellos también obtuvieron un coeficiente kappa de 0.47 entre los dos índices, cuando el índice de Norris lo recodificaron en cuatro categorías; en el presente estudio el coeficiente kappa fue calculado con el índice de Norris (natural) dando un valor de 0.20 (debe así ser notado, que los valores de la concordancia dependen del tipo de variable - continua u ordinal - utilizada en el cálculo del coeficiente de asociación).

Con el propósito de aumentar la validez de los dos índices, una estrategia alternativa de análisis podría ser el usar dichos índices como **pruebas en serie** [10]. Bajo este esquema, los pacientes que inicialmente se diagnostican con pronóstico favorable (vivirán) con alguno de los índices, son posteriormente clasificados con el otro índice. Esto lleva a mejorar la especificidad hasta el 88.7 por ciento, lo cual representa el porcentaje de pacientes que han sido correctamente identificados como muertos bajo la combinación de los dos índices.

REFERENCIAS

[1] Feinstein A. R. (1987) *Clinimetrics* Yale University Press New Haven.

[2] Killip T. & Kimball J. T. (1967) *Treatment of myocardial infarction in a Coronary Care Unit. A two year experience with 250 patients.* **Am J Cardiol 20:** 457-464.

[3] Norris R. M., Brandt P. W. T., Caughey D. E. et al (1969) *A new coronary prognostic index.* **Lancet, Feb 8:** 274-278.

[4] Horwitz R. I., Cicchetti D. V. & Horwitz S. M. (1984) *A comparison of the Norris and Killip Coronary Prognostic indices.* **J Chron Dis 37(5):** 369-375.

[5] Villa-Romero A. (1987) *Uso de anticoagulantes y riesgo de morir en pacientes con infarto agudo del miocardio.* **M Sc thesis** National Institute of Public Health, Mexico City.

[6] Schlesselman J. J. (1982) *Case-control studies. Design, Conduct and Analysis.* Oxford University Press. New York.

[7] Everitt B. S. (1977) *The analysis of Contingency Tables.* **Chapman and Hall, Ltd.** London.

[8] Press S. J. and Wilson S. (1978) *Choosing between logistic regression and discriminant analysis.* **J. Amer. Statist. Assoc. 73:** 699-705.

[9] Fienberg S. E. (1980) *The Analysis of Cross-Classified Categorical Data.* **The MIT Press**, Cambridge, Massachusetts.

[10] Fletcher R. H., Fletcher S. W. & Wagner E. H. (1985) *Clinical Epidemiology. The essentials.* **William and Wilkins.** Baltimore.

ELABORACION DE UN PRODUCTO TIPO JAMON COCIDO A PARTIR DE CARNES NO CONVENCIONALES (Conejo, carnero y pollo)

Arroyo López, Pilar E. y Muciño Yáñez, Maricela

Facultad de Química de la U.A.E.M.

Apartado postal no. 20, Z.P. 50,000. Toluca, Méx.

RESUMEN

Empleando carne de cerdo en combinación con carnes no convencionales (conejo, carnero y pollo), se elaboró un jamón cocido, en cuyo proceso de elaboración se emplearon diferentes combinaciones de nitritos y fosfatos. Los factores porcentaje de carne, cantidad de nitritos y cantidad de fosfatos, se propusieron a tres niveles, empleándose un diseño de bloques de tamaño tres para efectuar el experimento. Como variables de respuesta se emplearon las resultantes de una evaluación sensorial las cuales fueron analizadas mediante un análisis de varianza, determinándose en base a éste el producto de mayor aceptación para los potenciales consumidores.

INTRODUCCION

En los últimos años, el empleo de métodos estadísticos en la conducción de experimentos se ha difundido a todas las áreas. En revistas norteamericanas, como el Journal of Quality Control e Industrial Chemical Engineering, se tienen reportes sobre el empleo del diseño experimental en la industria química, el International Journal of Pharmaceutics reporta aplicaciones en farmacia y toxicología, y el Journal of Food Technology en el área de alimentos. Estos reportes muestran las ventajas que un buen

experimento proporciona en el análisis de resultados y en el número de experimentos necesarios para obtener información sobre el efecto de distintos factores en la calidad de un producto.

En el área de alimentos, el número 25 del Journal of Food Technology muestra una idea general del tipo de métodos estadísticos aplicables al análisis de experimentos con productos alimenticios, entre estos métodos se cuenta al Análisis de Varianza, Regresión (Korth, 1982), Superficies de Respuesta (Henika, 1982 y Myers, et. al., 1988) , Análisis de cúmulos y Análisis de correspondencia (Ennis, et. al., 1982). También se discuten las dificultades encontradas en la aplicación de tales métodos sobre todo para variables de respuesta resultantes de una evaluación sensorial empleando un panel de jueces (O'Mahony, 1982).

En este reporte, se presenta un experimento efectuado en planta piloto, el cual tuvo como objetivo desarrollar un jamón con un grado de aceptación apropiado y elaborado con carnes de mayor valor nutritivo que la carne de cerdo. En la primera parte, se presentan los factores en estudio y el tipo de diseño experimental usado, en la segunda parte, se mencionan las variables de respuesta analizadas. La tercera parte es la discusión de resultados en base a los análisis de varianza efectuados, los experimentos confirmatorios se dan en la cuarta sección del reporte y finalmente en una quinta sección las conclusiones y recomendaciones.

FACTORES EN ESTUDIO Y DISEÑO EXPERIMENTAL

En el experimento, se consideraron a los siguientes factores como los más importantes para determinar la calidad del jamón:

A = porcentaje de carne no convencional en el producto

Tres niveles: 25, 50 y 75% de carne de conejo, carnero ó pollo, el resto para tener 100% es carne de cerdo.

B = cantidad de fosfatos por kg. de producto.

Tres niveles: 2850, 3000 y 3150 mg. por kg. de carne.

C = cantidad de nitritos por kg. de producto.

Tres niveles: 50, 200 y 350 mgl por kg. de carne.

Por razones de economía y considerando que no más de 4 kgs. de carne pueden ser procesados por corrida en el equipo piloto, se optó por emplear lotes de 3 kgs. de carne, los cuales fueron divididos en tres porciones de 1 kg., que se procesaron a diferentes combinaciones de nitritos y fosfatos. El factorial 3^2 formado por los factores B y C se confundió en bloques de tamaño tres, empleando la parte BC^2 = J(BC) de la interacción BxC para generarlos, de tal manera que las tres combinaciones en cada uno de los bloques formados pudieran ser corridas empleando un lote de 3 kgs. con un porcentaje de carne no convencional específico. La asignación de factores al experimento muestra una forma análoga a un cuadro latino, con los nitritos en las columnas, los fosfatos en los renglones y el porcentaje de carne no convencional a las letras en el cuadro (Davis, 1978):

Combinaciones		Nitritos			
lote a_o: b_oc_o, b_1c_1, b_2c_2			0	1	2
lote a_2: b_oc_1, b_1c_2, b_2c_o	Fosfatos	0	a_o	a_2	a_1
lote a_1: b_oc_1, b_1c_2, b_2c_o		1	a_1	a_o	a_2
		2	a_2	a_1	a_o

En este arreglo, los niveles del factor A están confundidos con los bloques así como parte de la interacción BxC. Las

interacciones entre factores se asumieron despreciables en base al hecho de que nitritos y fosfatos tienen efectos sobre diferentes propiedades del producto tipo jamón. Los nitritos permiten el desarrollo del color rosado y dan sabor característico al jamón, en tanto los fosfatos controlan el pH y aumentan la capacidad de retención de agua en el producto. Con el propósito de tener una estimación del error para probar el efecto del factor A, dos réplicas del experimento descrito se corrieron en forma independiente.

VARIABLES DE RESPUESTA

Como variables de respuesta se obtuvieron las siguientes:

Evaluación sensorial: se incluyen textura, consistencia, color, olor, sabor, apariencia, sal y especias.

Análisis bromatológicos: humedad, cantidad de grasa, cantidad de proteínas, cantidad de nitritos y fosfatos en producto terminado.

Respecto a los análisis bromatológicos, las observaciones se registraron únicamente con fines de control del producto, dado que existiendo diferencias notables en la cantidad de proteínas y grasa en las carnes de pollo, conejo, carnero y cerdo, esto determinará definitivamente las cantidades observadas en el jamón.

Respecto a la evaluación sensorial, la adición de sal y especias se mantuvo constante en los productos elaborados y los valores obtenidos en la evaluación sensorial se emplearon básicamente para corregir por la adición al inicio del proceso. Para la evaluación sensorial se usó un grupo de cincuenta jueces no entrenados, a cada uno de ellos se les dieron a probar los 18 productos elaborados con cada tipo de carne y se les pidió calificaran en una escala de 0-100 la textura, consistencia y

demás caracterísstica de los productos (Merolli, 1980). El promedio de estas calificaciones constituyeron las variables de respuesta analizadas en base al análisis de varianza, aún cuando la variable es discreta, el uso del promedio de 50 observaciones si satisface la suposición de normalidad para el análisis. La interacción de tratamientos y jueces se considera despreciable en las evaluaciones sensoriales y por consiguiente no se consideró en el ANDEVA.

ANALISIS DE RESULTADOS

La forma general de la tabla de ANDEVA para cada tipo de carne, mostrando fuentes de variación y grados de libertad tiene la siguiente forma:

Fuente	gl
Repeticiones	1
A = % de carne	2
Error a = RxA	2
SUBTOTAL	5
Nitritos	2
Fosfatos	2
Error b	8
TOTAL	17

Tablas de ANDEVA para las seis variables relevantes en la evaluación sensorial fueron construídas con la información resultante. En la tabla I se muestran tres de estos cuadros como ejemplo. Considerando que sólo hay dos grados de libertad para estimar el error (a), se propuso un nivel de significancia de 0.1 para probar el efecto del factor A y de 0.05 para los factores B y C. En los cuadros de la tabla, el asterisco (*) indica los efectos

declarados significantes.

En el caso de conejo y carnero, para todas las variables de respuesta se declararon diferencias entre los tres porcentajes de carne empleados, favoreciéndose en general el 25% de carne no convencional y 75% cerdo. Respecto a pollo, las diferencias entre porcentajes de carne se observaron sólo para las variables apariencia y color, lo cual se debe al color cremoso en la carne de pollo. Una variación considerable entre las dos réplicas del experimento también fué notada.

Respecto a las cantidades de nitritos y fosfatos, no se obtuvo significancia para los siguientes casos:

POLLO respecto a consistencia y apariencia en el caso de los fosfatos.

CONEJO respecto al color.

CARNERO respecto a consistencia y aparte el sabor en el caso de los fosfatos.

En la tabla II se muestran los promedios de calificaciones otorgadas a las características del producto para todos aquellos casos en que se obtuvieron diferencias significantes entre las cantidades de nitritos y fosfatos.

Dado que una única cantidad de nitritos ó fosfatos desafortunadamente no maximiza las seis variables simultáneamente, se decidió elegir como cantidad óptima aquella que maximice el mayor número de respuestas. Para el producto con pollo, se propone la combinación b_1c_2, para conejo b_2c_2 y para carnero b_0c_0.

EXPERIMENTOS CONFIRMATORIOS

Se prepararon tres lotes con las tres mejores formulaciones definidas para pollo, conejo y carnero, en la proporción 25% de

estas carnes y 75% cerdo, además de tres lotes con 100% carne de cerdo, los resultados para aceptación total (promedio de los resultados de la evaluación sensorial) son los siguientes:

	X	CV
25% pollo, b_2c_1	58.5	32
25% conejo, b_2c_2	60.4	26
25% carnero, b_0c_0	57.2	28
100% cerdo	79.2	17

CONCLUSIONES Y RECOMENDACIONES

Todos los productos elaborados con carnes no convencionales tuvieron mayor contenido de proteína y menos grasa que el jamón 100% cerdo (los productos se calificaron como regulares), sin embargo el jamón con alto porcentaje de carne de cerdo fué el más aceptado por el potencial consumidor.

Diferentes cantidades de nitritos y fosfatos deben ser usadas dependiendo del producto, habiendo la necesidad de invesstigar otras variables en el proceso que permitan mejorar el producto y hacerlo comparable al de 100% cerdo. Considerando el factor económico en la elaboración de los productos, es factible elaborar aquellos con 25% conejo ó pollo y el 50% pollo. Para el producto con 75% pollo sería difícil el eliminar el efecto del color.

Se observó alta variabilidad entre los lotes elaborados, lo que también hace necesaria experimentación adicional para controlarla.

El experimento es de tipo exploratorio, el mínimo número de grados de libertad para estimar las fuentes de error (especialmente para % de carne), no da una alta confiabilidad a las pruebas de hipótesis efectuadas, las formulaciones definidas

se consideran la base para una investigación respecto al establecimiento de un proceso estandar de elaboración de jamón cocido a partir de carnes no convencionales.

REFERENCIAS

1.- Davies, O.L.(editor) (1978). "The Design and Analysis of Industrial Experiments." Longman Group Limited. Pag. 396-400.

2.- Ennis, D.M., Boelens, H., Haring, H., and Bowman, P. (1982). "Multivariate Analysis in sensory evaluation." Food Technology. Vol. 36, 83-90.

3.- Henika, R.G. "Use of response-surface methodology in sensory evaluation." Food Technology. Vol. 36, 96-101.

4.- Korth, B. "Use of regression in sensory evaluation." Food Technology. Vol. 36, 91-95.

5.- Merolli, A. "Sensory Evaluation in Operations." Food Technology. Vol. 34, 104-108.

6.- Myers, R.H., Khurl, A.I., and Carter, W.H.J. (1988) "Response Surface Methodology from 1966 to 1988." Technometrics. Vol. 12, 137-157.

7.- O'Mahony, M. (1982). "Some assumptions and difficulties with common statistics for sensory analysis." Food Technology. Vol.

TABLA I

PRODUCTO CON CARNE DE POLLO

ANDEVA

VARIABLE DE RESPUESTA: TEXTURA

FUENTE DE VARIACION	g.l.	SC	CM	F
Réplicas	1	22.205	22.205	
A=% de carne	2	38.8061	19.4031	5.38
Error a	2	7.2134	3.6067	
Subtotal	5	**68.2245**		
Fosfatos	2	204.5082	102.2541	9.28*
Nitritos	2	177.4461	88.7231	8.05 *
Error b	8	88.1773	11.0222	
Total	17	**538.3561**		

PRODUCTO CON CARNE DE CONEJO

ANDEVA

VARIABLE DE RESPUESTA: OLOR

FUENTE DE VARIACION	g.l.	SC	CM	F
Réplicas	1	49.0463	49.0463	
A=% de carne	2	49.7084	24.8542	9.8553*
Error a	2	5.0438	2.5219	
Subtotal	5	**103.7985**		
Fosfatos	2	237.1163	118.5581	4.4*
Nitritos	2	188.3379	94.1689	3.49*
Error b	8	215.86	26.9825	
Total	17	**745.1126**		

PRODUCTO CON CARNE DE CARNERO

ANDEVA

VARIABLE DE RESPUESTA: COLOR

FUENTE DE VARIACION	g.l.	SC	CM	F
Réplicas	1	45.0773	45.0773	
A=% de carne	2	46.795	23.3975	8.58*
Error a	2	5.4527	2.7264	
Subtotal	5	**97.325**		
Fosfatos	2	82.2645	46.1323	8.14*
Nitritos	2	73.7231	41.8616	7.39*
Error b	8	45.3417	5.6677	
Total	17	**298.6543**		

TABLA II

PRODUCTO CON CARNE DE POLLO

	OLOR			SABOR			CONSISTENCIA		
	0	1	2	0	1	2	0	1	2
Nitritos	56.54	58.92	60.63	50.60	54.08	56.75	54.47	56.88	54.25
Fosfatos	57.92	59.08	59.70	54.17	55.00	52.83			

	COLOR			TEXTURA			APARIENCIA		
	0	1	2	0	1	2	0	1	2
Nitritos	53.00	57.00	63.25	53.00	54.75	59.50	48.33	51.97	54.17
Fosfatos	53.00	63.50	56.75	53.00	61.50	52.75			

PRODUCTO CON CARNE DE CONEJO

	OLOR			SABOR			CONSISTENCIA		
	0	1	2	0	1	2	0	1	2
Nitritos	58.60	54.58	63.36	54.12	57.49	58.46	47.58	47.75	56.53
Fosfatos	58.60	56.05	61.84	55.00	59.70	56.37	48.11	51.88	52.27

	TEXTURA			APARIENCIA		
	0	1	2	0	1	2
Nitritos	48.44	48.96	56.64	51.88	51.65	55.64
Fosfatos	48.25	52.25	53.35	52.00	52.89	54.41

PRODUCTO CON CARNE DE CARNERO

	OLOR			SABOR		
	0	1	2	0	1	2
Nitritos	62.21	60.57	49.53	63.07	61.95	54.27
Fosfatos	63.30	48.50	60.51			

	COLOR			TEXTURA			APARIENCIA		
	0	1	2	0	1	2	0	1	2
Nitritos	65.82	65.30	56.01	58.03	60.58	52.46	63.72	65.56	46.06
Fosfatos	66.80	55.30	65.03	59.12	50.37	61.58	65.70	47.40	62.24

CARTAS DE CONTROL ROBUSTAS

Ernesto Barrios Zamudio
y
Jorge Domínguez Domínguez

Centro de Investigación en Matemáticas, A.C.
Apartado Postal 402
36000-Guanajuato, Gto.

Resumen

Se presenta el uso de medidas resistentes para la construcción de las tradicionales Cartas de Control de Shewhart. Se muestra también el uso de los diagramas de caja como una forma alternativa de graficación para evaluar procesos. Se comentan y comparan distintos procedimientos sugeridos por algunos autores utilizando medidas robustas para la construcción de las cartas de control.

1. Introducción.

Las cartas de control $\overline{X}$-R son una de las herramientas básicas empleadas en la industria con el objeto de analizar, mantener o llevar a control estadístico un proceso. Generalmente se elige una característica de interés del producto o proceso, y se estudia su tendencia central mediante la distribución de la variable $\overline{X}$, o bién, la mediana $\tilde{X}$ (medidas de posición), y la variabilidad del proceso se analiza usando la distribución del rango R (medida de dispersión), en algunas ocasiones se usa la desviación estándar S. El conocimiento de $\overline{X}$ y R se genera en la práctica seleccionando una muestra de cuatro a diez mediciones en diferentes periodos regulares -llamados *subgrupos homogéneos*- a lo largo de un intervalo de tiempo, es común considerar de 20 a 40 muestras.

Las *cartas de control* introducidas en 1924 por el Dr. Walter A. Shewhart, como se muestran en las figuras 1 y 2, consisten básicamente en una línea central y dos laterales equidistantes -llamadas *límites de control*- de modo que la banda definida por los límites tiene una amplitud de 6 veces la desviación estándar del estadístico correspondiente. Así, para una carta $\overline{X}$ la línea central es el promedio de las $\overline{X}$'s, denotada por $\overline{\overline{X}}$ y la distancia entre límites de control, conocidos como *límite superior de control* LSC y *límite inferior de control* LIC y representados por las líneas punteadas en la figura 1, es de $6\sigma_{\overline{X}}$. Los puntos en la carta son el valor de $\overline{X}$ calculado en cada muestra. La misma idea da lugar a la carta R (en la figura 2, para el límite inferior de control se considera el cero puesto que no tiene sentido un rango negativo).

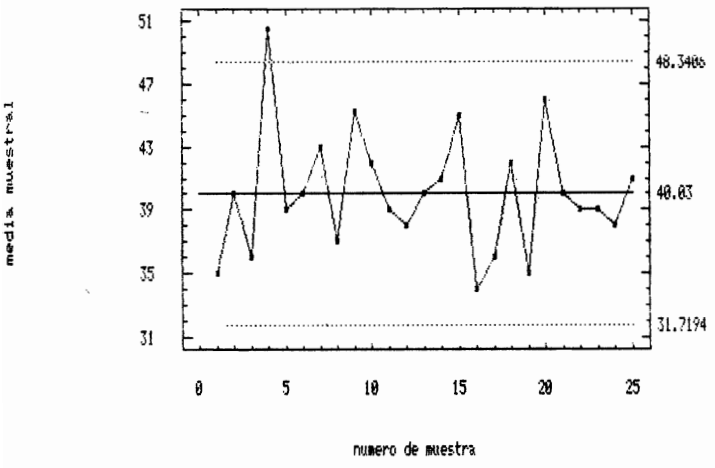
En el caso de una carta $\overline{X}$, suponiendo que el proceso se ha estabilizado (tendencia y variabilidad constantes en el tiempo) y las mediciones se distribuyen normalmente alrededor de la media, es de esperar que sólo aproximadamente un 0.3% de los puntos en la carta quedarán más allá de los límites de control. Así, un punto fuera de los límites se interpreta como si el nivel del proceso hubiese variado; es decir, ya no se encuentra bajo *control estadístico*, por lo que habría que indagar la razón del por qué y en su caso realizar las correcciones necesarias para regresar el proceso a control.

En la práctica se recomienda primero la estabilización de la variabilidad para después controlar la tendencia central. Así, para un proceso se construye una carta R, o bien una carta S, y hasta después de mostrar que está bajo control se comienza con la correspondiente carta de medias o de medianas. Finalmente, para fines de control, ambas cartas se analizan simultaneamente, siendo los puntos fuera de los límites una de las llamadas pruebas para *causas especiales*, es decir,comportamientos sospechosos de que el procesos se encuentra fuera de control estadístico (véase Nelson (1984)). Las líneas de control (línea central y límites) se recalculan después de 30 ó 40 muestras.

El gran valor de las Cartas de Control de Shewhart radica en la sencillez de su construcción e interpretación, por lo que resulta ser una herramienta muy eficaz para el usuario no versado en la Estadística.

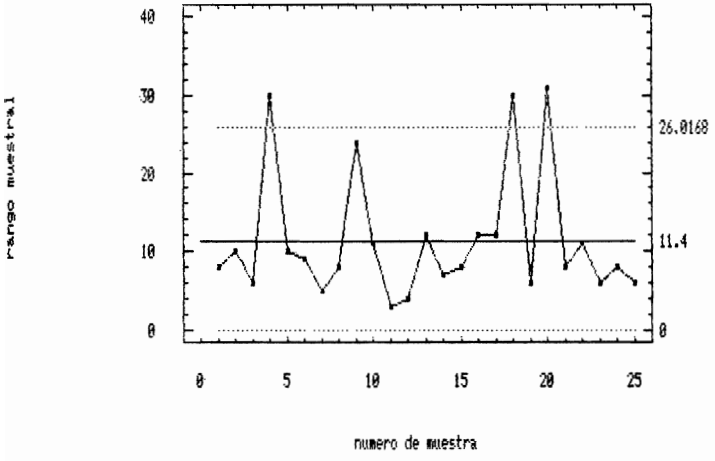
CARTA X

fig 1.



CARTA R

fig 2.



Por cuestiones de costo, en la práctica se seleccionan muestras de tamaño pequeño -rara vez superior a 10. Por ello, las observaciones grandes en la muestra dan lugar a que $\overline{X}$ se vea afectada sensiblemente, a la vez que influyen haciendo más ancha la banda determinada por los límites de control. Esto último dificulta la posibilidad de detectar cambios en el proceso. Analogámente, en el caso de R, observaciones con valores muy grandes o pequeños harán crecer significativamente el rango. En este sentido se dice que $\overline{X}$ y R son no resistentes.

Distintos autores han propuesto el uso de medidas robustas para construir cartas de control como alternativa a las tradicionales $\overline{X}$ y R. Informalmente, dichas medidas son menos sensibles ante valores grandes o pequeños (dan resultados más consistentes), o a ligeras desviaciones del supuesto de la normalidad de $\overline{X}$. La idea entonces, es considerar estadísticas más "resistentes" a observaciones cuyos valores son grandes. En este sentido decimos que las medidas resistentes son *robustas*. Sin embargo, el uso de estadísticas robustas tiene su costo. Se ha mostrado en varios estudios que las medidas no resistentes son más eficientes, en el sentido que hacen menos llamadas en falso, cuando se satisfacen los supuestos distribucionales, veáse por ejemplo Langenberg e Iglewicz(1987), Rocke(1989).

Para estudiar los casos de observaciones discrepantes, o bien, cuando se tienen distribuciones no simétricas, se han sugerido distintos procedimientos simples que emplean medidas resistentes. Con esta misma idea, se han propuesto también la construcción de cartas de control usando diagramas de caja, como puede verse en Iglewicz y Hoaglin (1987). Con ellas se puede evaluar el proceso de tal manera que se conozca a la vez la medida de tendencia central y la variabilidad en una sola gráfica, veáse también White y Schroeder (1987).

El objetivo de este trabajo es dar a conocer los procedimientos robustos propuestos para construcción de cartas de control, y señalar algunas de las conclusiones obtenidas.

2. Procedimientos robustos para cartas de control.

Los primeros estudios desarrollados para el empleo de medidas resistentes en la construcción de las cartas de control, como alternativa al uso de las tradicionales $\overline{X}$ y R, se deben a Ferrell en el año de 1953, donde recomienda

usar las medianas y el rango medio. Clifford en 1959, considera mejor calcular las medianas del rango y del rango medio en lugar del uso del promedio para determinar los límites de control. Ambas referencias aparecen citadas en el artículo de White y Schoeder (1987).

Los distintos procedimientos considerados para construir cartas de control "robustas" se apoyan en la estadística de orden.

Sean $X_1, X_2, \ldots, X_n$ las observaciones de una muestra de tamaño n, la estadística de orden, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, son las observaciones X_i ordenadas en magnitudes de orden creciente. Denotaremos por e un entero positivo y definimos:

$$X_{(e+.5)} = \frac{X_{(e)} + X_{(e+1)}}{2}$$

para muestras de tamaño n, definimos también el índice i como:

$$i = \frac{\left[\left(\frac{n+1}{2}\right) + 1\right]}{2} = \frac{\left[\frac{n+3}{2}\right]}{2},$$

donde [x] expresa el menor entero mayor o igual a x.

Las medidas de tendencia central (localización) consideradas son:

la media: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

la mediana: $\tilde{X} = X_{((n+1)/2)}$

La F-media:

$$\bar{X}_F = \begin{cases} \bar{X} & \text{para } n = 4 \\ \sum_{k=1}^{n+1-i} \frac{X_{(k)}}{n+2-2i} & \text{para } i \text{ entero} \\ \left\{X_{(1)} + X_{(n+1-1)} + \sum_{k=e+2}^{n-e-1} X_{(k)}\right\}/(n-2e), & i=e+.5 \end{cases}$$

la media "recortada":

$$\bar{X}(\alpha) = \frac{1}{m(1-2\alpha)} \left[(1-t)(X_{(r+1)} + X_{(m-r)}) + \sum_{k=r+2}^{m-r-1} X_{(k)}\right]$$

donde: $t=\alpha m - r$, $r=[\alpha m]$, $0<\alpha<1$. Es decir, de una muestra de tamaño n, se elimina el $100\alpha\%$ de cada cola de las observaciones ordenadas y se calcula la media aritmética del resto (definida por Hoaglin, Mosteller y Tukey (1983)).

Las medidas de dispersión consideradas son:

La Q-Dispersión:

$$R_Q = \begin{cases} X_{(n+1-i)} - X_{(i)} & \text{para } i \text{ entero} \\ X_{(n-e)} - X_{(e+1)} & \text{para } i = e + \frac{1}{2} \end{cases}$$

la expresión anterior es una aproximación a la definida por Tukey (1977)

La F-Dispersión:

$$R_F = X_{(n+1-i)} - X_{(i)}$$

que en caso de $i = e + \frac{1}{2}$, tomamos la convención considerada para la mediana.

El Rango Inter-Cuartil:

$$R_{IC} = X_{(k_1)} - X_{(k_2)}$$

donde $k_2 = \left[\frac{n}{4} \right] + 1$, y, $k_1 = n - k_2 + 1$.

Las medidas de localización y variabilidad citadas anteriormente han sido combinadas en la construcción de distintas cartas de control. Así, Langenberg e Iglewicz (1986) proponen construir cartas usando la media recortada como medida de posición y el rango recortado para la dispersión. Cabe observar que $E(\overline{R})=kE(\overline{R}(\alpha))$, por lo que en el cálculo de los límites se sustituye a $\overline{R}$ por un múltiplo del rango recortado $\overline{R}(\alpha)$. Este método lo comparan con el de las cartas $\overline{X}$-R, simulando distintas distribuciones en la generación de las muestras. Dicho método resultó superior que el tradicional cuando las condiciones de normalidad no se satisfacen y/o cuando existen observaciones discrepantes. Rocke (1989) propone construir dos nuevas cartas, una emplea $\overline{\overline{X}}$ para la posición y el rango inter-cuartil para la variabilidad ($\overline{X}$-R_{IC}), la otra está relacionada con la media recortada $\overline{X}(.25)$ y el rango inter-cuartil ($\overline{X}(\alpha)$-R_{IC}). Estos dos procedimientos los compara con el propuesto por Ferrell en 1953, el de Langenberg e Iglewicz (1986) y la carta $\overline{X}$-R. Cuando no hay causas de variación y las mediciones se distribuyen normalmente los procedimientos antes descritos se comportan de manera similar. Por otro lado, en presencia de observaciones discrepantes, la carta cuya línea central es la media de las medianas y los límites son calculados a partir de R_{IC} resultó ser la menos sensitiva pero con inferior eficiencia. Bajo las mismas condiciones se observó que las cartas $\overline{X}$-R_{IC} son robustas y en general eficientes. Así también, mediante un proceso de simulación Iglewicz y Hoaglin (1987) concluyen que la

F-media es una medida más eficiente que la mediana y la recomiendan como una medida resistente alternativa a la media. En el caso de las medidas de dispersión, la F-dispersión mostró mejores resultados que la Q-dispersión.

3. Líneas de Control.

El planteamiento general del problema es el siguiente: se desea construir cartas de control del tipo 3σ a partir de que disponemos de N subgrupos (homogéneos) con n observaciones cada uno, $X_{i1},\ldots,X_{in}$, (i=1,...,N), que se distribuyen independiente e idénticamente, una estadística de localización G_i y una estadística de dispersión S_i para cada subgrupo y un método T que resume las estadísticas de localización y dispersión a utilizar.

Rocke (1989) define un método general para la construcción de las cartas de control 3σ, a partir genéricamente de las estadísticas del subgrupo $\mathbf{G}=(G_1,\ldots,G_N)$ y $\mathbf{S}=(S_1,\ldots,S_N)$. Los límites de control están dados por:

$$T(G) \pm 3T\left(\sigma_G/\mu_{T(S)}\right) \qquad (1)$$

donde σ_G es la desviación estándar de la estadística del subgrupo G y $\mu_{T(S)}$ es el valor esperado del estadístico resumen de la dispersión.

Usando la expresión (1) podemos encontrar las líneas de control para las distintas cartas comentadas. Por ejemplo, si G y S representan la media y el rango del subgrupo respectivamente y T la media, entonces se tiene que

$$\mathbf{G} = \left(\bar{X}_1, \bar{X}_2,\ldots, \bar{X}_n\right) \qquad \text{y} \qquad \mathbf{S} = \left(R_1, R_2,\ldots,R_n\right)$$

y

$$T(G) = \bar{\bar{X}} \qquad \text{y} \qquad T(S) = \bar{R}$$

Si además suponemos que la distribución de las observaciones es normal de media μ y varianza σ^2,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad \text{y} \qquad \mu_{T(S)} = \mu_{\bar{R}} = E(\bar{R}) = d_2\sigma$$

donde d_2 es la constante correctora del sesgo, siguiendo la notación usual en Cartas de Control. Sustituyendo en (1) obtenemos los límites de control para la carta $\bar{X}$

$$\overline{\overline{X}} \pm 3\overline{R}\left(\frac{\sigma}{\sqrt{n}}/d_2\sigma\right) = \overline{\overline{X}} \pm 3\frac{\overline{R}}{\sqrt{n}\,d_2} = \overline{\overline{X}} \pm A_2\overline{R}$$

donde $A_2 = \dfrac{3}{\sqrt{n}\,d_2}$.

De manera similar en la carta R, G y S representan el rango del subgrupo y el método T es la media. Así,

$$\mathbf{G} = \left(R_1, R_2, \ldots, R_n\right), \qquad \mathbf{S} = \left(R_1, R_2, \ldots, R_n\right)$$

y de la expresión (1), los límites están dados por

$$\overline{R} \pm 3\overline{R}\frac{d_3\sigma}{d_2\sigma} = \overline{R} \pm 3\overline{R}\frac{d_3}{d_2} = \left(D_3\overline{R},\ D_4\overline{R}\right)$$

donde $D_3 = 1 - 3\dfrac{d_3\sigma}{d_2\sigma}$ y $D_4 = 1 + 3\dfrac{d_3}{d_2}$, siguiendo nuevamente la notación usual. Véase por ejemplo el libro de Burr (1976) donde las constantes A_2, D_3, D_4, d_2, d_3 se encuentran tabuladas.

De la misma forma, siguiendo la expresión (1) se pueden calcular los límites para las demás cartas. En la tabla 1 mostramos algunos casos.

C A R T A S	LIMITES DE CONTROL
Media recortada $\overline{X}(\alpha)$ Rango recortado $\overline{R}(\alpha)$	$\overline{\overline{X}}(\alpha) \pm A_2(k\ \overline{R}(\alpha))$ (1) $D_3(k\ \overline{R}(\alpha)),\ D_4(k\ \overline{R}(\alpha))$
Media de la mediana Rango	$\overline{m} \pm A_7\overline{R}$ (2) $D_3\overline{R},\ D_4\overline{R}$
Mediana Q-dispersión	$\overline{m} \pm M_2(\overline{Q})$ (3) $Q_3(\overline{Q}),\ Q_4(\overline{Q})$
F-media F-dispersión	$\overline{X}_F \pm A_{2F}\overline{R}_F$ (4) $D_{3F}\overline{R}_F,\ D_{4F}\overline{R}_F$

Tabla 1.

(1) $E(\overline{R}) = kE(\overline{R}(\alpha))$. Los valores de k apropiados para distintos tamaños de muestra se presentan tabulados en Langenberg e Iglewicz(1986).

(2) A_7, D_3, D_4, son constantes usuales para las cartas $\overline{X}$-R, véase por ejemplo Burr (1976)

(3) M_2, Q_3 y Q_4 son constantes calculadas por White y Schroeder (1987).

(4) A_{2F}, D_{3F} y D_{4F} son constantes que reportan calculados Iglewicz y Hoaglin (1987).

4. Ejemplo Numérico.

Con el objeto de ilustrar los distintos procedimientos comentados, en el apéndice se presentan los datos de mediciones practicadas a la profundidad de cilindros automotrices. Las dimensiones se ajustaron para efectos del ejemplo. Se consideraron N=25 grupos de n=4 mediciones por grupo.

En el caso de la carta de control utilizando la *media recortada* los cálculos necesarios se presentan a continuación:

1. Consideramos $\alpha = 0.10$
2. Calculamos $r = [(25)(.10)] = 2$
3. Los valores extremos de $\overline{X}$'s y R's eliminados de las 25 muestras son:
   Inferiores $\overline{X}_{(1)}$=34mm., $\overline{X}_{(2)}$=35mm., $R_{(1)}$=3mm., $R_{(2)}$=4mm.
   Superiores $\overline{X}_{(25)}$=50.5.m., $\overline{X}_{(24)}$=46mm., $R_{(25)}$=31mm., $R_{(24)}$=31mm.
4. $\overline{\overline{X}}(\alpha) = 39.79$ y $\overline{R}(\alpha) = 10.33$
5. Los valores de las constantes son k=1.0171, A_2=0.729, D_3=0 y D_4=2.282.
6. $\overline{\overline{X}} \pm A_2(k\overline{R}(\alpha)) = 39.79 \pm 7.66 = (32.14 , 47.46)$
7. $(D_3(k\overline{R}(\alpha)) , D_4(k\overline{R}(\alpha))) = (0.00 , 23.96)$

Para el resto de las cartas presentadas en la tabla 1, el cálculo de las líneas de control, incluyendo los correpondientes a la carta $\overline{X}$-R, se resumen en la tabla 2:

5. Comentarios.

Vemos en la tabla 2 que la distancia entre los límites superior e inferior para las cartas que corresponden a las medidas de posición es más angosta con respecto a la de la carta $\overline{X}$, excepto para la carta $\tilde{X}$. Análogamente en las cartas de dispersión los límites son más angostos que en la carta R. Esta si-

tuación se presentará en general cuando existan observaciones discrepantes, o bien que la distribución de las X_i se aleje de la distribución normal.

CARTAS	RESUMEN	LIMITES SUPERIOR	INFERIOR
Media $\overline{X}$	40.04±8.34	48.34	31.72
Rango R		0.00	26.02
Media recortada $\overline{X}(\alpha)$	39.80±7.66	47.46	32.14
Rango recortado $\overline{R}(\alpha)$		0.00	23.96
Media de la mediana	39.22±9.10	48.32	30.12
Rango		0.00	26.02
Mediana	39.80±7.24	47.03	32.56
Q-dispersión		0.00	23.15
F-media	40.01±8.01	48.02	32.00
F-dispersión		0.00	16.46

Tabla 2.

El resultado práctico de tener bandas de control más angostas, que es el caso general en las cartas robustas, es que permite detectar más fácilmente puequeños cambios de nivel en el proceso; esto es, son más sensibles que las tradicionales $\overline{X}$-R. Este es el caso de la novena muestra en nuestro ejemplo. En la carta R (fig 2) el noveno punto se encuentra dentro de los límites mientras que para cualquiera de las otras cartas de la tabla 2 quedaría fuera.

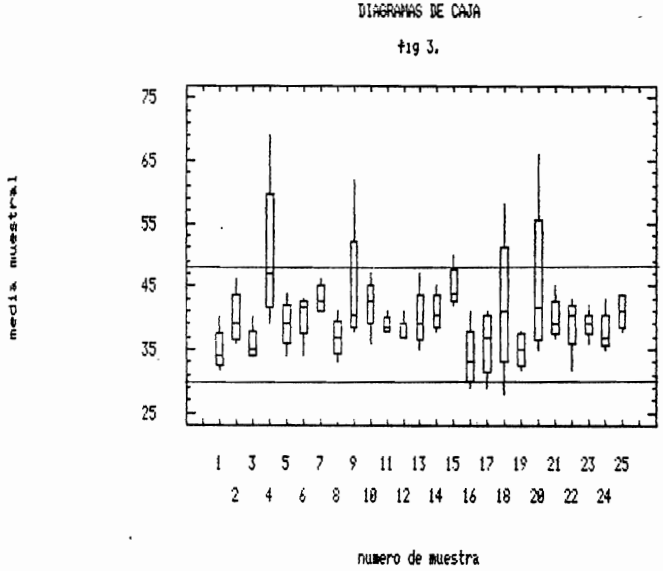
Otro aspecto interesante que podemos observar en la tabla 2, es que la línea central de las cartas robustas queda por abajo de $\overline{\overline{X}}$. Esta situación se presenta en general. El uso de medidas resistentes permite tener un equilibrio entre los puntos que están arriba y abajo de la línea central.

Por otro lado, Tukey (1977) introduce el uso de los diagramas de caja como una forma de describir la información. Esta idea la toman Iglewicz y Hoaglin (1987) para presentar en una sola carta la localización y variabilidad

del proceso. Los diagramas de caja resumen en cinco números la característica de los datos, las tapas de las cajas representan los cuartiles superior e inferior, la línea central indica la mediana y las líneas que parten de las tapas de las cajas se trazan hasta encontrar el punto máximo o mínimo de los datos. Las cajas graficadas en el tiempo nos dan idea del estado del proceso.

Como se indicó antes, la línea central de las cajas es la mediana -que es una medida resistente-, por lo que si incluimos en la gráfica límites de control robustos como los de la tabla 1, tendremos como resultado una carta de control robusta. La figura 3 corresponde a la carta de nuestro ejemplo, habiendo usado el rango para el cálculo de los límites.

Resulta de interés notar que mientras que en la figura 1 el cuarto punto sale de los límites, en la figura 3 ningún punto queda fuera de control. Esto se debe a la diferencia entre medias y medianas. Nótese también que los puntos fuera de control en la carta R (fig. 2), resultan evidentes en la figura 3.



APENDICE

muestra	profundidad (mm)			
1	35	40	32	33
2	46	37	36	41
3	34	40	34	36
4	69	39	44	50
5	38	34	44	40
6	42	41	43	34
7	44	41	41	46
8	33	41	38	36
9	38	62	39	42
10	47	43	36	42
11	38	41	39	38
12	37	37	41	37
13	40	38	47	35
14	38	39	45	42
15	43	45	42	50
16	31	35	29	41
17	41	40	29	34
18	38	44	28	58
19	33	32	37	38
20	66	45	35	38
21	38	40	45	37
22	40	43	32	41
23	42	39	39	36
24	43	36	35	38
25	39	38	43	44

Tabla A

Referencias.

1. BURR, I. W. (1976). "Statistical quality control methods". Marcel Dekker, Inc. New York.
2. IGLEWICZ, B., HOAGLIN, D. (1987). "Use of boxplots for process evaluation". Journal of Quality Technology 19, pag. 180-190.
3. HOAGLIN, D., MOSTELLER, F. y TUKEY, J.W. (1983). "Understanding robust and exploratory data analysis". John Wiley and Sons, Inc. New York.
4. LANGENBERG, P., IGLEWICZ, B. (1987). "Trimmed mean $\overline{X}$ y R charts". Journal of Quality Technology 18, 152-161.
5. NELSON, L. S. (1984). "The Shewhart Control Chart - Test for Special Causes". Journal of Quality Technology 16, 237-239.
6. TUKEY, J.W. (1977). "Exploratory data analysis". Addison-Wesley Reading M.A.
7. ROCKE, D. (1989). "Robust control charts". Technometrics 31, 173-184.
8. WHITE, E.M., SCHROEDER, R. (1987). "A simultaneous control chart". Journal of Quality Technology 19, 1-10.

SIMEC : SISTEMA DE MUESTREO PARA ECOLOGIA.

Graciela Bueno A.
Osvaldo Camacho C.
Consuelo Diaz T.

Centro de Estadística y Cálculo
Colegio de Postgraduados
Chapingo, Mex.

RESUMEN

En este trabajo se presenta el SIMEC que es un sistema interactivo para microcomputadoras IBM-PC y compatibles, especializado en el diseño y análisis de estudios por muestreo bajo los esquemas de transectos y cuadrantes.

INTRODUCCION

Muchas investigaciones ecológicas y agronómicas tienen como objetivo conocer el número de individuos en un área determinada. Esta información puede desearse para conocer índices de crecimiento de la población, avances de enfermedades o plagas, proporción de individuos que muestran alguna caracteristica de interes.

En la mayoría de los casos es impráctico o demasiado costoso realizar un censo en el área de estudio, por lo que la mejor opción es utilizar las técnicas de muestreo para estimar los parámetros de interés.

En este tipo de estudios se utilizan básicamente los esquemas del muestreo de cuadrantes y transectos.

El objetivo de este trabajo fue desarrollar un sistema interactivo de fácil uso que ayude al investigador en la selección de una muestra piloto o definitiva, asi como en la estimación del total de una población, en aquellos casos en los que se utiliza un esquema de muestreo por cuadrantes o por transectos, esto es, cuando la unidad experimental son cuadrados o franjas del terreno donde se encuentra la población a estudiar.

SIMEC es una versión modificada del sistema SIMETAP (1) programado en el lenguaje Pascal y compilado con la version 5.5 de Turbo Pascal para microcomputadoras IBM-PC y compatibles.

PRESENTACION DEL SISTEMA

SIMEC es un sistema interactivo que funciona con menúes donde el usuario debe elegir una de las opciones desplegadas en el monitor, permite el diseño y análisis de la muestra para los esquemas de cuadrantes y transectos.

En el menú principal se presenta las siguientes opciones:

1) Descripción del sistema
2) Esquema de cuadrantes
3) Esquema de transectos

DESCRIPCION DEL SISTEMA

Cuando se elije esta opción solo se presenta una descripción de la facilidades y funciones del sistema.

ESQUEMA DE CUADRANTES

En este esquema la unidad muestral puede estar constituida por cuadrados, rectángulos, hexagonos o círculos. Debemos señalar sin embargo, que los estudios anteriores, han demostrado que es mejor utilizar rectángulos y cuadros que no sean muy pequeños, y que además, en el caso de rectángulos, no sean muy angostos en alguna de sus dimensiones para eliminar hasta donde sea posible el efecto de orilla(En SIMEC se utilizan cuadros).

Este esquema de muestreo se puede utilizar cuando se cumplen satisfactoriamente las siguientes suposiciones:

a) La población está uniformemente distribuida.
b) La presencia de un individuo no influye en la posición ocupada por otro.

Bajo estas suposiciones, n, el número de individuos encontrados en el área muestreada tiene una distribución binomial.

Con la opción de cuadrantes en el menú principal de SIMEC, se presenta a elegir las opciones siguientes:

(1) Conocer conceptos básicos
(2) Diseño
(3) Análisis

Con la opción (1) "conocer conceptos básicos", se muestra una descripción del esquema de muestreo por cuadrantes.

Cuando se elige la opción (2) "diseño", es necesario proporcional al sistema una serie de puntos extremos que definen el contorno del terreno y el área del mismo.

Estos datos son guardados automáticamente en un archivo bajo un nombre que le es solicitado al usuario para su posterior utilizarse.

A continuación se presenta un menú con las siguientes opciones:

a) Muestra piloto
b) Muestra total
c) Analizar muestra piloto para obtener una muestra total.

Con la opción de "muestra piloto" (a) el usuario puede probar varias alternativas de cuadros de distinta longitud de lado (l) que son dibujados sobre la gráfica del terreno desplegado en pantalla, para que el usuario elija la opción más adecuada a la especie que desea estudiar. Una vez escogida una cuadrícula, se encuentra el tamaño de muestra con la ecuacion:

$$m=PA/a$$

Donde "m" Tamaño de la muestra piloto.

"A" es el área del terreno.

"a" es el área del cuadro.

"P" es la proporción del área que se desea observar a través de la muestra piloto,

Se obtiene aleatoriamente la muestra piloto y se despliega ésta proporcionando las coordenadas superior izquierda e inferior derecha de cada cuadro a muestrear, y marcando los cuadros seleccionados en la gráfica del terreno. Esta información puede adicionalmente imprimirse o enviarse a disco.

Con la opción de "Muestra total" (b) el usuario debe proporcionar una estimación previa del tamaño de la población $\hat{N}$ y un coeficiente de variación, con ello se estima "$\hat{P}$" la proporción del área que se debe muestrear y "m" el número de cuadros que constituyen la muestra.

Con la opción "Analizar muestra piloto para obtener muestra total" (c), se tecleará el coeficiente de variación, el número de cuadros en la muestra y el número de individuos observados en cada cuadro. Se analizará a continuación esta información para obtener un tamaño de muestra total.

Si el tamaño de muestra obtenido a través del análisis es mayor que el tamaño de muestra piloto, se seleccionará aleatoriamente los cuadros que deberán agregarse para la muestra definitiva.

Al igual que con las opciones anteriores los cuadros seleccionados son exhibidos en la gráfica del terreno y sus coordenadas desplegadas.

Con la opción "Análisis" (3), del menú de Muestreo por Cuadrantes, el sistema analiza la información recabada en el muestreo definitivo y proporciona el estimador del total de la población, junto con su desviación estándard.

La información que se debe proporcionar al sistema, en este caso, puede simplemente ser un complemento de la existente en un archivo previamente creado (Muestra piloto) o bien ser proporcionada totalmente en este misma etapa.

ESQUEMA DE TRANSECTOS

En este esquema de muestreo, las unidades muestreadas son franjas del terreno, donde se encuentra la población a estudiar. Un obsevador recorrerá las franjas o transectos y al hacerlo contará los individuos observados y tomará una o dos de las siguientes medidas: 1) distancia perpendicular del individuo al transecto, 2) distancia radial (r_i) entre el observador y el individuo y el ángulo (θ_i) formado por el observador, el individuo y el transecto.

Para una mejor utilización de este esquema de muestreo se requiere que se cumplan satisfactoriamente las siguientes suposiciones:

1) Los individuos se encuentran uniforme e independientemente distribuidos.
2) Observar un indiviuo es independiente de observar otro.
3) Ningún individuo se cuenta más de una vez.
4) Cada individuo es observado en la posición exacta en que se encontraba cuando el observador comenzó el recorrido.
5) El comportamiento de la población como un todo no cambia sustancialmente durante el recorrido del transecto.
6) Los individuos son homogéneos respecto a su comportamiento sin importar sexo, edad, etc.
7) La probabilidad de que un individuo sea observado, dado que está a una distancia perpendicular x del transecto, es una función $g(x)$ tal que $g(0) = 1$.

Bajo estas suposiciones el estimador de la población $\hat{N}$ se obtiene como:

$$\hat{N} = \frac{nA}{2Lw}$$

donde "w" es la mitad del verdadero ancho de la franja de terreno cubierta por el observador cuando camina a lo largo del transecto.

Dependiendo de la medida registrada (distancia perpendicular o radial) y de la forma de la función g(x) se tienen diferentes modelos para la estimación de N.

Con la opción "Esquema de transectos" del menú principal de SIMEC, el sistema le pregunta qué medida utilizará: Distancia perpendicular o distancia radial, o si se desea conocer conceptos básicos.

Con la opción de "Conocer conceptos básicos" se desplegará en pantalla una descripción del esquema de muestreo por transectos.

Com la opción de "distancia perpendicular", se proporcionará al sistema el número de observaciones y los valores de las "x_i" (distancia perpendicular de individuo observado al transecto) y presentará un menú para elegir el tipo de estimador que se desea utilizar para estimar N y que puede ser:

1) Gates
2) Eberhart
3) Amman y Baldwin

En el estimador de Gates en la suposición 7 del esquema de transectos, se tiene que la función g(x) sigue una distribución exponencial con parámetro λ, que se estima en función de los datos x_i observados y cuyo valor participa en la estimación de w, el verdadero ancho del transecto que cubre el observador.

Con la opción de Gates, se proporciona el estimador del total de la población $\hat{N}$ y su varianza.

Si se utiliza el estimador de Gates es posible realizar un muestreo piloto con un transecto de longitud L_o; con los valores observados x_i y un coeficiente de variación, obtener la longitud definitiva L, para posteriormente obtener $\hat{N}$.

Con el estimador de Eberhart, la función g(x) de la suposición 7 es de la forma

$$g(x)=\begin{cases}1-\left(\frac{x}{w}\right)^b & 0 \leq x \leq w \\ 0 & w < x\end{cases}$$

En este caso, en el que existen dos parámetros, se tiene un estimador si "b" es conocida y otro si "b" es desconocida.

En el estimador de Amman y Baldwin la función g(x) de la suposición (7) es

$$g(x)=\begin{cases}10 & \leq x \leq w \\ 0 & x > w\end{cases}$$

Con la opción de "distancia radial" el sistema presentará la alternativa de utilizar los estimadores de King, Webb o Gates.

Con el estimador de Webb sera necesario teclear los los angulos Θ_i.

Cada uno de estos estimadores plantea una alternativa diferente para estimar w, el verdadero ancho del transecto, lo que produce diferentes estimadores para N.

En el estimador de King $\hat{w} = \bar{r}$, mientras que en el estimador de Webb, $\hat{w} = \bar{r}$ sen $(\bar{\Theta})$ y en el estimador de Gates $\hat{w} = G$, la media geométrica de las r_i.

REQUERIMIENTOS DEL SISTEMA

Una microcomputadora IBM-PC o compatible.
Una unidad de video.
Dos unidades de disco.
Una impresora, sólo si se desean resultados impresos.

REFERENCIAS

Torres D.C.G., G. Bueno de A., R. Gómez A., L. Landois P. (1989). Sistema para el muestreo y estimación de totales en poblaciones de animales y plantas. Agrociencia 75.

FAC2K-P UN SISTEMA INTERACTIVO PARA EL DISEÑO Y ANALISIS DE EXPERIMENTOS FACTORIALES A DOS NIVELES 2^{K-P}

Osvaldo Camacho Castillo
Guillermo P. Zarate de Lara
Colegio de Postgraduados, Chapingo Mex.
Centro de Estadística y Cálculo,

RESUMEN

Este trabajo consistió en la elaboración de un sistema interactivo para microcomputadoras IBM-PC y compatibles especializado en el diseño y análisis de experimentos factoriales a dos niveles.

INTRODUCCION

Los experimentos factoriales a dos niveles son de gran utilidad en las primeras etapas de una investigación secuencial, cuando se tiene una gran cantidad de factores y el investigador está interesado en saber cuales de estos factores tienen efecto sobre la variable respuesta, aún cuando no está interesado por el momento en conocer como es este efecto. Con el propósito de reducir los costos y hacer más factible la realización de un experimento factorial con muchos factores, se han generado los experimentos factoriales fraccionados, en los que sólo se prueba una parte de las combinaciones de los niveles de los factores bajo estudio.

El utilizar una fracción del experimento implica que tendremos algunos efectos confundidos, por lo que para hacer inferencias sobre algún efecto se tendrá que suponer la inexistencia de otros (generalmente de interacciones entre muchos factores).

OBJETIVO

El objetivo de este trabajo es la elaboración de un sistema interactivo para microcomputadoras especializado en el diseño y análisis de experimentos factoriales a dos niveles.

La salida del programa está compuesta por uno o mas de los siguientes resultados:

1.-Un conjunto de generadores con los que se obtiene la maxima resolución posible.

2.-La resolución alcanzada.

3.-La relación de identidad y el patrón de confusión.

4.-La matriz diseño generada.

5.-Los promedios de los efectos principales y las interacciones de segundo orden.

6.-El análisis de varianza.

7.-Las estimaciones de cada efecto e interacción, su cuadrado medio, F calculada y el nivel de significancia.

ESTRUCTURA DEL SISTEMA

La capacidad del programa es de hasta 11 factores y 128 tratamientos. Está compuesto de 27 procedimientos principales y algunos mas de apoyo.

En la figura 1 se muestra el diagrama de bloques del sistema.

SUBPROGRAMAS PRINCIPALES

A continuación se listan los procedimientos describiendo brevemente las acciones principales que realizan:

-Procedimiento dimensiones.

Lee el valor de k (número de factores) y de p (número de generadores).

-Procedimiento aleatorio.

Este procedimiento utiliza un parámetro de entrada entero y proporciona una lista de enteros entre uno y el valor dado, en orden aleatorio; se utiliza en el sistema para proporcionar la aleatorizacion del diseño; la lista se almacena en el arreglo de enteros "número".

-Procedimiento genera_matriz.

Requiere los valores de k y p y construye la matriz del diseño completo 2^{k-p} con valores -1 (para el nivel 'bajo') y 1, (para el nivel "alto") almacenándola en el arreglo bidimensional "mat",cuando p > 0 llama al procedimiento m_k_p para construir la parte faltante de la matriz diseño.

-Procedimiento m_k_p.

Procedimiento auxiliar llamado desde genera_matriz paracompletar la matriz diseño contruyendo las columnas de los factores generados. Requiere los valores de k, de p y los p generadores.

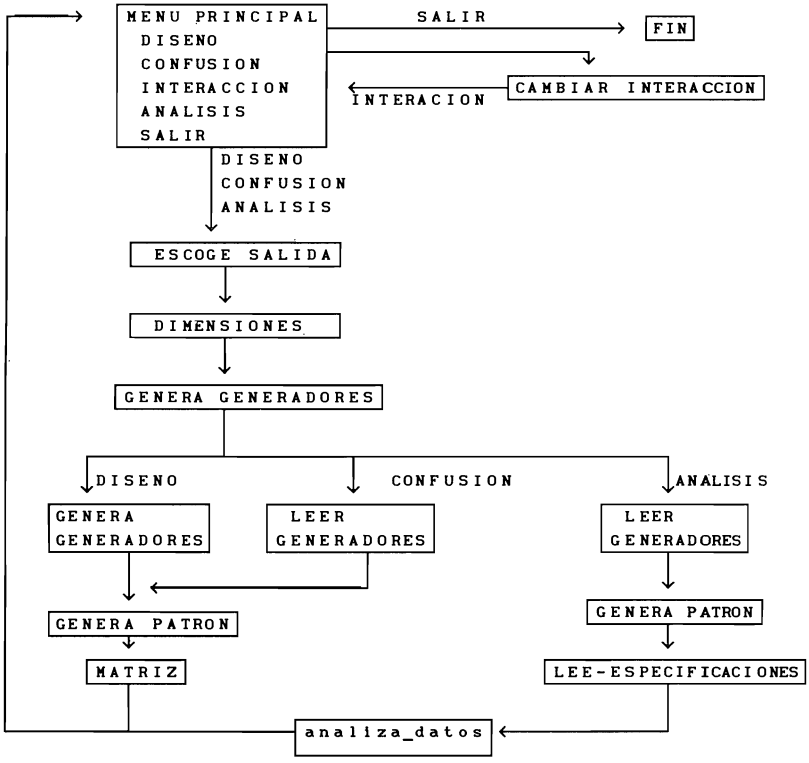


Fig. 1 DIAGRAMA DE BLOQUES DEL SISTEMA FAC2^K-P

-Procedimiento genera_factores.

Procedimiento que asigna al arreglo "factores" los valores 'A', 'B',...,'L', para denotar a los factores; en caso de que

el usuario decida no darles nombre; en caso contrario lee el nombre de dichos factores que deben constar de un solo caracter.

-Procedimiento lee_generadores.

Pide al usuario los p generadores que requiere el diseño, y los coloca en el arreglo de caracteres "generadores". Revisa que estos generadores sean válidos.

-Procedimiento genera_generadores.

Con este procedimiento, el sistema, produce un conjunto de p generadores con los que el diseño alcansa su resolución maxima.

Procedimiento genera_idr.

Con este procedimiento se obtiene la relacion de identidad, con todas la combinaciones posibles de los p generadores.

-Procedimiento genera_patrón.

Este procedimiento produce el patrón de alias del diseño.

-Procedimiento analiza_datos.

Este es un procedimiento de gran tamaño que requiere de varios auxiliares, se encarga de enlazarlos para realizar el análisis de los datos de un factorial 2^{k-p}.

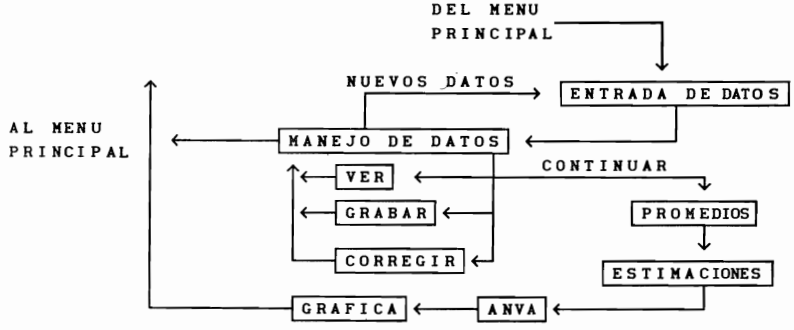


FIG. 2.- ESTRUCTURA DEL PROCEDIMENTO ANALIZA_DATOS

-Procedimiento gráfica.

Contruye y muestra la gráfica de probabilidad normal.

-Procedimiento acumular.

Este procedimiento calcula los promedios para los efectos principales e interacciones de dos factores.

-Procedimiento promedios.

Despliega los promedios de cada efecto principal e interacciones de primer orden .

-Procedimiento lee_esp.

Lee las especificaciones del diseño que se va a analizar tales como:

Nombre de la variable respuesta

Número de repeticiones

Número de puntos centrales

Nombre del archivo de datos

-Procedimiento estimación.

Utiliza los totales para cada combinación de los niveles de los factores y encuentra las estimaciones para cada uno de los efectos usando el método de Yates.

-Procedimiento anva.

Realiza el análisis de varianza para los datos analizados en caso de tener una estimación de la varianza, si no se tiene dicha estimación sólo presenta las estimaciones y los cuadrados medios.

-Procedimiento estimación_varianza.

Obtiene una estimación de la varianza con las siguientes opciones:

Por medio de las repeticiones cuando se tienen.

Si no se tienen repeticiones del diseño pero si del punto central, la estimación de la varianza se hará por medio de éstas.

Si no se tienen repeticiones ni del punto central entonces se le pedirá al usuario una estimación histórica para la varianza de su variable respuesta.

-Procedimiento matriz.

Procedimiento que da salida a la matriz de tratamientos básica en orden estándar, ésto es, la k-ésima columna está formada por 2^{k-1} signos negativos, seguidos de 2^{k-1} signos positivos.

MANUAL DEL USUARIO

FAC2^K-P .- Es un sistema interactivo que no requiere de un manual para su uso , sin embargo, para describir las facilidades con que cuenta se mostrarán algunos ejemplos.

Para ejecutar el programa se requiere de un disco que contenga el sistema FAC2^K-P, con la unidad correspondiente a la terjeta de fraficación con que se cuente (CGA.BGI, EGAVGA.BGI etc) del TURBO-PASCAL (versión 4.0 o posterior) y la unidad graphics del MS-DOS (versión 3.13 o posterior).

Para iniciar solo hay que teclear

A>FACTOR

Con el disco del sistema en la unidad A.

Al ejecutar el programa lo primero que aparece en pantalla es el menú principal

```
DISEÑO.............1
CONFUSION..........2
INTERACCION........3
ANALISIS...........4
SALIR..............5

ELIJA SU OPCION ===>
```

Aquí se deberá presionar el número de la opción seleccionada, presionando enseguida la tecla "ENTER"

OPCION DISEÑO.

La opción DISEÑO genera los experimentos factoriales y fracciones de factoriales a 2 niveles, usando generadores proporcionados por el programa. La salida consiste de la matriz diseño, la relación de identidad y el patrón de confusión.

Un ejemplo de un factorial fraccionado un 2^{5-1}

Se teclea :

A>FACTOR

Se elige la opción diseño del menú principal presionando las teclas 1 y "ENTER", enseguida aparece el siguiente mensaje:

TECLEE EL NUMERO DE FACTORES BAJO ESTUDIO ===>

Después de presionar 5 y "ENTER" aparece :

SU EXPERIMENTO ES UN 2^(5-P) TECLEE EL VALOR DE P ===>

A lo que se responde presionando 1 y "ENTER"; aparece entonces la opción para decidir darle o no nombre a los factores, en este último caso los nombres que se asignan son "A", "B",...,"L'.

DESEA DAR NOMBRE A SUS FACTORES S/N ===>
POR OMISION SE TOMA " A " " B "..." L "

Presionamos N y "ENTER" y a continuación aparece el menú para seleccionar el dispositivo de salida, que es el siguiente:

IMPRESORA	< 1 >
ARCHIVO	< 2 >
PANTALLA	< 3 >

ELIJA DONDE QUIERE SU SALIDA ===>

Si se desea dar salida a pantalla se presiona 3 y "ENTER" y la sailda será la siguiente:

EL GENERADOR PARA ESTE DISENO ES:

ABCDE

ESTE ES UN DISENO DE 5 VARIABLES Y UN GENERADOR

LA RELACION DE IDENTIDAD

I = ABCDE

RESOLUCION = 5

PATRON DE CONFUSION

A
B
AB + CDE
C
AC + BDE
BC + ADE
DE + ABC
D
AD + BCE
BD + ACE
CE + ABD
CD + ABE
BE + ACD
AE + BCD
E

LA MATRIZ DE TRATAMIENTOS BASICA EN ORDEN ESTANDAR

TRAT	A	B	C	D	E	ALEATORIZACION
1	-	-	-	-	+	4
2	+	-	-	-	-	6
3	-	+	-	-	-	16
4	+	+	-	-	+	15
5	-	-	+	-	-	10
6	+	-	+	-	+	7
7	-	+	+	-	+	5
8	+	+	+	-	-	13
9	-	-	-	+	-	2
10	+	-	-	+	+	9
11	-	+	-	+	+	14
12	+	+	-	+	-	12
13	-	-	+	+	+	11
14	+	-	+	+	-	1
15	-	+	+	+	-	3
16	+	+	+	+	+	8

E = ABCD

"CUIDADO" SOLO SE HAN CONSIDERADO INTERACCIONES DE ORDEN 3
Y MENORES. CON LA OPCION "INTERACCION" SE PUEDE CAMBIAR ESTO

LIMITACIONES DE LA OPCION DISEÑO

Los diseños disponibles para la opción diseño se señalan en la siguiente tabla; nótese que se requiere que $2^{k-p}>=k$.

K	P 0	1	2	3	4	5	6	7
3	SI	III						
4	SI	IV						
5	SI	V	III					
6	SI	VI	IV	III				
7	SI	VII	IV	IV	III			
8	NO	VIII	V	IV	IV			
9	NO	NO	VI	IV	IV	III		
10	NO	NO	NO	V	IV	IV	III	
11	NO	NO	NO	NO	V	IV	IV	III

SI Indica que el diseño factorial es posible y puede ser generado por la opción DISEÑO.

NO Indica que el diseño factorial es posible pero no puede ser generado por la opción DISEÑO ya que el número de puntos en el cubo pasa de 128.

NUMERO ROMANO Indica la resolución lograda con la opción DISEÑO, solo se genera la fracción principal.

ESPACIOS EN BLANCO Indica que el diseño no es posible.

OPCION CONFUSIÓN.

Con la opción CONFUSIÓN se pueden generar otros diseños diferentes a los que proporciona la opción DISEÑO tecleando el usuario un conjunto específico de generadores; sólo se pueden teclear generadores positivos.

Como ejemplo se presenta el diseño de un experimento 2^{7-4}

Se elige la opción confusión del menú pricipal.

```
TECLEE EL NUMERO DE FACTORES BAJO ESTUDIO ===> 7
SU EXPERIMENTO ES UN 2^(3-P) TECLEE EL VALOR DE P ===> 4

TECLEE SU GENERADORES CONFORME SE LE PIDAN

GENERADOR 1 ===>ABD
GENERADOR 2 ===>ACE
GENERADOR 3 ===>BCF
GENERADOR 4 ===>ABCG
```

Una vez que se ha tecleado lo anterior la salida es la siguiente:

```
ESTE ES UN DISENO DE 7 VARIABLES Y 4 GENERADORES


LA RELACION  DE IDENTIDAD

 I = ABD = ACE = BDCE = BCF = ADCF = AEBF = DEF = ABCG = DCG =
     EBG = DEAG = FAG = DFBG = EFCG = DEFABCG

RESOLUCION =  3

PATRON DE CONFUSION

A + BD + CE +  DCF  + EBF + BCG + DEG + FG
B + AD + DCE + CF + AEF + ACG + EG +  DFG
D + AB + BCE + ACF + EF + CG + AEG + BFG
C + AE + BDE + BF + ADF + ABG + DG + EFG
E + CBD + AC + ABF + DF + BG + ADG  +  CFG
F  + CAD + BAE + DE + BC + AG + BDG + CEG
G + CD + BE + ADE + AF  + BDF + CEF + ABC


LA MATRIZ  DE  TRATAMIENTOS  BASICA  EN  ORDEN ESTANDAR

TRAT A B C D E F G     ALEATORIZACION
 1  - - - + + + -      3
 2  + - - - - + +      6
 3  - + - - + - +      4
 4  + + - + - - -      2
 5  - - + + - - +      5
 6  + - + - + - -      7
 7  - + + - - + -      1
 8  + + + + + + +      8

   D = AB
   E = AC
   F = BC
   G = ABC
LOS SIGUIENTES EFECTOS ESTAN CONFUNDIDOS CON LA
MEDIA GENERAL Y NO PUEDEN ESTIMARSE DIRECTAMENTE

ABD   ACE   BCF   DEF   DCG   EBG   FAG

"CUIDADO" SOLO SE HAN CONSIDERADO INTERACCIONES DE ORDEN 3
Y MENORES  CON LA OPCION  "INTERACCION" SE PUEDE CAMBIAR ESTO
```

OPCION ANALISIS

Con la opción ANALISIS se analizan experimentos factoriales a dos niveles así como fracciones (el programa no puede trabajar con porciones de la matriz diseño sino que requiere de los datos completos), puede analizar diseños con puntos centrales.

Se presentan estimaciones de todos los efectos así como las sumas de cuadrados. Si el diseño tiene repeticiones, puntos centrales o una estimación histórica de la varianza, da también como salida la tabla del ANVA con los valores de F-calculada y su significancia; también da el error estándar de la estimaciones.

En casos en que no se tiene una estimación de la varianza (histórica, repeticiones del diseño o de los puntos centrales) se puede recurrir a la gráfica de probablidad normal que se presenta con la opción ANALISIS.

DATOS

Los datos pueden ser introducidos a partir del teclado o de un archivo creado por el mismo programa.

Para crear el archivo se usa la ANALISIS, proporcionando la información conforme el programa lo señala.

Después de teclear los datos aparece el menú de manejo endonde se elige la opción que se desea.

UN EXPERIMENTO FACTORIAL FRACCIONADO

En este ejemplo se considera un 2^{5-1}.

Se elige la opción análisis del menú principal.

TECLEE EL NUMERO DE FACTORES ===> 5

SU EXPERIMENTO ES UN 2^(3-P) TECLEE "P" ===> 1

TECLEE SU GENERADOR

GENERADOR 1 ===>ABDE

CUAL ES EL NOMBRE DE SU VARIABLE RESPUESTA ===> REND.

CUANTOS PUNTOS CENTRALES TIENE PARA RENDIMIENTO ===> 4

```
   CUANTAS REPETICIONES ===> 2

TECLEE EL NOMBRE DE SU ARCHIVO DE DATOS O "TECLADO" ===>
TECLADO

PROPORCIONE LOS DATOS PARA REND.

TRAT A B C D E
 1  - - - - +   REP 1   212
                REP 2   186
 2  + - - - -   REP 1   206
                REP 2   180
 .               .       .
 .               .       .
 .               .       .
16  + + + + +   REP 1   338
                REP 2   300
```

Una vez que se han tecleado todos los valores aparece el siguiente menú:

```
MENU DE MANEJO DE DATOS

   CORREGIR DATO (S).................< 1 >
   CARGAR NUEVOS DATOS...............< 2 >
   GRABAR EN UN ARCHIVO..............< 3 >
   VISUALIZAR SUS DATOS..............< 4 >
   REGRESAR AL MENU PRINCIPAL.........< 5 >
   CONTINUAR CON EL ANALISIS..........< 6 >

         ELIJA SU OPCION ===> 6
```

Una vez que se teclea esto, la salida es la siguiente:

```
RESULTADOS DE LA VARIABLE  RENDIMIENTO
PROMEDIO DE LOS EFECTOS PRINCIPALES

EFECTO          -1       PUNTOS CENTRALES      +1
     A       172.7500         266.7500      243.1875
     B       193.6875         266.7500      222.2500
     C       202.1875         266.7500      213.7500
     D       206.0000         266.7500      209.9375
     E       200.0625         266.7500      215.8750

 PROMEDIO DE LAS INTERACCIONES

                    INTERACCION  AB

  + A       225.3750             261.0000
                    266.7500
  - A       162.0000             183.5000
             - B                  + B
```

INTERACCION AC

+ A	240.0000		246.3750
		266.7500	
- A	164.3750		181.1250
	- C		+ C

Aquí aparecen los promedios para cada interacción.

INTERACCION DE

+ D	198.5000		221.3750
		266.7500	
- D	201.6250		210.3750
	- E		+ E

ANALISIS DE VARIANZA

EFECTO	G L	ESTIMACION DEL EFECTO	CUADRADO MEDIO	Fo	P>F
A	1	70.4375	39691.5312	170.235	0.000
B	1	28.5625	6526.5312	27.992	0.000
AB	1	7.0625	399.0312	1.711	0.209
C	1	11.5625	1069.5312	4.587	0.048
AC	1	-5.1875	215.2812	0.923	0.351
BC	1	16.6875	2227.7812	9.554	0.007
ABC	1	7.9375	504.0312	2.161	0.161
D	1	3.9375	124.0312	0.532	0.476
AD	1	5.9375	282.0313	1.209	0.288
BD	1	8.3125	552.7812	2.370	0.143
E	1	15.8125	2000.2812	8.579	0.010
CD	1	-12.1875	1188.2812	5.096	0.038
ACD	1	-10.9375	957.0313	4.104	0.060
BCD	1	-7.3125	427.7812	1.834	0.194
CE	1	81.4375	53056.5312	227.557	0.000

"CUIDADO" SOLO SE HA COLOCADO EL EFECTO DE ORDEN MENOR
CONSULTE EL PATRON DE CONFUSION PARA VER SUS ALIAS

PROMEDIO DE PUNTOS CENTRALES= 266.7500
PROMEDIO GENERAL = 214.5000
PROMEDIO DEL DISENO = 207.9687

ERROR ESTIMADO	G L	
233.1562	16	USADO EN EL CALCULO DE " F "

OPCION INTERACCION

Se elige la opción INTERACCION del menú principal, y aparece en la pantalla el siguiente texto:

```
SE ESTAN CONSIDERANDO LAS INTERACCIONES DE ORDEN 3 Y MENORES
SI DESEA CAMBIAR ESTO TECLEE EL ORDEN DE LAS INTERACCIONES QUE
LE INTERESAN ===> 5
A PARTIR DE ESTE MOMENTO EL PROGRAMA LE MOSTRARA
INTERACCIONES DE ORDEN 5 O MENORES.
```

REFERENCIAS

Box, G.E.P. y W.G. Hunter & J.S. Hunter (1978), Statistics For Experimenters, John Wiley and Sons, New York.

Camacho C.,O. (1988) Fac2k-p: Un sistema interactivo para el diseño y análisis de experimentos factoriales a dos niveles. Tesis de Maestria, Centro de Estadística y Cálculo, Colegio de Postgraduados, Chapingo México.

DEDUCCION DE UNA FUNCION DE RENDIMIENTO PARA SIEMBRAS MATEADAS AGRICOLAS

Francisco Camacho Morfin
Lab. de Semillas Forestales
CIFAP-DF. INIFAP
Av. Progreso No. 5, Coyoacán, D. F.
C. P. 04000

RESUMEN

Se plantean los supuestos en los que se fundamenta la obtención de una ecuación que relacione la cantidad de propágulos sembrados con el número de plantas que se producirán; se encontró que la función de rendimiento es en este caso la distribución binomial de probabilidades. Con base en ella se demostró que cuando se admite una distribución aleatoria de las plantas en el sembradío, las necesidades de propágulos requeridos para obtener una población dada, se determinan relacionando dicha población con la capacidad de los propágulos para producir plantas; en cambio, si la distribución debe ser regular es necesario el cálculo de probabilidades y costos de las labores. Se discute la optimización económica de las siembras.

INTRODUCCION

La población de plantas por hectárea y las distancias entre éstas influyen notoriamente en los rendimientos de los cultivos; para alcanzar los óptimos en estos aspectos es necesario considerar el material empleado para obtener las plantas, la estrategia de siembra que se realice y los costos de las labores.

En la producción agrícola es frecuente que se ignoren estos aspectos, tanto por los campesinos como por los profesionistas encargados de asesorarlos.

En este punto es vital la optimización económica del uso de insumos; para ello se requiere disponer de una función de rendimiento, que en este caso es evidente que se trata de una distribución de probabilidades.

La aplicabilidad de la función deducida depende del cumplimiento de los supuestos en que se fundamenta, por ello su planteamiento permite evaluar la validez de los resultados que se obtendrán. Otro aspecto que no debe olvidarse son las repercusiones estadísticas que tienen las labores que se realizan en el campo para la distribución de los propágulos y el ajuste de la población, esto es la estrategia de siembra usada.

El problema que se aborda en la presente contribución, es el primero con el que se enfrenta un productor; no obstante, ha merecido poco análisis tanto en los cursos de propagación de plantas como en los de estadística, en los que constituye un buen ejemplo del empleo del uso práctico de la distribución binomial de probabilidades, el teorema central del límite y la adaptación de las fórmulas para hacer accesible su empleo.

ANTECEDENTES

Los propágulos de las plantas son órganos vegetales o fragmentos de éstos que se emplean para obtener nuevas plantas; algunas de las opciones más usadas son: bulbos, estacas, semillas y tubérculos.

Se llama prendimiento al hecho de que un propágulo produzca una planta después de que ha sido sembrado, o sea colocado en condiciones que permitan el crecimiento vegetal.

Para comercializar los propágulos de una especie, generalmente se les maneja en grupos que provienen de la misma localidad y ciclo de cosecha; a estos grupos se les denomina lotes comerciales de propágulos. Estos lotes se diferencian de los lotes para siembra, en que estos últimos son una fracción de los primeros que se usan para establecer un cultivo en un área determinada.

Como se acostumbra manejar los lotes por peso, un dato importante es el número de plantas que se pueden obtener de un Kg del lote, cantidad que se determina así, en el caso de las semillas (Boyd, 1969):

$$V = S\ I\ p \qquad (1)$$

Donde:

S = propágulos por Kg de lote.
I = proporción de pureza del lote, es decir, la fracción constituida por propágulos, ya que suelen estar acompañados por basuras.
p = proporción de prendimiento de los propágulos respecto al total de ellos.

La determinación del tamaño del lote de siembra requerido para obtener una población dada, es el primer problema que se tiene para establecer un cultivo; para solucionarlo es necesario considerar el balance de gastos en insumos y labores por una parte y por otra el beneficio que se obtiene del cultivo. Esto es buscar los óptimos económicos, para lo que es necesario analizar las estrategias de siembra.

Estas estrategias incluyen, además de una forma de distribuir los propágulos, una serie de labores para el ajuste de la población; las estrategias de siembra pueden ser masivas o dirigidas de acuerdo con la unidad de siembra empleada, la cual es el sitio en el que se requiere establecer un conjunto de una o más plantas.

La magnitud del cultivo que se maneja con ambas estrategias es la misma; centenares o miles de plantas, pero en una estrategia masiva, la unidad de siembra es una parcela o un surco donde se establece la población, sin que importe que se distribuya aleatoriamente, como ocurre en las siembras al voleo y a chorrillo. En cambio, en una estrategia dirigida, la unidad de siembra es un punto en una parcela o incluso un recipiente con tierra, pues se requiere distribuir los propágulos y las plantas en forma regular en el terreno, como ocurre en las siembras mateadas y en las realizadas en macetas (Berlijn, 1982).

Para tratar de asegurar que se tenga una planta en cada punto del terreno o en cada maceta, en las siembras dirigidas es frecuente que se deposite más de un propágulo por unidad, lo que obliga a realizar aclareos de las plantas de más que se obtengan.

Como el obtener una o más plantas en cada unidad de siembra es un evento aleatorio, es frecuente que sea necesario transplantar en las unidades que carezcan de plantas (Tinus y Mac Donald, 1979), así como efectuar resiembras.

NOMENCLATURA EMPLEADA

Considerando que es posible que en una siembra se coloquen distintas cantidades de propágulos por unidad (Camacho, 1989) y que cuando se eliminan las unidades que carezcan de plantas es que no se efectuará trasplante, se estableció la siguiente nomenclatura:

M = magnitud de cultivo, o sea, el número de matas que se va a establecer (mata es una unidad de manejo agrícola constituida por varias plantas que crecen juntas o por una sola planta).

n = número de propágulos sembrados por unidad de siembra.

y_i = número de plantas que se obtiene del i'ésimo propágulo depositado en la unidad de siembra.

X_n = cantidad de plantas que se obtiene en la unidad de siembra en la que se colocaron n propágulos.

$P_{X_{n_1} = i}$ = probabilidad de obtener i plantas al sembrar n_1 propágulos.

U = total de unidades de siembra = $\sum_{i=1}^{n} U_{n_i}$

B_{n_i} = total de plantas que se obtienen al sembrar n_i propágulos en U_{n_i} unidades de siembra.

$W_{X=i}$ = cantidad de unidades que tienen i plantas.

A_{n_i} = plantas a aclarear, o sea plantas de más que se tienen al sembrar U_{n_i} unidades con n_i propágulos.

T = total de propágulos requeridos para una siembra expresado en número.

K = Kg de propágulos requeridos para una siembra.

DESARROLLO

La función de rendimiento que se requiere debe relacionar el número de propágulos sembrados con la cantidad de plantas que se obtienen, por lo que se expresa así:

$$X = \sum_{i=1}^{n} y_i$$

Para desarrollar la función se hacen los siguientes supuestos:

a) Cada propágulo que prende produce sólo una planta, por lo que la variable y_i tiene dos resultados mutuamente excluyentes:

$$y_i = 0 \quad y \quad y_i = 1$$

b) El valor que tenga y_i en el iésimo propágulo no es afectado por lo que ocurra en los demás propágulos colocados en la misma o en otra unidad de siembra.

Como consecuencia se tiene que, dependiendo de cuáles propágulos prendan, hay $m \geq 1$ formas de que X tome un valor entero en el intervalo de

cero a n. Por lo tanto, considerando todas las combinaciones, se tienen 2^n resultados posibles al sembrar n propágulos.

Lo anterior se hace más evidente al determinar los valores de m dado n, para lo que se usa el triángulo de Pascal que se tiene en los textos de estadística general.

c) En todos los lotes comerciales y para siembra, los propágulos que prenden y los que no, sólo se diferencian en esta característica, y están bien mezclados unos con otros. Por lo tanto, los valores de y_i y de X son dados por el azar y cada resultado tiene una probabilidd de ocurrir.

d) Aunque el número de propágulos que integran el lote de siembra se puede considerar infinito, para fines prácticos, se conoce la proporción p de propágulos que pueden prender.

e) La probabilidad de prendimiento de todos los propágulos que participan en una siembra es constante y no es afectada por el número de éstos que se tengan en la unidad de siembra.

Además de la colocación de los propágulos en las unidades de siembra, lo que impide una distribución uniforme de las plantas en un cultivo es la falta de prendimiento de algunos propágulos; la reducción de las fallas mediante el incremento de: los propágulos colocados por unidad de siembra, la cantidad de éstas y el número de transplantes a realizar influye en los costos de producción, por ello se procedió al planteamiento de criterios de optimización económica concordantes con las características de los tipos de siembra.

La seguridad de obtener una cantidad de plantas cercana a la media de la distribución de probabilidades obtenida como función de rendimiento, se evaluó para el caso de una varianza máxima con valores de n desde algunos

centenares hasta varios miles de propágulos. Para hacer esto se consideraron valores cercanos a la media, a los que diferían de ésta desde cero hasta "f", una cantidad menor o igual al 5% de la media.

La determinación de probabilidades se hizo aplicando el teorema central del límite con el fin de usar una normal con media cero y varianza uno (Infante y Zárate, 1984); las áreas bajo esta curva se calcularon con el programa en BASIC presentado por Poole y Cols (1981).

RESULTADOS

Como en la práctica lo que importa es el número de propágulos obtenidos y no cuáles prendan, la función de rendimiento requerida es:

$$P_{X=i} = \binom{n}{i} p^i q^{n-i} \qquad (3)$$

Donde:

q = 1-p = la probabilidad de que un propágulo no prenda.

Esta fórmula corresponde a la función binomial de probabilidades o de Bernoulli. A pesar que en su aplicación requiere el laborioso cálculo de combinaciones, la determinación del número de unidades sin plantas y el de las que tienen cuando menos una planta es sencillo ya que:

$$P_{X_{n_i} = 0} = \binom{n_i}{0} p^0 q^{n_i} = q^{n_i} \qquad (4)$$

y por lo tanto:

$$P_{X_{n_i} > 0} = 1-q^{n_i} \qquad (5)$$

Esta última ecuación da la clave para el establecimiento de los criterios de optimización económica, pues indica que la probabilidad de tener unidades que tengan cuando menos una planta, se aproxima asintóticamente a la unidad conforme se incrementan los propágulos sembrados por unidad. Así,

si no se hace transplante o se aumenta el número de unidades de siembra no es posible asegurar una magnitud de cultivo dada, el incremento del número de propágulos sembrados hace aumentar los costos tanto por la mayor cantidad de insumos requerida como por el incremento de las plantas de más que se tienen que aclarear.

Por lo tanto en las siembras mateadas en las que no es posible asegurar una magnitud de cultivo mediante el transplante, el criterio de optimización económica es la maximización de la ganancia; mientras que en las siembras en macetas en las que la magnitud de cultivo se asegura transplantando algunas plantas que estén demás o sembrando un número mayor de unidades, el criterio de optimización es la minimización de costos para un nivel de producción dado.

Respecto a la seguridad de obtener cantidades de plantas cercanas a la media de la distribución en condiciones de máxima varianza, se encontró que la probabilidad de obtener valores cercanos a la media en un 5% de ésta, es superior al 90% cuando se trabaja con unos cuantos miles de propágulos (Cuadro 1). Es evidente que con prendimientos distintos al 50%, se alcanzan altos valores de probabilidad con un menor número de propágulos que los que se requieren en condiciones de máxima varianza.

DISCUSION

Es costumbre que el cálculo de las necesidades de propágulos se realice con la siguiente fórmula (Boyd, 1969; Hartmann y Kester, 1971 y Patiño y Cols., 1983):

$$K = M/V \qquad (6)$$

Con lo que se demuestra en el cuadro 1, se tiene que esta ecuación es útil cuando se realizan siembras masivas, en las que importa más el número de plantas obtenido que su distribución en el terreno y no se hacen labores para ajustar la población. Para estas siembras no tiene sentido el uso de la distribución binomial.

En cambio, para las siembras dirigidas, esta función de rendimiento es importante para determinar el valor de n_i óptimo, pues lógicamente la cantidad de propágulos requerida está dada por:

$$T = U_{n_i} n_i \quad . \tag{7}$$

y por

$$K = T/(S\ I) \tag{8}$$

La determinación del valor de n_i óptimo requiere del cálculo de los costos de la siembra, los cuales están dados por la siguiente ecuación general obtenida de los ejemplos que presentan Tinus y Mac Donald (1979):

$$C = c_k K + c_t T + c_e E + c_a A + c_r R + c_u \sum U_{n_i} \tag{9}$$

Donde:

C = costo total de la siembra.
c = costo parcial, esto es el del concepto que expresa el subíndice.
T = propágulos a sembrar (se considera el costo de la siembra de cada propágulo).
E = unidades sin plantas o fallas (se considera el costo de eliminarlas)
A = plantas de más (se considera el costo de aclararlas).
R = plantas a transplantar en unidades vacías. (Este concepto es excluyente del E; lo contrario también se cumple).

Una cantidad necesaria para obtener los valores de estos conceptos es el total de plantas que se obtienen, para lo que Tinus y Mac Donald (1979) usan una sumatoria de los valores de $U_{n_1} P_{X_{n_1}=i}$; la cual no es necesaria ya

Cuadro **1.** Efecto de n sobre la probabilidad de obtener un valor cercano a np en f, con p = 0.5.

Semilla por unidad de siembra (n)	np	$J^* = (npf + 0.5) / (npq)^{1/2}$ f = 0.01		f = 0.05	
		npf	P (-J ≤ Z ≤ J)	npf	P (-J ≤ Z ≤ J)
400	200	2	0.197	10	0.706
1,000	500	5	0.272	25	0.893
5,000	2,500	25	0.529	125	0.999
10,000	5,000	50	0.688	250	0.999
50,000	25,000	250	0.975	1250	0.999
100,000	50,000	500	0.998	2500	0.999

* Se respetó la corrección por continuidad apesar de su poco efecto en la probabilidad.

que se cumple que:

$$B_{n_i} = U_{n_i}\, n_i\, p \qquad (10)$$

tiene una probabilidad de ocurrencia alta con unos cuantos miles de propágulos.

Con lo obtenido en el presente se tienen las bases para el cálculo de los costos de siembra sin recurrir a las tablas de probabilidad que presentan Tinus y Mac Donald (1979), pues las fórmulas 4, 5 y 10 son fáciles de memorizar y operar. Consecuentemente las cantidades requeridas de determinan así:

I) Unidades a sembrar:

a) Sin transplante: $U_{n_i} = M$ (11)

b) Con transplante: $U_{n_i} = M/(1-q^{n_i})$ (12)

II) Unidades sin plantas, para determinar E y T:

$$W_{X_{n_i}=0} = U_{n_i} q^{n_i} \qquad (13)$$

III) Plantas que se aclarearán:

$$A_{n_i} = B_{n_i} - U_{n_i}(1-q^{n_i}) - R \qquad (14)$$

La nomenclatura empleada para el estudio de las siembras dirigidas puede parecer innecesariamente compleja, aceptando que sólo se emplee un número de propágulos para sembrar las unidades de siembra; no obstante, se presentó esta nomenclatura con el fin de que en futuros trabajos se explore la conveniencia de usar diferentes valores de n, como lo efectuó Camacho (1989).

BIBLIOGRAFIA

Boyd, c. w. (1969) "A better estimation of nursery survival used in sowling formula". *Tree Plant*. Not 20(3): 21-25.

Berlijn, J. (1982) "Maquinaria para fertilización, siembra y transplante". *Manuales para la Educ. Agrop.* SEP/Trillas, México, 74 p.

Camacho, M. F. (1989) "Fórmula para reducir requerimientos de semillas en siembras directas en viveros forestales". Congr. Forestal Mexicano. PROTIMBOS/ANCF. Tomo II, pp. 359-361.

Hartmann, H. T. y Kester, D. E. (1971). *Propagación de plantas; principios y prácticas*. Trad. Marino, A. CECSA, México, pp. 216.

Infante, G. S. y Zárate, G. (1984) *Métodos Estadísticos; un enfoque interdisplinario.* Trillas. México, pp. 260-264.

Patiño, V. F.; Garza, P. de la; Villagómez, A. Y.; Talavera, A. I. y Camacho M. F. (1983) *Guía para la recolección y manejo de semillas forestales.* INIF Bol. Div. No. 63. México, pp. 161-163.

Poole, L.; Barchers, M. y Castlewitz, D. (1981) *Algunos programas de uso común en BASIC*; edición para Apple II. Osborne/McGraw-Hill, México, pp. 257.

Tinus, R. W. y Mac Donald, S. E. (1979) *How grow tree seedings in containers in greenhouses.* USDA. For. Serv. Gral. Tech. Rep. RM-60, USA, pp. 142-148.

EL PROBLEMA DE LOS COMPONENTES DE VARIANZA NEGATIVOS EN ESTUDIOS DE VARIACION DE ESPECIES FORESTALES

Francisco Camacho Morfin
Felipe Nepamuceno Martínez
Pilar de la Garza López de L.
Coord. de Germoplasma Forestal
CIFAP - D. F. INIFAP
Avenida Progreso No. 5, Coyoacán, D. F.
C. P. 04000

RESUMEN

Una situación que suele presentarse en los estudios de variación de las especies, es que al despejar los componentes de variación de algún nivel de muestreo, se tengan datos con signo negativo, lo que causa desconcierto, pues la suma de cuadrados que se hace para obtener una varianza no permite que se obtengan resultados negativos. En este trabajo se presenta una interpretación estadística del problema, en la que se argumenta que los componentes negativos de varianza resultan de violaciones de los supuestos requeridos para los análisis de varianza; se sugiere el uso de transformaciones para solucionar estas irregularidades.

INTRODUCCION

Una fase fundamental para llevar a cabo la mejora genética de las especies forestales es determinar cómo se distribuye la variación, tanto de los caracteres de interés económico como la de los de importancia biosistemática (Nepamuceno y Cols. 1989). Una aproximación a este conocimiento se realiza a través de los análisis estadísticos de componentes de la varianza, los cuales asumen variables totalmente aleatorias.

Una situación que suele presentarse en los estudios de variación de las especies, es que al despejar los componentes de variación de algún nivel de

muestreo, se tengan datos con signo negativo, lo que causa desconcierto, pues la suma de cuadrados que se hace para obtener una varianza no permite que se obtengan resultados negativos.

Ante la presencia de componentes de varianza negativos se ha asumido que: a) No es procedente calcular los componentes de varianza para la variable estudiada (Moreno, 1985) y, b) Consignar con un valor de cero el componente respectivo y desarrollar los otros componentes sobre esta base (Morgenstern 1969, Pérez y Eguiluz 1985).

En este trabajo se presenta una interpretación adecuada del problema con base en consideraciones estadísticas.

REVISION DE DATOS RELACIONADOS CON EL PROBLEMA

Los datos usados en el estudio proceden de un muestreo de estructuras vegetativas y reproductivas de **Pinus montezumae** Lamb., las cuales se tomaron de 10 árboles de cada una de cuatro procedencias (Chiautzingo, Pue., Río Frío, Villa Victoria y San Felipe del Progreso, Méx.) El material fue colectado y evaluado por De la Garza y Nepamuceno (1989).

Los resultados obtenidos se sometieron al análisis de varianza de acuerdo con el modelo II para un muestreo jerárquico (Snedecor y Cochran, 1971).

El despeje de los componentes de variación entre poblaciones y entre árboles dentro de poblaciones, se realizó de acuerdo con las esperanzas de los cuadrados medios del modelo usado (Cuadro 1).

Entre las variables evaluadas, el ancho del ala de la semilla, el de la apófisis de las escamas del estróbilo, el número de cotiledones y el largo del hipocotilo en plantas de tres meses de edad, obtuvieron componentes de varianza negativos para alguno de los niveles de muestreo estudiados, que en

todos los casos presentó relaciones de varianzas (F) menores de uno. En otras variables medidas en las semillas, sus alas, los conos y plántulas, los componentes de varianza fueron positivos y la F calculada tuvo valores mayores a la unidad, independientemente de su significancia. En todos los análisis realizados, nunca se obtuvieron resultados negativos para el nivel de muestreo más bajo.

Cuadro 1

Esperanzas de cuadrados medios para niveles de muestreo jerárquico (De Snedecor y Cochran 1971)

FUENTE DE VARIACION	ESPERANZA DE CUADRADOS MEDIOS
Poblaciones (A)	$V_E^2 + NV_B^2 + BNV_A^2$
Arboles en poblaciones (B)	$V_E^2 + NV_B^2$
Dentro de árboles (E)	V_E^2

V = Varianza respectiva del subíndice
N = Repeticiones realizadas en el nivel más bajo de muestreo

La presencia de una F menor a la unidad cuando se tienen componentes de varianza negativos, también se encontró en los resultados de Moreno (1985) y en los de Pérez y Eguiluz (1985).

La asociación de los componentes de varianza negativos y valores de F menores a la unidad no es sorprendente, ya que para que ocurran ambos es necesario que el cuadrado medio tomado como numerador sea menor que el usado como denominador.

CONSIDERACIONES ESTADISTICAS

Ostle (1965) menciona que cuando en un análisis de varianza se obtiene una F menor a uno, se puede decir simplemente que no es significativa, es decir, que el componente analizado no aporta una cantidad importante de variación. Sin embargo, no es prudente una excusa tan simple del problema, ya que se está ignorando una advertencia valiosa, pues suele suceder que el inverso de la F sí resulte significativo con los grados de libertad correspondientes, lo que implica que debe rechazarse algo, que en este caso resulta ser el modelo, pues no se satisfacen las suposiciones en las que se fundamenta; este autor considera que se deben revisar éstas en cuanto a: el procedimiento de toma de muestras y datos, la normalidad de éstos y la homogeneidad de varianzas, entre los puntos principales.

En muchos casos es posible que el análisis de los datos transformados a: arco seno $(\text{porcentaje}/100)^{0.5}$, logaritmo o raíz cuadrada ayude al cumplimiento de los requerimientos del modelo estadístico empleado. El uso de transformaciones de los datos no debe despertar suspicacia acerca de una manipulación dolosa de la información, pues no hay argumento científico sólido que rechace un cambio de escala, si éste hace válido el uso de un modelo matemático útil (Sokal y Rohlf 1979). Incluso se ha evaluado el empleo de transformación de los datos a rangos o sea, trabajar con información ordinal, con objeto de satisfacer los supuestos requeridos en el análisis de varianza (Conover e Iman, 1981).

Quizá aún con el empleo de las transformaciones se sigan obteniendo componentes de varianza negativos, éstos podrían tomarse como cero de manera válida si: a) Se ha cumplido con los supuestos, lo que se debe evaluar con las pruebas pertinentes; b) El dato es ligeramente negativo, considerando

como condición para que esto se cumpla, que el inverso de la relación de varianzas con sus respectivos grados de libertad no sea significativo.

Es importante señalar que la ausencia de componentes negativos y por tanto de valores de F menores a la unidad, no asegura el cumplimiento de los supuestos requeridos para realizar un análisis de varianza válido.

EJEMPLO NUMERICO

Como los muestreos realizados incluyen gran cantidad de datos (De la Garza y Nepamuceno, 1989), se diseñó un ejemplo del manejo de los componentes negativos de varianza que ocupará poco espacio. (Anexo).

En el cuadro 2 se observa que los datos violan el supuesto de homogeneidad de varianzas, que se obtiene un componente de varianza negativo y es evidente que una de las F es menor a la unidad. Al aplicar la tranformación logaritmo base 10 de los datos se consiguió que las varianzas se homogeneizaran y que ningún componente resultara negativo, así como que no se tuvieran F menores a la unidad.

SUGERENCIAS

De los resultados presentados obtenidos a partir del ejemplo se tiene un nuevo llamado de atención acerca de:

a) Evaluar el cumplimiento de los supuestos del análisis de varianza como parte integral del procesamiento estadístico de los estudios de variación. Esta recomendación se aplica a cualquier estudio que requiera de realizar análisis de varianza.

b) Ante la presencia de componentes negativos y F menores a la unidad, probar el uso de transformaciones de los datos junto con nuevas evaluaciones del cumplimiento de los supuestos.

Cuadro 2 Componentes de varianza para los datos del anexo.

Fuente de variación.	Grados de libertad.	Datos no transformados		Transformación a logaritmo	
		Cuadrado medio	Varianza del componente	Cuadrado medio	Varianza del componente.
Poblaciones	2	5416,058.00	675,371.75	10.74	1.34
Arboles en poblaciones	3	13,084.00	-692.51	0.01	6.53×10^{-4}
Dentro de árboles.	18	15,854.03	15,854.03	4.12×10^{-3}	4.12×10^{-3}
Prueba de homogeneidad de varianzas*		81 . 87		3.15	

*Prueba de Barttlet.

c) Si la situación no se regulariza con lo anterior, hay que revisar el procedimiento de acopio de los datos para tomarlos nuevamente de manera correcta.

BIBLIOGRAFIA

Conover, W. J. and Iman, R. L. (1981). "Rank transformations as a bridge between parametric and no parametric statistics". JASA 35(3):124-133

De la Garza, L. P. y Nepamuceno, M. F. (1989)."Evaluación de la variabilidad genética en coníferas mexicanas: **Pinus montezumae** Lamb". I Congr. Forestal Mexicano. ANCF. Tomo II México, pp. 883-891.

Moreno, B. G. (1985) "Variación morfológica en **Pinus pseudostrobus** Lind". Mem. de la III Reunión Nal. sobre Plant. Forestales. SARH. Public. Esp. No. 48, México, pp. 219-244.

Morgenstern, E. K. (1969) "Genetic variation in seedlings of **Picea mariana** (Mill.), B.S.P. II Variation Patterns". *Silvae Genet.* 18:5-6.

Nepamuceno, M. F.; Cuevas, R. A.; de la Garza, L. P. y Camacho, M. F. (1989) "Prospección, conservación y evaluación de germoplasma forestal". Congr. Forestal Mexicano. PROTIMBOS/ANCF. Tomo II. México, pp. 951-958.

Ostle, B. (1965), *Estadística aplicada; técnicas de la estadística moderna, cuándo y dónde aplicarlas*. Td. de la Serna, V. D. Limusa. México, pp. 245-270.

Pérez, R. P. M. y Eguiluz, P. T. (1985) "Variación morfológica en **Pinus hartwegii** del eje neovolcánico". Mem. de la III Reun. Nal. sobre Plant. Forestales. SARH. Public. Esp. No. 48, México, pp. 245-270

Sokal, R. R. y Rohlf, F. J. (1979) *Biometría; principios y métodos estadísticos en la investigación biológica*. Td. Lahoz, L. M. Blume, España, pp. 402-426.

Snedecor, G. W. y Cochran, W. G. (1975). *Métodos Estadísticos*. Td. Reinosa, F. J., CECSA, México, pp. 354-358

ANEXO. Datos utilizados para el cálculo de componentes de varianza.

		Mediciones en individuos.			
		I	II	III	IV
Poblaciones	Individuos				
I	1	1765	1485	1535	1275
	2	1655	1355	1405	1085
II	1	9.4	7.8	8.0	6.8
	2	8.2	9.0	9.2	9.6
III	1	33	23	35	25
	2	38	30	38	30

LAS TRANSFORMACIONES LOGARITMICAS EN ARREGLOS TABULARES DE DATOS Y SUS FUNCIONES FACTORIALES

CASANOVA del ANGEL Francisco
Instituto Politécnico Nacional
Acueducto de Guadalupe No. 125
Col. Acueducto de Guadalupe
07270 - México, D. F.

1. INTRODUCCION

Se presentan aquí, con toda generalidad, las aplicaciones logarítmicas a las cuales las tablas de datos pueden someterse. La sola condición consiste en tener datos más grandes que cero y diferentes de uno; es decir, la aplicación se hace sobre tablas de datos no lógicas.

Después de la presentación del modelo logarítmico general, se encuentra su característica más particular. La tabla de perfiles que se obtiene, a partir de una aplicación de tipo logarítmico, es siempre la misma. La elección de la base de la transformación no es relevante.

2. EL MODELO LOGARITMICO MULTIDIMENSIONAL EN BASE a Y BASE e

Consideremos una tabla de valores positivos V_{IP} . Se definen las notaciones y en seguida, algunas propiedades del modelo logarítmico en base a.

Sea

$$V_{IP} =: \{ v_{ip} \mid i \in I, p \in P \} \qquad 2.1$$

la correspondencia de dos conjuntos I y P. Sea L una aplicación logarítmica en base a, definida por:

$$L : V_{IP} \text{ ---------> } V^{LOG} \qquad 2.2$$

Los nuevos conjuntos I_L y P_L son definidos por:

$$I_L = \{ \text{Log}_a(V_{IP}) \mid i \in I \} = \{ \text{Log}_a(v_{ip}) \mid i \in I \} \qquad 2.3$$

$$P_L = \{ \text{Log}_a(V_{IP}) \mid p \in P \} = \{ \text{Log}_a(v_{ip}) \mid p \in P \}$$

Se designa por:

$$f_I^{LOG_a} = \{ f_{i.} \mid i \in I_L \} \qquad 2.4$$

$$f_P^{LOG_a} = \{ f_{.p} \mid p \in P_L \}$$

a los margenes de la tabla de probabilidades del modelo logaritmico en base a. Recuérdese que $f_{i.}$ (o $f_{.p}$) son las masas, los pesos o las frecuencias de los individuos (o variables) del estudio.

En el análisis factorial de correspondencias de V_{IP}, el centro de gravedad del elemento i_1 de I_L (resp. p_1 de P_L) afectado de las masas $f_{i.}^{LOG_a}$ (resp. $f_{.j}^{LOG_a}$) es idéntico al centro de gravedad de todas las i de I (resp. p de P).

Se denota el conjunto de las f^i para cada p de P como el perfil:

$$f^i_{P_L} = \{ \text{LOG}_a(f^i_p) \mid p \in P_L \} = \{ \frac{\text{LOG}_a(v(i,p))}{\text{LOG}_a(v(i))} \mid p \in P_L \} \qquad 2.5$$

después de la transformación 2.2. $f_P^{LOG_a}$ y $f_I^{LOG_a}$ son los perfiles

de la columna de margen y línea de margen.

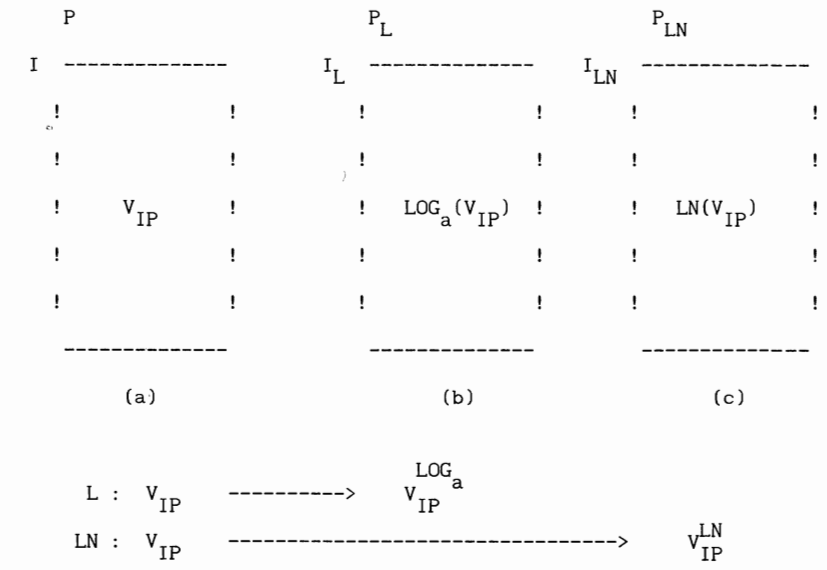


Figura 1: a) Tabla de datos iniciales. b) Tabla bajo la aplicación logarítmica en base a. c) Tabla bajo la aplicación logarítmica en base e.

Tomando a = e, se designa por f_I^{LN} y por f_P^{LN} los márgenes de la tabla de probabilidades del modelo logarítmico en base e, de la misma manera que en 2.4, también conocidos como los perfiles de la línea y de la columna de margen.

En el análisis factorial de correspondencias de V_{IP} , el centro de gravedad de i de I (resp. p de P) afectados de las masas $f_{i.}$ (resp. $f_{.j}$) es idéntico al centro de gravedad de todos los i de I (resp. p de P).

3. CARACTERISTICAS DEL MODELO LOGARITMICO MULTIDIMENSIONAL .

Teorema. El conjunto de perfiles obtenidos bajo una transformación logarítmica aplicada a la tabla de valores positivos y estrictamente diferentes de uno, es siempre la misma sin tener en cuenta la base de la transformación logaritmica.

Demostración:

Sea V_{IP} una tabla de valores positivos definidos según la ecuación 2.1. Sean L_a& L_b las aplicaciones logaritmicas en base a y en base b, definidas como en la ecuación 2.2, respectivamente. Sea $v(i,p) > 0$; a y b son valores positivos diferentes de 1. Recuérdese que por definición:

$$\text{Log}_a\ v(i,p) = \{ \frac{LN\ (\ v(i,p))}{LN\ (a)} \mid i \in I,\ p \in P \} \qquad 3.1$$

aplicando 2.2 a 2.5, se tiene que:

$$f^i_{P_{L_b}} = \{ \text{Log}_b(f^i_p) \mid p \in P \} = \{ \text{Log}_b(v(i,p)) \ / \ \text{Log}_b(v(i))) \mid p \in P \} =$$

$$= \frac{(\ \text{Log}_a(v(i,p)) \ / \ \text{Log}_a(b))}{(\ \text{Log}_a(v(i,p)) \ / \ \text{Log}_a(b))} = \{ \frac{\text{Log}_a(v(i,p)\ )}{\text{Log}_a\ (\ v(i)\ )} \mid p \in P \} = f^i_{P_{L_a}}$$

es decir, las tablas de perfiles son iguales: ($f^i_{P_{L_a}} = f^i_{P_{L_b}}$)

QED

Corolario 1. El conjunto de perfiles obtenidos bajo la aplicación logaritmica puede obtenerse tomando a e como la base de la aplicación. A partir del teorema anterior:

$$f^i_{P_{L_a}} = f^i_{P_{L_e}} = f^i_{P_{LN}}$$

de donde, es suficiente calcular $f^{i}_{P_{L_e}}$

Demostración:

Dado un valor más grande que uno, se tiene que:

$$f^{i}_{P_{L_a}} = \{ (\mathrm{Log}_a (v(i,p)) / \mathrm{Log}_a (v(i))) \mid p \in P \} =$$

$$= \{ \frac{(LN (v(i,p)) / LN (a))}{(LN (v(i)) / LN (a))} \mid p \in P \}$$

$$= \{ \frac{LN (v(i,p))}{LN (v (i))} \mid p \in P \} = f^{i}_{P_{LN}}$$

QED

Corolario 2. La aplicación logarítmica no puede hacerse sobre una tabla de descripción lógica.

Demostración:

Sea V_{IP} una tabla de valores positivos definida según la ecuación 1.1. Sea L una aplicación logaritmica en base a, definida como en la ecuación 2.2.

Sea:

v(i,p) = 1 si el individuo i posee la propiedad p

y

v(i,p) = 0 si el individuo no posee la propiedad p

Como $a^0 = 1$ entonces $\mathrm{Log}_a 1 = 0$. Lo que significa que si se aplica la transformación L_a sobre los valores 1 se obtiene una tabla de valores indefinidos.

QED

4. UN EJEMPLO DE APLICACION

Los datos y las variables en estudio.

Los datos muestran el total de votos obtenidos por cada partido político en las elecciones presidenciales en México, el 6 de junio de 1988. Con ellos, han sido calificadas las elecciones, según el artículo 24 de la Ley Orgánica del Congreso General del país.

Las variables en análisis son diez, los ocho partidos políticos: Partido Acción Nacional (PAN), Partido Revolucionario Institucional (PRI), Partido Popular Socialista (PPS), Partido Mexicano Socialista (PMS), Partido Revolucionario de los Trabajadores (PRT), Frente Democrático Cardenista (FDC), Partido Auténtico de la Revolución Mexicana (PARM) y Partido Demócrata Mexicano (PDM), una variable llamada NREG, o de no-registro, donde se encuentran incluidos los votos para los candidatos comunes y una variable conteniendo la cantidad de votos por distrito (VTOT). Los valores numéricos y su transformación logarítmica, son la votación obtenida por cada partido, en cada uno de los 300 distritos electorales en los cuales el país fue dividido de antemano. Se han adicionado a la tabla los subtotales, por Estado Federal con la idea de ver las posibles influencias de éstos sobre el total nacional.

Los análisis.

Los datos han sido cuidadosamente revisados, encontrándose que algunas sumas hechas con la computadora (y rectificadas a mano), no están de acuerdo con las emitidas oficialmente. Esto, tanto por distrito como para en el total nacional. Un ejemplo lo es la suma, distrito por distrito, que obtuvo el Partido Revolucionario Institucional, PRI,

de 9'628,319 (nueve millones seiscientos veintiocho mil trescientos diecinueve) votos. La suma de los subtotales da 9'890,926 (nueve millones ochocientos noventa mil novecientos veintiseis) votos en total. Es decir, una diferencia de 262,607 (doscientos sesenta y dos mil seiscientos siete) votos. Y la cifra marcada como total de votos del PRI, es de 9'687,926 (nueve millones seiscientos ochenta y siete mil novecientos veintiseis) votos.

Los datos son el universo de la votación presidencial obtenida por los partidos políticos, y no se pueden tomar como una muestra de la población total cuyo marco demográfico es: población empadronada 38,075 (en miles) y población en edad de votar 42,104 (en miles). Los votos emitidos fueron, sumando distrito a distrito, 19'000,632. Sumando los subtotales, 19'324,189. Sumando los totales, 19'106,176. Aunque no está claro si los votos declarados irregulares por el Colegio Electoral se descontaron de los totales publicados, la abstención aproximada es: para la votación distrito por distrito, de 48.803 %; para totales, de 50.080%.

El promedio de las desviaciones de los distritos alrededor de la media de la distribución de votos sirve para medir variabilidad. Comparándolo con el promedio de las desviaciones de los distritos, pero de toda la serie distrital de votos, se ve que obtiene una relación positiva únicamente el Partido Acción Nacional y el Partido Revolucionario Institucional. Siendo mayor en éste último. La relación más negativa la tiene el Partido Mexicano Socialista.

Otra estadística elemental es la varianza relativa a la muestra que es una estimación no sesgada de la varianza de la serie distrital de votos. Quien tiene una varianza muestral menor que la varianza de la serie de votos es el Partido Demócrata Mexicano.

Los momentos centrales de la distribución de votos en el país nos dan una idea de la dispersión de dicha votación presidencial. El primer momento ayuda a describir la distribución de la frecuencia de votos. En este estudio se ha centrado la serie de votos distritales razón fundamental por la que se obtiene el primer momento con valor cero para todos los partidos políticos. Pero en el momento dos, que representa la varianza de la serie distrital de votos presidenciales se puede ver que el Partido Revolucionario Institucional acusa la mayor variación de votos seguido por los partidos Mexicano Socialista y Acción Nacional. Quienes menor variación muestran son el Partido Revolucionario de los Trabajadores y el Partido Demócrata Mexicano, lo que puede deberse a su baja votación distrital y nacional.

Con respecto a las correlaciones obtenidas: el PRI no se correlaciona con ningún partido. El PMS y el PPS son los más altamente correlacionados (0.786). Después, el FDC se correlaciona con el PMS (0.690) y por último el FDC con el PPS (0.675). Los menos correlacionados son el

FDC y el PAN asi como el PRI y el PDM seguidos por la no correlación entre el PARM y el PDM.

La figura 2 muestra el círculo de correlaciones de los primeros dos ejes. La primera componente la hacen los tres partidos de izquierda: PMS, PPS y FDC seguidos por la extrema izquierda mexicana, el PRT. Se les contrapone el PRI pero no logra polarizar el eje. Esto se puede definir como la izquierda-centro. En la segunda componente se manifiestan el PAN y el PDM. Se les contrapone el PARM. No muy claramente pero se puede definir el eje como el de derecha.

El círculo parece expresar la tendencia actual y el movimiento del voto popular, afirmación que se basa en la cercanía que tienen los partidos al valor extremo derecho del eje horizontal. La tercera componente (no significativa en el análisis y que no se muestra aquí) la domina ampliamente el PRI y se le contrapone el PARM. Contraposición bastante curiosa, principalmente por tener una tradición de afinidad partidista.

Ahora veamos la interpretación de los resultados del análisis factorial de la correspondencia de votos entre partidos políticos y distritos electorales del país pero a valores logarítmicos.

En la figura 3, y sólo para fines de interpretación de resultados, pueden identificarse las siglas de los partidos políticos, y los nombres de los distritos electorales más importantes (para cada partido político). Las figuras 2 y 3, han sido obtenidas de dos análisis diferentes, ACP y AFC a partir de tablas a valores logarítmicos.

De entrada se nota la forma triangular de la figura. En cada vértice del triángulo se encuentra un partido político rodeado por los distri-

tos electorales que le fueron significativos el seis de julio pasado, así como por otros con menor contribución al resultado final. Al igual que en el caso del círculo de correlaciones, se presenta aquí un primer factor que se puede denominar de tendencia centro-izquierda que es quien dominó las votaciones.

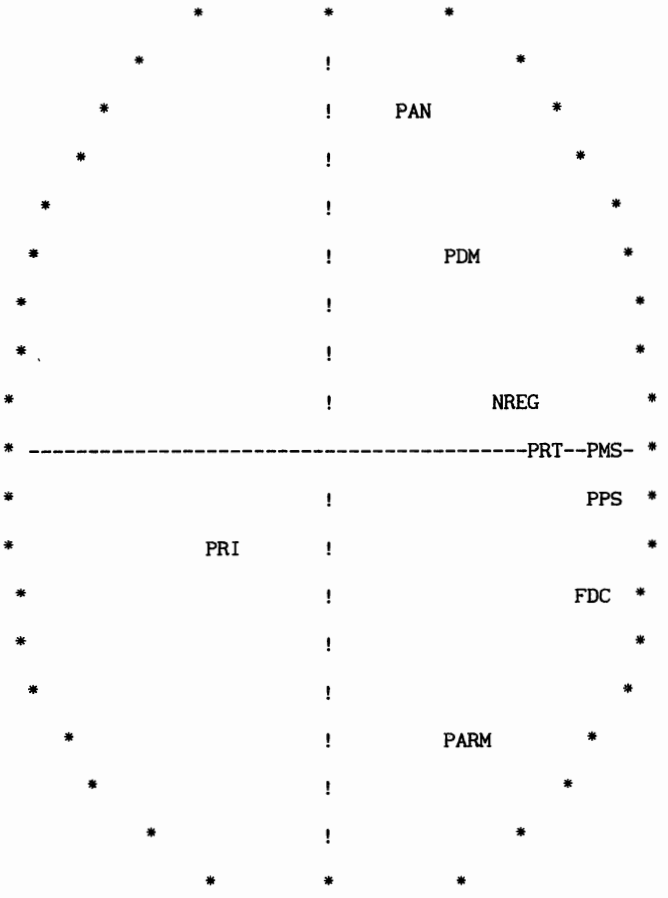


Figura 2: Círculo de correlaciones sobre datos centrados y reducidos. Primeras dos componentes principales. 1era. componente, la izquierda-centro. 2a componente la derecha

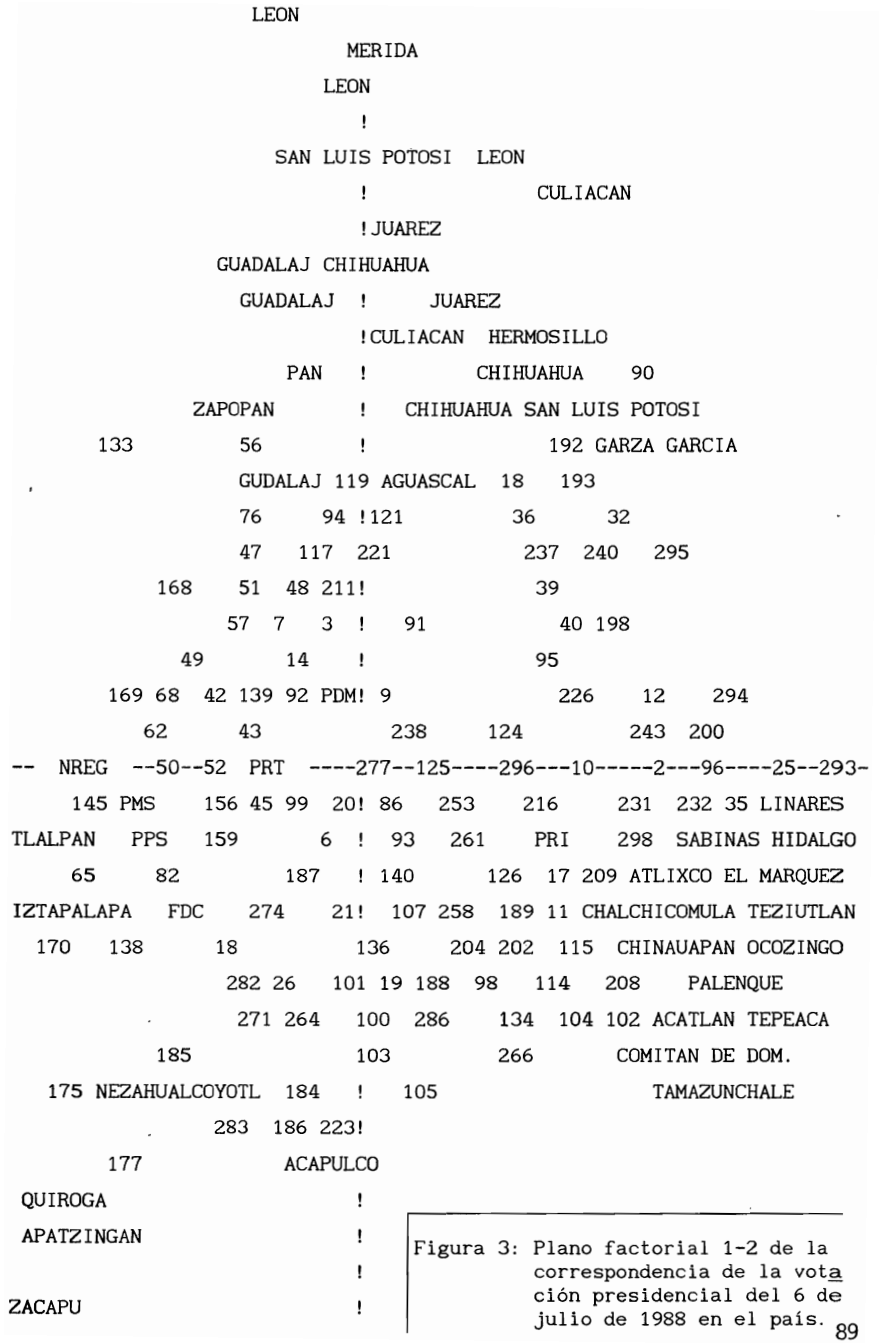


Figura 3: Plano factorial 1-2 de la correspondencia de la votación presidencial del 6 de julio de 1988 en el país.

El centro político, representado por el PRI, ha sido avalado por gran des conjuntos o áreas rurales del país. Entre ellos se pueden mencionar a Sabinas Hidalgo y Linares del estado de Nuevo León; Ocozingo, Palenque, Comitán de Domínguez y San Cristóbal del estado de Chiapas; Tepeaca, Chinahuapan, Teziutlán, Atlixco, Acatlán y Chalchicomula del estado de Puebla, así como El Marquez del estado de Querétaro y Tama zunchale del estado de San Luis Potosí.

La izquierda política mexicana está representada por el FDC, el PPS y el PARM, siendo este último quien hace verdaderamente este factor. Cimentan su avance histórico en las votaciones, y contraposición al actual partido en el poder, en regiones urbanas y suburbanas de no muchos recursos económicos como Iztapalapa, Nezahualcóyotl y Tlalpan en el Distrito Federal; Zacapu, Apatzingán y Quiroga en el estado de Michoacan. Algo que llama mucho la atención es el valor o peso que el PARM obtiene en el estudio. Su votación total es la cuarta, en orden jerárquico, con aproximadamente un 6.32 % (o 6.22 % o 6.29 % según las sumas totales de votos mencionadas anteriormente). Es el 60.01% de la obtenida por el Frente Cardenista. Este fenómeno electoral puede deberse a la uniformidad de su cómputo distrital pues es el que menos se dispara distritalmente al interior de cada entidad.

El segundo factor está dado por el PAN cuya fuerza electoral se circunscribe a tres regiones del país. El norte, representado por Culiacán, Sinaloa; Juárez y Chihuahua, del estado de Chihuahua, y Hermosillo, Sonora. El Occidente en Guadalajara y Zapopan Jalisco así como León, Guanajuato y Aguascalientes, Aguascalientes. Y el sureste

con Mérida, Yucatán. Garza García en Nuevo León tiene también un lugar preponderante en este factor que se le puede definir como el de derecha. Lo atrayente del factor es que al PAN se le opone el PARM y no el PRI.

5. FUNCIONES FACTORIALES LOGARITMICAS

Las funciones que establecen la relación entre los sufragios, que cada partido político ha obtenido, y la cantidad de población en edad de sufragar son hechas a partir de una transformación de funciones de consumo desarrolladas en algunas publicaciones (ver referencias (2) y (3)). Se ha adaptado y experimentado la transformación factorial logarítmica sobre cinco funciones tipo de consumo, con los factores obtenidos de un análisis factorial donde las variables independientes son los factores. Tales funciones (veáse la tabla 1) se someten al ajuste de la regresión una vez hechas las tranformaciones. Ahora bien, si a la tabla de datos solo se le aplica un análisis log-lineal (que es un método para estudiar las relaciones estructurales entre las variables en una tabla de contingencia) significa restringir el análisis de datos.

La obtención teórica de las funciones de la tabla 1, se hace a partir de encuestas realizadas en un momento determinado o sobre el universo mismo (en nuestro caso, se trata de sufragios hechos sobre las urnas). Es evidente que la elección definitiva de la función factorial logarítmica depende de ciertos parámetros, como el ajuste simple y cuadrático, la desviación estándar, los tests y la función de verosi-

Cuadro No 1: Funciones que establecen la relación entre los sufragios y la cantidad de población en edad de votar.

Nombre de la función	Función tipo	Función factorial
LINEAL		
Original	$y = ax + b$	$g(x) = a_0 + \sum_j a_j G_j(x)$
Transformada		
DOBLE LOGARITMICA		
Original	$y = e^a x^b$	$g(x) = e^{a1} \left(\sum_j G_j(x)\, a_j \right)$
Transformada	$\log y = a + b \log x$	$\ln g(x) = a_1 \sum_j a_j \ln G_j(x)$
SEMI-LOGARITMICA		
Original	$e^y = e^a x^b$	$e^{g(x)} = e^{a1} \left(\sum_j G_j(x)\, a_j \right)$
Transformada	$y = a + b \log x$	$g(x) = a_1 + \sum_j a_j \ln G_j(x)$
LOGARITMICA INVERSA		
Original	$y = e^{(a + b/x)}$	$g(x) = e^{(a_1 + a_j / \sum_j G_j(x))}$
Transformada	$\log y = a + \frac{b}{x}$	$\ln g(x) = a_1 - (a_0 / \sum_j G_j(x))$
LOG-LOG INVERSA		
Original	$y = x^c e^{(a + \frac{b}{x})}$	$g(x) = x\, e^{(a_1 + a_0 / \sum_j G_j(x))}$
Transformada	$\log y = a + \frac{b}{x} + c \log x$	$\ln(g(x)) = a_1 - (a_0 / \sum_j G_j(x)) + \sum_j a_j + \ln(G_j(x))$

militud. Se han desarrollado las fórmulas factoriales logarítmicas relativas a un coeficiente de elasticidad y a un índice de crecimiento de votos (que no se muestran en esta comunicación).

De tales funciones se puede decir brevemente que la función lineal se ajusta mal a causa del coeficiente de elasticidad, porque converge a uno, cuando los votos aumentan. Recuerde que los votos se caracterizan por el total de la población en edad de votar.

La función semi-logarítmica, donde el coeficiente de elasticidad calculado de votos es inversamente proporcional a la cantidad de votos emitidos, es utilizada cuando los sufragios son expresados en gran cantidad en relación a los otros partidos políticos. La función log-inversa da buenos resultados cuando se tiene una pequeña diferencia entre los sufragios realizados y la cantidad de población del distrito. Las otras funciones han sido construidas y aplicadas pero como los primeros resultados obtenidos fueron un poco anormales, no se les tomó en cuenta.

Comentarios

La técnica mostrada aquí es extremadamente simple de realizar. Ha sido deducida a partir de una aplicación del análisis factorial de correspondencias simples sobre datos obtenidos sobre las votaciones presidenciales que tuvieron lugar el 6 julio de 1988 en México (la votación más disputada desde la revolución mexicana de 1910).

La característica particular del modelo de datos es la posibilidad de aplicarle un modelo logarítmico, lo que permite la obtención de perfiles y leyes marginales iguales, cualquiera que sea la aplicación logarítmica utilizada.

La interpretación que se hace de los resultados, es más rica que la que se obtiene a partir de los resultados de un análisis factorial sobre datos brutos.

Referencias

1: BENZECRI, J. P: L'analyse des données, T2, l'analyse des correspondances. Deuxième édition. Dunod. 1976. Paris.

2: GOREUX, L. M: Ingresos y consumo de alimentos. Estudios de la FAO sobre Economía y Estadísticas Agrícolas 1952-1977. FAO, Roma. 1978.

3: HOUTHAKKER, H. S: An international Comparation of Household Expediture Patterns. Econometrica, Vol 2, julio de 1957, pp. 532-551.

4: PETER G. M. VAN DER HEIJDEN AND JAN DE LEEUW. Correspondence analysis used complementary to loglinear analysis. Psychometrika Vol 50, no. 4, pp 429-447. December 1985.

5: SECRETARIA DE AGRICULTURA Y RECURSOS HIDRAULICOS-MEXICO: El Desarrollo Agropecuario de México: Pasado y Perspectivas. Tomo XIII. Perspectivas de la demanda y de la oferta de productos agropecuarios. pp. 1-505. Informe 1982.

6: VARAN CHOULAKIAN. Exploratory analysis of contingency tables by loglinear formulation and generalizations of correspondence analysis. Psychometrika Vol 53, no. 2, pp 235-250. June 1988.

7: WEBSTER WELLS, S. B: Nueva Trigonometría Plana y Esférica. D. C. Heath & Co., Ed. 1917.

Optimizacion de multiples respuestas: un enfoque algoritmico

Por

Juan Gaytan Iniestra

Facultad de Ingenieria

Universidad Autonoma del Estado de Mexico

y

ITESM-Campus Toluca

Septiembre de 1989

Resumen

En muchas situaciones experimentales, se estudian varias respuestas para cierta combinación de un grupo de variables de diseño. Cuando se habla de optimización en el contexto de múltiples superficies de respuesta, se desea determinar condiciones que son óptimas para todas las respuestas, aunque esto difícilmente se logra debido a que usualmente las respuestas están en conflicto. En la literatura se han reportado varias estrategias y algoritmos para determinar condiciones "óptimas". En este reporte se hace una breve descripción de las estrategias más conocidas que han sido reportadas para la solución de un problema con más de una respuesta, así como la descripción de los algoritmos de optimización mas relevantes que resuelven los problemas generados de esas estrategias.

Introducción.

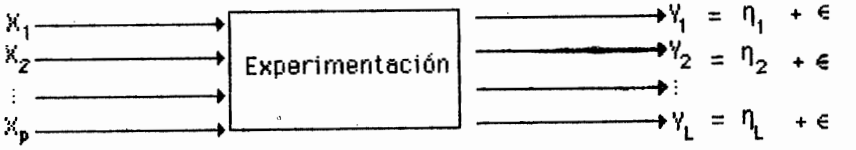
Este reporte es un resumen de las estrategias y sus algoritmos que han sido propuestos durante los últimos años para determinar el óptimo de un conjunto de múltiples superficies de respuesta.

Debido a que no existe una estrategia única para la solución de un problema con más de una respuesta, se describirán las estrategias más populares así como los algoritmos utilizados que dan solución a los modelos generados por esas estategias.

Varios autores han efectuado revisiones de la literatura existente para resolver el problema con una sola superficie, por ejemplo Farrell (1977), Glynn (1986), Meketon (1983), Montgomery y Bettencourt (1976), Rustagi (1981), Smith (1973), Jacobson y Schruben (1987). Aparentemente sólo la revisión de Montgomery y Bettencourt (1977), la cual se enfoca a simulación, ha revisado la literatura existente para el caso de más de una superficie de respuesta.

Aunque no ha sido encontrado un método general que funcione adecuadamente en un conjunto general de varias superficies de respuesta, han sido propuestas varias estrategias en la literatura. Ejemplos de ellos son Biles (1977), Montgomery y Bettencourt (1977), Biles (1978), Harrington (1965), Derringer y Suich (1980), y más recientemente Khuri y Conlon (1981). Las aplicaciones reportadas son en una gran mayoría en el área de simulación y escasas en el área de control de calidad. Ejemplos de esas aplicaciones son los reportes de Derringer y Suich (1980), Taraman (1972), y (1974), Hendrix (1970).

El problema consiste en optimizar un conjunto de respuestas esperadas η_j $(j=1,2,\ldots,L)$, sobre una región S con respecto a un conjunto de P factores de entrada, $X = [x_1\ x_2\ ,\ldots,x_p]^T \in S \subseteq \mathbb{R}^p$. El siguiente diagrama muestra esquemáticamente el problema.



Formalmente, el problema que se desea resolver tiene la estructura mostrada en el siguiente problema (P_o) de Programación No Lineal (PNL) con objetivos múltiples:

$$\begin{aligned} &\text{Max } \eta_1(X) \\ &\text{Max } \eta_2(X) \\ &\quad \vdots \\ &\text{Max } \eta_L(X) \\ &\text{sujeto a} \\ &X \in S \subseteq \mathbb{R}^P \end{aligned}$$

Por lo general el conjunto S es de la forma siguiente: $S=\{X \in \mathbb{R}^P: a_i \leq x_i \leq b_i,\ i=1,2,...,P\}$, donde los números a_i y b_i son escalares reales.

Las funciones n_j indican las respuestas esperadas. La estrategia experimental en el análisis de superficies de respuesta supone que las respuestas η_j son funciones de las variables de diseño $x_1, x_2, ...,x_n$ y que las funciones pueden ser aproximadas en la región S por modelos lineales, cuadráticos, y cúbicos. Cada observación Y_{kj} de la respuesta η_j está formada de una función g_j más una componente aleatoria ε_{kj} como se indica a continuación

$$Y_{kj} = g_j(x_{1k}, x_{2k}, ...,x_{Pk}) + \varepsilon_{kj}; \qquad \begin{matrix} j = 1,2,...,L \\ k = 1,2,...,n_j, \end{matrix}$$

donde g_j denota la relación funcional entre Y_j y $x_1,x_2,...,x_p$. Si hacemos la suposición usual que $E(\varepsilon_{kj}) = 0$ para cada j y k, entonces podemos relacionar la respuesta esperada η_j con las P variables de diseño mediante

$$\eta_j = g_j(x_1,x_2,...,x_P); \qquad j = 1,2,..,L.$$

En la práctica las funciones η_j son desconocidas por lo que son estimadas experimentalmente en la región S por $\hat{Y}_j$ mediante los modelos de primer orden

$$\hat{Y}_j(X) = \hat{\beta}_{oj} + \hat{\beta}_{1j}x_1 + ... + \hat{\beta}_{Pj}x_p \qquad j = 1,2,...,L$$

o bien de segundo orden

$$\hat{Y}_j(X) = \hat{\beta}_{oj} + \sum_i \hat{\alpha}_{ij}x_i + \sum_i \hat{\beta}_{iij}x_i^2 + \sum_k \sum_l \hat{\gamma}_{klj}\, x_k\, x_l \qquad j =1,2,...,L.$$

La estimación de los parámetros es efectuada a través de las técnicas de regresión usuales.

El problema (P_o) tiene dos dificultades importantes:

- Deseamos optimizar $\eta_1(X),\ldots,\eta_L(X)$ pero sólo Y_{kj} puede ser observado, lo cual coloca al problema en uno de optimización estocástica, problema que es sabido muestra dificultad en la determinación de su solución óptima.
- Se desea maximizar simultáneamente varias respuestas que pueden estar en conflicto, por lo que un óptimo en el sentido usual tal vez no exista.

Por lo tanto, la solución de un problema con varios objetivos no solo debe depender de una técnica numérica eficiente, lo cual ya es difícil, sino que debe eliminar o reducir el efecto del ruido para que el algoritmo de optimización pueda localizar la combinación óptima de los niveles de las variables de diseño.

Es conveniente recordar los pasos que son estándar en el análisis de superficies de respuesta: a) ejecutar el diseño estadístico de experimentos, b) estimar los coeficientes de la superficie de respuesta, c) probar la validez de los modelos vía pruebas de falta de ajuste, y d) estudio de la respuesta en la región de interés.

La discusión presentada en este reporte se concentra en el punto d) por lo que nuestro objetivo será mostrar las estrategias de ataque del caso de múltiples objetivos y citar algunos algoritmos reportados en la literatura, ningún intento será hecho aquí en estudiar las restantes etapas.

Clasificamos las diferentes estrategias de optimización en 8 categorías básicas: 1) estrategias gráficas, 2) métodos de los objetivos acotados, 3) método de los objetivos ponderados, 4) criterios globales, 5) métodos basados en preferencias, 6) programación por metas, 7) método min-max, 8) estrategias interactivas.

Métodos Gráficos.

Los métodos gráficos son históricamente los primeros en ser utilizados para analizar el problema de mas de una superficie de respuesta. La estrategia usual consiste en superimponer las respuestas de interés en el mismo conjunto de ejes coordenados y determinar por inspección los valores óptimos de las variables

independientes. Para ilustrar el procedimiento consideramos la figura 1.

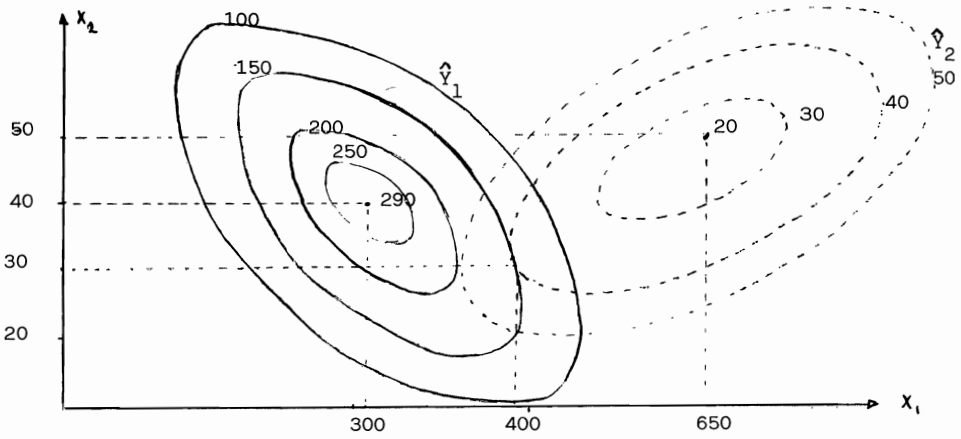


Fig. 1

El óptimo (máximo) de la superficie $\hat{Y}_1$ ocurre en (300,40), y el óptimo (mínimo) de la superficie $\hat{Y}_2$ ocurre en (650,50). Por inspección, el experimentador eligió la mejor solución de "compromiso" (400,30) para un $\hat{Y}_1 = 150$ y $\hat{Y}_2 = 50$.

Cuando se tienen más de dos variables la representación se complica. Una alternativa es construir las curvas de nivel de las respuestas expresadas en función de dos de las variables y dejar fijas las restantes en ciertos valores que se sospecha son adecuados. La estrategia anterior no garantiza que se encuentre una solución satisfactoria.

Entre los autores que presentan ejemplos de la estrategia gráfica se encuentran Lind, Goldin y Hickman (1960), Davies (1956), Hunter (1958), Smith (1963), Lind y Young (1965), Taraman y Lambert (1972).

Por las limitaciones que se tienen a medida que el número de variables independientes se incrementa y a la falta de un método formal para pesar las diferentes respuestas esta estategia es raramente utilizada.

Método de de los objetivos acotados.

Una estrategia popular consiste en seleccionar una de las respuestas como la *respuesta base* y optimizarla sujeta a las demás respuestas (*respuestas secundarias*).

Suponga que la respuesta 1 es la base y las restantes 2,...,L son las secundarias, entonces el problema (P_1) consiste en

$$\text{max o min} \quad \eta_1 = g_1(x_1,\ldots,x_p) \tag{1}$$

$$\text{sujeto a} \quad a_i \leq x_i \leq b_i, \quad i = 1, 2,\ldots,P \tag{2}$$

$$\eta_j = g_j(x_1,\ldots,x_p) \left\{ \begin{matrix} \leq \\ = \\ \geq \end{matrix} \right\} d_j, \quad j = 1,2,\ldots,L \tag{3}$$

Las constantes a_i y b_i en (2) son las cotas de la variable controlable x_i, y las constantes d_j en (3) son cotas superiores o inferiores de valores de las respuestas secundarias propuestas por el experimentador.

La mayor desventaja de esta estrategia radica en la filosofia en la que se basa. A menudo la meta del estudio es obtener el mejor balance entre las diferentes respuestas, pero esto no podría ser alcanzado debido a que las respuestas no tienen asignado el mismo peso. Además, el seleccionar una respuesta base forza al experimentador a asignar valores específicos a las cotas d_j en las respuestas secundarias. Esto es desventajoso si se toma en cuenta que la mayoría de los objetivos tienen diferentes unidades y estan en conflicto en la mayoría de las ocasiones, lo cual hace difícil la elección de valores adecuados de las cotas d_j; elecciones inadecuadas de esas cotas podrían resultar en que el conjunto definido por las restricciones sea vacío.

Varios algoritmos han sido propuestos para resolver el problema (P_1), la gran mayoría hacen uso de la teoría y algoritmos ya existentes de la programación lineal y no lineal. Por ejemplo, si las funciones respuesta son lineales (en las variables de diseño), el problema (P_1) es uno de programación lineal el cual puede ser resuelto eficientemente mediante el algoritmo simplex de G. Dantzig. Hartmann y Beaumont, (1968); Nicholson y Pullen (1969) hacen uso de esta estrategia.

Umland y Smith (1959) aplican los multiplicadores de Lagrange cuando las respuestas son cuadráticas.

Myers y Carter (1973) consideran el problema de dos respuestas cuadráticas y utilizan los multiplicadores de Lagrange junto con el análisis de crestas (ridge analysis) para determinar una gráfica bidimensional de la cual se obtienen las condiciones del máximo restringido independientemente del número de variables de

diseño (aunque limitado a dos respuestas).

Otra estrategia de solución del problema (P_1) que proporciona la programación no lineal es la de penalización. Carroll (1961) utiliza esta estrategia la cual consiste en incorporar las restricciones en la respuesta base por medio de una función de penalización. La estrategia consiste en penalizar la función objetivo cuando las variables independientes se aproximan a las fronteras de la región. es interesante notar que esta estrategia es precursora de los métodos de penalización y barrera desarrollados hacia los fines de los 60's por Fiacco y MacCormick.

Biles (1975, 1977) y Biles y Swain (1980) optimizan la respuesta base sujeta a las respuestas secundarias restringidas dentro de ciertas cotas. El algoritmo de optimización que utiliza es el de la proyección del gradiente de Rosen (Rosen 1961), el cual consiste en utilizar la usual dirección del gradiente en la búsqueda del óptimo. Cuando esa dirección conduce a puntos fuera de la región factible se efectúa una proyección del gradiente sobre la variedad lineal definida por las restricciones activas linealizadas. El algoritmo inicia en un punto factible $\mathbf{X}^k$. Una dirección factible $\mathbf{S}^k$ es definida y un paso de longitud λ^k es tomado en esa dirección a partir de $\mathbf{X}^k$. Ladeterminación de λ^k se logra maximizando la función base cuidando que ninguna de las respuestas secundarias sean violadas. La figura 2 indica el comportamiento del algoritmo (iniciando en el punto $\mathbf{X}^1$= (1,1)) para el siguiente problema hipotético con 2 variables de diseño y tres respuestas debido a Biles (1975). Las tres respuestas se maximizan.

$$\eta_1 = 3x_1^2 + 2x_2^2$$

$$\eta_2 = x_1^2 + x_2^2$$

$$\eta_3 = 9x_1 - x_2^2$$

Eligiendo a la función η_1 como la base el problema restringido correspondiente es

$$\text{Max } \eta_1 = 3x_1^2 + 2x_2^2$$

sujeto a

$$\eta_2 = x_1^2 + x_2^2 \leq 25$$

$$\eta_3 = 9x_1 - x_2^2 \leq 27$$
$$x_1, x_2 \geq 0$$

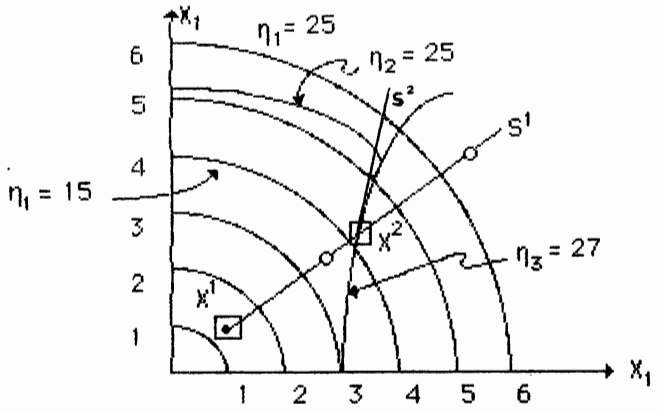


Fig. 2

Otro algoritmo de programación no lineal también elegible para resolver el problema (P_1) es el de Hooke y Jeeves (Hooke y Jeeves 1961), el cual efectúa una exploración local en las direcciones paralelas a los ejes coordenados desde el punto base actual. La búsqueda se inicia intentando un movimiento (hacia adelante o hacia atrás) en la dirección de la variable x_1. Si existe un mejoramiento de la función base, se efectúa un nuevo intento de movimiento en la dirección de la variable x_2 iniciando en el punto que mejoró en la dirección de x_1. Esto se repite para todas las P variables. A continuación se efectúa un paso de aceleración el cual consiste en intentar mejorar la respuesta base en la dirección definida por el punto donde se inició la exploración y el último punto obtenido de la fase explotratoria. La figura 3 muestra el comportamiento del algoritmo para el caso de dos variables de diseño y una respuesta.

Fields (1974) utiliza una variación del método de Hooke y Jeeves para el caso multivariado resolviendo el problema (P_1).

Método de los objetivos ponderados

La dificultad que se tiene al seleccionar la respuesta base en el método de los objetivos acotados así como las cotas d_j en los valores de las respuestas puede ser eliminado asignando un peso w_j a cada una de las respuestas y formulando una sóla función

como sigue

$$(P_2) \qquad \text{Max} \quad F(X) = w_1 \eta_1 + w_2 \eta_2 + \ldots + w_L \eta_L$$

donde los pesos $w_1, w_2, \ldots, w_L$ son cantidades no negativas tales que suman la unidad.

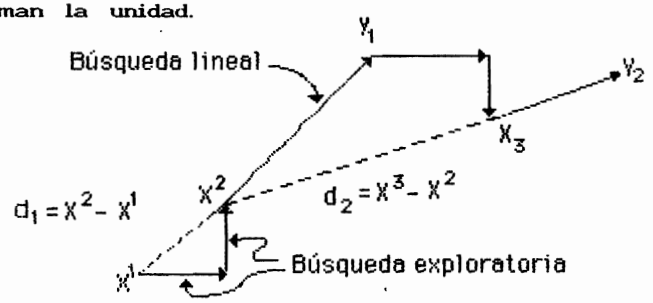


Fig. 3

La ventaja de esta estrategia es el hecho que se maximiza una sóla función, para la cual existen muchos algoritmos para encontrar el óptimo. Su desventaja principal es la elección adecuada de los pesos (w_j). Dado que el resultado obtenido al resolver el modelo anterior depende significantemente de la asignación de los valores de los pesos, y dado que poco se sabe que valores pueden tener esos pesos (fundamentalmente debido a que las unidades de los objetivos son distintas) es necesario resolver el modelo anterior para varios conjuntos de pesos. Únicamente cuando esto ha sido efectuado es posible tener una idea de los valores de los pesos más convenientes. Esto sugiere que para que sea exitoso el método, el analista debe poder interactuar para así poder elegir una combinación satisfactoria. Fields (1974) proporciona un programa de computadora para seleccionar los pesos. Montgomery, Talavage y Mullen (1972) aplican el método para la optimización de una red de tráfico señalizada.

Debido a que las unidades de las respuestas son distintas, una mejora numérica de la solución se obtiene expresando las respuestas aproximadamente en los mismos ordenes de magnitud. esto se obtiene transformando (P_2) como sigue

$$(P_2') \qquad \text{Max} \quad F(X) = \Sigma\, w_j\, \eta_j\, c_j$$

donde a la constante c_j usualmente sele asigna el valor $1/\eta_j(X^*)$, donde $\eta_j(X^*)$ es el valor óptimo irrestricto de la respuesta j.

Se puede demostrar que si los pesos son estrictamente

positivos las soluciones que se encuentran al resolver el problema (P_2) son Pareto óptimas (este concepto se define la siguiente sección).

Soluciones Pareto óptimas

Recordemos que el problema de múltiples respuestas (P_0) puede ser establecido formalmente como sigue

$$\text{Max } (\eta_1(X), \eta_2(X), \ldots, \eta_L(X))$$
$$\text{sujeto a} \quad a_i \leq x_i \leq b_i \quad i=1,2,\ldots,P,$$

para el cual la programación matemática con objetivos múltiples tiene desarrollada una teoría abundante, incluyendo condiciones de optimalidad y algoritmos. Para que sea más provechosa y clara la presentación de las estrategias de solución directa de ese problema consideremos las siguiente definición.

Definición. Una solución factible X^* del problema (P_0) es llamada *no inferior, Pareto óptima, o eficiente* si no existe otra solución factible que produzca un mejoramiento en una de las respuestas sin causar una degradación en al menos otro criterio. En términos matemáticos:

Una solución factible X^* del problema (P_0) es llamada Pareto óptima si no existe otra solución factible, digamos Y, tal que

$$\eta_j(Y) \geq \eta_j(X^*), \; j = 1,2,\ldots,L \qquad (4)$$

y que satisfaga a (4) como una estricta desigualdad para al menos un valor de j.

El concepto de eficiencia queda ilustrado en la figura 4 para el caso de dos respuestas.

La gráfica representa la región factible de un problema arbitrario de dos respuestas que se desean maximizar. La región mostrada corresponde a la región factible en el espacio de los criterios (las respuestas). Esta región corresponde a la transformación de la región factible en el espacio de los criterios. La definición de eficiencia nos permite asegurar que cualquier punto interior a la región no es Pareto óptimo ya que ambos criterios (η_1 y η_2) pueden ser mejorados por cualquier punto en el noreste de él. Consideremos el punto interior A de la figura 4 el cual corresponde en el espacio de los criterios a uno que no

es Pareto óptimo. La alternativa B tiene un mayor valor de η_1 que la alternativa A sin decrementar el valor de η_2. Similarmente, la alternativa C proporciona un valor mayor de η_2 que A aunque tiene el mismo valor de η_1. Mas aún, cualquier alternativa localizada en el noreste de A domina a A. El encontrar puntos Pareto óptimos se reduce a encontrar puntos factibles en el espacio de los criterios para los cuales la intersección de un ortante positivo (con vértice en el punto en cuestión) y la región factible es vacía. Uno de tales puntos es el D, al igual que C y todos los marcados en la figura localizados entre D y B.

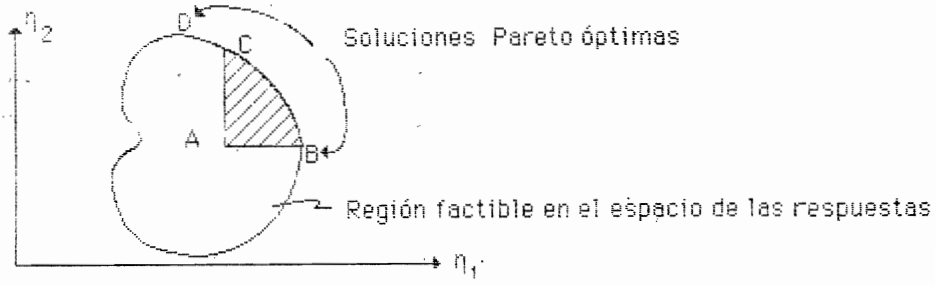


Fig. 4

Existe una teoría muy rica que permite expresar condiciones similares a las de Karush-Kuhn-Tucker para los problemas con una respuesta (Luc 1989) o Sawaragi (1887). Las estrategias que describiremos a continuación determinan soluciones Pareto óptimas.

Una característica importante del concepto arriba definido es que no es único; es decir existe por lo general un conjunto de soluciones Pareto óptimas. Esto en la gran parte de los casos es desventajoso pues el experimentador debe todavía indicar su preferencia por alguna de esas soluciones.

Criterios Globales

En este método, una solución óptima es un vector de variables de decisión que minimiza algun criterio global. La función que describe el criterio global representa una medición de "que tan cerca se puede acercar el experimentador a un vector ideal [η_1^o, η_2^o,..., η_L^o] préviamente definido". Varios autores han propuesto

diferentes tipos de funciones que describen el criterio global.

Derringer y Suich (1980) definen la siguiente función criterio

$$D = (d_1 \times d_2 \cdots d_L)^{1/k}$$

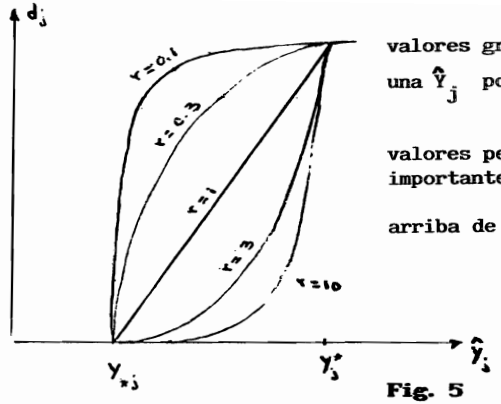
donde el valor d_j ($0 \leq d_j \leq 1$) indica la "deseabilidad" de la respuesta j, (j =1,2,..., L). El valor de d_j se incrementa a medida que la deseabilidad de la correspondiente respuesta se incrementa. La función D indica la deseabilidad global de las respuestas combinadas. Es evidente que D $\in$ [0,1], y a medida que se incrementa, se incrementa la deseabilidad de las demás respuestas. La función D es igual a cero si alguno de los valores d_j es igual a cero. Esto es deseable si alguna de las respuestas toma algun valor no permitido.

Se pueden tener restricciones en los valores máximos y mínimos de las respuestas. Aquí presentaremos el caso de restricciones en los valortes mínimos.

Suponga que estimamos la respuesta j mediante $\hat{Y}_j$, y que una cota inferior en su valor es Y_{*j}. Entonces el valor de d_j se define como sigue

$$d_j = \begin{cases} 0 & \text{si } \hat{Y}_j \leq Y_{*j} \\ \left[(\hat{Y}_j - Y_{*j})/(Y_j^* - Y_{*j})\right]^r & \text{si } Y_{*j} < \hat{Y}_j < Y_j^* \\ 1 & \text{si } \hat{Y}_j \geq Y_j^* \end{cases}$$

De esta manera $d_j = 0$ si $\hat{Y}_j \leq Y_{*j}$, lo que implica que el criterio global D sea cero, tal como se desea. El valor de Y_j^* es el valor más grande permitido de $\hat{Y}_j$, y puede ser igual al valor máximo individual, o un valor de $\hat{Y}_j$ tal que una cantidad mayor no tiene un mérito adicional. El valor de r es una cantidad especificada por el usuario. La figura 5 indica el comportamiento de d_j para varios valores de r. Note que a medida que r se incrementa, el valor de $\hat{Y}_j$ se hace más deseable al alejarse de Y_{*j}. En cambio, para valores pequeños de r, la deseabilidad de $\hat{Y}_j$ aumenta para valores cercanos a Y_{*j}.



valores grandes de r si es deseable una $\hat{Y}_j$ por arriba de Y_{*j}

valores pequeños de r si no es tan importante tener valores de $\hat{Y}_j$ por arriba de Y_{*j}

Fig. 5

La transformación original con la idea de deseabilidad es debida a Harrington (1965), y es de la forma $d_j = \exp(-\exp(-\hat{Y}_j))$ (un caso especial de una curva de Gompertz) para el caso de una función de un sólo lado y $d_j = \exp(-|-\hat{Y}_j|^r)$ para transformaciones de dos lados donde r es una constante definida previamente.

Otra transformación propuesta con el enfoque deseabilidad es la de Gatza y McMillan (1972)

$$d_j = \{\exp[-\exp(-\hat{Y}_j)]-\exp(-1)\}/[1-\exp(-1)],$$

que produce valores negativos de d_j para valores inadmisibles de $\hat{Y}_j$. Por ejemplo, si $\hat{Y}_j = 0$, entonces $d_i = 0$. Sin embargo, si $\hat{Y}_j$ es negativo, digamos, -1, entonces $d_i < 0$.

Zeleny (1974) propone (aunque en otro contexto de aplicaciones) la siguiente función

$$\text{Min } d_\alpha = \left[\sum_j | \hat{Y}_j(X) - Y_j^* |^\alpha \right]^{1/\alpha}$$

donde d_α es una métrica con $1 \leq \alpha \leq \infty$. La función d_α es un compromiso entre las los óptimos independientes y las funciones respuesta. Así d_α es un criterio de "que tan cerca" el experimentador se encuentra de la solución ideal.

Osyczka (1985) recomienda utilizar las desviaciones relativas en lugar de las absolutas debido a que tienen un significado directo en cualquier contexto.

Cuando el criterio global d_α se elige con $\alpha = \infty$, el criterio también es llamado minimax, dado que para esa métrica el óptimo $\mathbf{X}^*$ está definido como sigue

$$d^*_\alpha = \min_{x} \max_{j} \left| \left(\hat{Y}_j(X) - Y^*_j\right) / Y^*_j \right|$$

Métodos basados en preferencias

Existe una abundante cantidad de métodos basados en las preferencias del decisor para problemas con más de una respuesta. Todos esos métodos estan basados en la meta común de determinar la mejor solución a través de la representación matemática de las preferencias del decisor. Los métodos difieren en la manera en que son tomadas en cuenta las preferencias del decisor. Si las preferencias son articuladas apriori se tienen fundamentalmente los métodos basados en una función de utilidad , programación por metas y programación lexicográfica. En cambio si el decisor vá articulando sus preferencias a medida que vá obteniendo información de las respuestas se tienen fundamentalmente los métodos interactivos; ejemplos de los cuales se encuentran el método interactivo de Geoffrion, programación por metas interactiva, método de las metas satisfactorias (Benson 1975)) y Zionts-Wallenius (1976), entre otros.

Métodos basados en la articulación a priori de las preferencias

Supongamos que las preferencias del decisor se declaran antes de resolver el problema. Suponga también que el decisor proporcionó la información necesaria para construir el modelo (P_o). La información que puede proporcionar puede ser: a) cardinal, b) mixta (cardinal y ordinal). En el primer caso el decisor debe proporcionar alguna información acerca de sus preferencias de los distintos objetivos. En cambio si la información es mixta, el decisor debe ordenar sus respuestas en orden de su preferencia.

El método más conocido de la categoría a) es el método de la función de utilidad. En este caso el problema (P_o) es convertido a

Max $U(\eta_1(X), \eta_2(X),\ldots,\eta_L(X))$

sujeto a $a_i \le x_i \le b_i \quad i=1,2,\ldots,P$

donde $U(\cdot)$ es la función de utilidad de las L respuestas. Esta estartegia es similar a la que se maneja en la economía, y requiere que $U(\cdot)$ sea conocida. Varios autores han utilizado y revisado esta estrategia, por ejemplo Keeney y Raiffa (1976), Farquhar (1977), Huber (1974). Hwang y Masud (1979), Cohon (1978) proporcionan una amplia revisión de esos métodos.

De los métodos cuya preferencia es mixta tal vez el más conocido es el de programación por metas debido a Charnes y Cooper (1961), aunque ampliado de muchas maneras. Ignizio proporciona una de las revisiones más completas que existen de esta estrategia

La programación por metas requiere que el decisor haya articulado una *meta* para cada respuesta y un peso que refleja la importancia relativa de desviaciones de esa meta. Con esa información, el problema se reduce a determinar la solución que satisfaga lo máximo posible las preferencias del decisor. Así el problema (P_a) es

$$\text{Min } \Sigma\ w_j\ (d_j^+ + d_j^-) \qquad (5)$$

sujeto a $\quad a_i \le x_i \le b_i \qquad i = 1,2,\ldots,P \qquad (6)$

$$\hat{Y}_j(X) + d_j^- - d_j^+ = G_j \qquad j = 1,2,\ldots,L \qquad (7)$$

$$d_j^+,\ d_j^- \ge 0 \qquad j = 1,2,\ldots,P \qquad (8)$$

donde G_j representa la meta para la respuesta j, y d_j^+, d_j^- representan las desviaciones en exceso y defecto respectivamente de la meta G_j. La restricción (6) es la del problema original. La restricción (7) relaciona la respuesta j-ésima con las desviaciones d_j^- y d_j^+ al valor propuesto G_j. Note que si d_j^+ es positiva, entonces se logró un valor de la respuesta $\hat{Y}_j$ por arriba de la meta propuesta G_j; mientras que si d_j^-, entonces se obtuvo un valor de la respuesta $\hat{Y}_j$ por abajo de lo propuesto. Las restricciones (8) aseguran que las desviaciones no sean negativas. Los pesos w_j en la función (5) afectan a las desviaciones de G_j. Son especificados préviamente e indican la importancia relativa de las diferentes respuestas. Existen muchas otras variantes de este modelo básico, algunas de las cuales son revisadas por Ignizio (1976) o Steuer (1986) donde además se

presentan algoritmos especiales para resolver eficientemente dichas variantes.

Biles (1977b) utiliza la formulación (P_3) anterior aplicada a simulación con objetivos múltiples y la resuelve aplicando técnicas de programación lineal.

(Métodos para la articulación progresiva de las preferencias (métodos interactivos).

Estos métodos se basan en la definición progresiva de las preferencias del analista a madida que se explora el espacio definido por los criterios. Gran atención ha recibido esta estrategia durante lod últimos años, y la tendencia es efectuar un diálogo iterativo entre el decisor y una computadora. En cada uno de esos diálogos el analista es interrogado sobre sus preferencias de la solución actual (o conjunto de soluciones) para poder así determinar una nueva solución.

Pocos trabajos han sido hechos en el área de múltiples superficies de respuesta. Uno de los mas conocidos es el Montgomery y Bettencourt (1977), donde se aplica el método de Geoffrion (Geoffrion, Dyer y Feinberg 1972) para simular el comportamiento de una batalla entre tanques. El problema puede ser formulado como sigue:

$$\text{Max } \mathbf{U}(n_1(\mathbf{X}), n_2(\mathbf{X}),\ldots,n_L(\mathbf{X}))$$
$$\text{sujeto a } \mathbf{X} \in S$$

donde **U** es la función de utilidad del tomador de desiciones, S es el espacio factible donde se ubican las variables.

Se supone que la función **U** es conocida implícitamente. (Si fuera explícitamente conocida, sería un problema de la estrategia donde se conoce la información a priori. Geoffrion utiliza el algoritmo de Frank-Wolfe para la solución del problema anterior debido a su simplicidad y robusta convergencia. La idea es efectuar aproximaciones lineales. La parte crucial del método es la determinación de la dirección de mejoramiento. Dado que la función **U** no es onocida , se solicita al decisor información en la forma de preferencia entre cualesquiera dos objetivos. Con esta información el gradiente es estimado y por consiguiente la dirección de mejoramiento. Para una presentación más amplia del método sugerimos el libro de Hwang y Masud (1979).

Referencias

Benson, R.G., "Interactive Multiple Criteria Optimization Using Satisfactory Goals", Ph.D. Thesis, University of Iowa, 1975.

Biles, W.E., "A Response Surface Method for Experimental Optimization of Multiple response Processes", Ind. Eng. Chem. Processes Des. Dev., Vol. 14, No. 2, 1975.

Biles, W.E. (a), "Strategies for Optimization of Multiple Response Simulation Models", Proceedings of the 1977 Winter Simulation Conference, Pp. 134-142.

Biles, W.E. (b) "Optimization of Multiple-Objective Computer Simulations: A Non-Linear Goal Programming Approach", Proceedings of the 1977 AIDS Conference, Chicago, 1977.

Biles, W.E. y Swain, J.J., "Optimization and Industrial Experimentation", Wiley-Interscience, 1980.

Carroll, C.W., "The Created Response Surface Technique for Optimizing Nonlinear Retrained Systems", Operations Research, Vol. 9, pp. 169-184, 1961.

Charnes, A., y Cooper, W., "Management Models and Industrial Aplications of Lineal Programming", Vol. 1, Wiley, 1961.

Cohon, J.L., "Multiobjective Programming and Planning", Academic Press, 1978.

Davies, O. L., "The Design and Analysis of Industrial Experiments", Hafner Publishing Company, New York, 1956.

Derringer, G. y R. Suich, "Simultaneous Optimization of Several Response Variables", Journal of Quality Technology, Vol. 12, No 4, October 1980.

Farell, W., "Literature Review And Bibliography of Simulation Optimization", Proceedings of the 1977 Winter Simulation Conference, pp. 117-124.

Farquhar, P.H., "A Survey Multi-Attribute Utility Theory and Applications". en Starr, M.K. y Zeleny, M., (Eds.), Multiple Criteria Decision Making, North Holland, 1977.

Fiacco, A.V., y G.P. McCormick, "Nonlinear Programming, Sequential Unconstrained Minimization Techniques", Wiley, 1968.

Fields, T.G., "Nonlinear Programming Techniques for the Multiple Response Problem", MS Thesis, Georgia Tec., February 1974.

Gatza, P.E. y McMillan, R.C., "The Use of Experimental Design and

Computerized Data Analisis in Elastomer Development Studies", Division of Rubber Chemistry, American Chemical Society Fall Meeting, Paper No. 6, Cincinnati, Ohio, October 3-6, 1972.

Geoffrion, A.M., J.S. Dyer y Feinberg, A., "An Interactive Approach for Multi-Criterion Optimization, with an Application to the Operation of an Academic Department", Management Science, Vol. 19, No 4 (Part 1), pp. 357-368, 1972.

Glynn, P.W., "Optimization of Stochastic Systems", Proceedings of the 1986 Winter Simulation Conference, Washington, D.C.

Harrington, E.C. Jr., The Desirability Function", Industrial Quality Control, Vol. 21, No. 10, pp. 494-498, 1965.

Hartmann, N. E. Beaumont y R. A., "Optimum Compounding by Computer",Journal of the Intitute of the Rubber Industry, Vol. 2 No. 6, 1968, pp. 272-275.

Hendrix, C.D., "Experimental Design-An Efficient Route to Process and Product Development", Technical Report, Union Carbide Corp., S. Charleston, WV, 1970.

Hickman, J. B., "Fitting Yield and Cost Response Surfaces", Chemical Engineering Progress Vol. 56 1960.

Hooke, R. y Jeeves, T.A., "A Direct Search Solution of Numerical and Statistical Problems", JACM, Vol. 8, No. 2. pp 212-229, 1961.

Huber, G.P., "Multi-Attribute Utility Models: A Review of Fields and Fields Like Studies", Management Science, Vol. 20, No. 10, pp. 1393-1402, 1974.

Hunter, J. S., "Determination of Optimum Operating Conditions by Experimental Methods: Part II-1,2,3, Models and Methods", Industrial Quality Control Vol. 15, pp 6-14, 1958-1959.

Hwang, C.L. y A.S.M., Masud, "Multiple Objective Decision Making-Methods and Applications: A State of the Art Survey", Lecture Notes in Economics and Mathematical System, Vol. 164, Springer-Verlag, 1979.

Ignizio, J.P., "Goal Programming and Extensions", Lexington Books, 1976.

Jacobson, S.H. y L.W. Schruben, "A Review of Techniques for Simulation Optimization", Technical Report No. 715, Cornell University, Ithaca, N.Y. September 1987.

Keeney, R. y Raiffa, H., "Decision with Multiple Objectives: Preference and Value Tradeoffs", Wiley, 1976.

Khuri, A.I. y M. Conlon, "Simultaneous Optimization of Multiple Responses Represented by Polynomial Regression Functions", Technometrics, 30, pp 95-104, 1981.

Lind, E. E., Goldin, J. Hickman, J. B., "Fitting Yield and Cost Response Surfaces", Chemical Engineering Progress vol. 56 1960.

Lind, E. E. y Young, W. R. "Con-man: a 3-D Device for the Representation of Responses Surfaces", Transactions Nineteenth Annual Convention American Society for Quality Control 1965 p. 545.

Meketon, M., "Optimization Methods in Simulation: A Survey of Recent Results", Presented at the 1983 Winter Simulation Conference.

Montgomery, D.C. y Bettencourt, V.M, Jr., "A Review of Multiple Response Surface Methods in Computer Simulation", ORSA\TIMS Meeting, 1976.

Montgomery, D.C. y Bettencourt, V.M, Jr., "Multiple Response Surface Methods in Computer Simulation", Simulation, Vol. 29, No. 4, p. 113-121, 1977.

Montgomery, D.C., Talavage, J.J., y Mullen ,C.J., "Response Surface Approach to Improving Traffic Signal Settings in a Street Network", Transportation Research, Vol. 6, No. 1, pp. 69-80, 1972.

Myers, R.H. y Carter, W.H., "The Use of Lagrange Multipliers with Response Surfaces", Technometrics Vol. 1, pp. 289-292, 1959.

Nicholson, T. A. y Pullen, R. D., "Statistical Optimization Techniques in the Design of Rubber Compounds", Computer Aided Design Vol 1 No. 1, 1969 pp. 39-47.

Osyczka, A., "Multicriteria Optimization for Engeneering Design", Cap. 7 de "Design Optimization", Ed. Gero J.S., Notes and Reports in Mathematics in Science and Engineering, 1985.

Rosen, J.B., "The Gradient Projection Method of Nonlinear Programming, II, Non-Linear Constraints", Journal of the Society of Industrial and Applied Mathematics, Vol. 9, pp. 514-532, 1961.

Rustagi, J.S., "Optimizizing Methods in Simulation", Technical report, Department of Statistics, Ohio State University, july 1981.

Sawaragi. "Theory of Multiobjectives", Academic Press, 1987.

Smith, D.E., "An Empirical Investigation of Optimum Seeking in the Computer Simulation", Operations Research, Vol. 21, pp. 475-479.

1973.

Smith, H. y Rose, A., "Subjective Responses in Process Investigation", Industrial and Engineering Chemistry Vol. 55, p. 25, 1963.

Steuer, R.E., "Multiple Criteria Optimization: Theory, Computations, and Applications", John Wiley & Sons, 1986.

Taraman, K.S., "Multi Matching Output-Multiple Independent Variable Turning Research by Response Surface Metodology", International Journal of Production Research, Vol. 12, No. 2, pp. 233-245, 1974.

Taraman, K. S. Lambert, R. K. "Application of Response Surface Methodology to the Selection of Machining Variables", AIIE Transactions, Vol. 4 No. 2, pp. 111-115, June 1972.

Umland, A.W. y Smith, W.N., "The use of Lagrange Multipliers with Response Surfaces", Technometrics Vol. 1, pp. 289-292, 1959.

Zeleny, M., "A Concept of Compromise Solutions and the Method of the Displaced Ideal", Computers and Operations Research, Vol. 1, p 479, 1974.

Zionts, S. y Wallenius, J., "An Interactive Programming Method for Solving the Multiple Criteria Problem", Manegement Science, Vol. 22, No. 6, pp. 652-663, 1976.

Pronósticos ARIMA con Restricciones Derivadas de un Cambio Estructural

Víctor M. Guerrero
Instituto Tecnológico Autónomo de México
01000 México, D. F.

RESUMEN

En este trabajo se presenta un modelo de series de tiempo que tiene en cuenta un cambio estructural, ya sea en la parte determinista o en la parte estocástica de un modelo ARIMA. El cambio estructural se supone que ocurre durante el horizonte de pronóstico de la serie y la única información disponible acerca de ese cambio, además del momento de su ocurrencia, es proporcionada por tan solo una o dos restricciones impuestas sobre los valores futuros de la serie. Se derivan aquí las fórmulas para el cálculo de los pronósticos y de sus varianzas; se presentan también los procedimientos que surgen de la aplicación de dichas fórmulas, los cuales se ilustran mediante algunos ejemplos teóricos.

1. INTRODUCCION

En ocasiones ocurre que a un analista de series de tiempo se le proporciona información adicional a la del registro histórico de la serie. Dicha información debe ser incluída para mejorar la calidad de los pronósticos, tanto en lo que se refiere a la exactitud como a la precisión. Por ejemplo, puede darse el caso de que se conozca una meta al final del año, la cual debe ser satisfecha por una serie cuyos datos son mensuales. Aquí se considerará el caso en el cual las cifras mensuales a pronosticar se obtienen con la ayuda de un modelo ARIMA (autorregresivo, integrado y de promedios móviles); evidentemente, si se sabe que un cambio estructural ocurrirá en el modelo durante el horizonte de pronóstico, la probabilidad

de alcanzar la meta anual con los pronósticos ARIMA convencionales es cero, a menos que la información adicional provista por la meta al final del año, sea considerada como una restricción que deben satisfacer los pronósticos.

En este artículo se propone un modelo en el cual se supone que el cambio estructural afectará específicamente, ya sea a la parte determinista o a la parte estocástica del modelo ARIMA. También se supone que se cuenta con suficientes datos históricos como para construír dicho modelo mediante la estrategia de Box y Jenkins (véase a este respecto Guerrero, 1983). Por otro lado, se hace la consideración de que la información adicional consta tan solo de una o dos restricciones lineales que deben satisfacer los pronósticos de la serie, esto se apega a lo que realmente ocurre en la práctica.

Conviene resaltar que quizás en algunas aplicaciones prácticas, no se tenga la certeza de que la información adicional implica un cambio estructural en el comportamiento histórico de la serie. En tal caso es aconsejable realizar una prueba de compatibilidad entre la información histórica y la adicional, mediante algún estadístico de prueba como puede ser el proporcionado por Guerrero (1989). Si en realidad no se anticipa un cambio estructural en el horizonte de pronóstico, la información adicional puede ser incorporada en los pronósticos sin alterar la estructura del modelo ARIMA, según se indica también en el citado trabajo de Guerrero.

2. MODELO PARA REPRESENTAR CAMBIOS EN LA ESTRUCTURA ARIMA

Sea $\{Z_t\}$ una serie de tiempo observada durante el período $t = 1, \ldots, N$, la cual admite una representación de proceso ARIMA. Su pronóstico con Error Cuadrático Medio Mínimo (ECMM), dado el vector columna de datos históricos $\mathbf{Z}_0 = (Z_1, \ldots, Z_N)'$ y expresado

en términos de los coeficientes de la representación de promedios móviles pura, produce para $t = N+1, N+2, \ldots$, el error de pronóstico

$$Z_t - E(Z_t | Z_0) = \sum_{j=0}^{t-1} \psi_j a_{t-j} \quad , \quad \psi_0 = 1 \qquad (2.1)$$

en donde $\{a_t\}$ representa un proceso de Ruido Blanco Gaussiano con media cero y varianza σ_a^2, y donde las ponderaciones $\psi_1, \psi_2, \ldots$ se suponen conocidas. La expresión (2.1) se puede validar en el caso estacionario mediante el Teorema de la Descomposición de Wold y en el caso no-estacionario por los resultados de Bell (1984, sección 2).

Un cambio estructural que afecte la estructura determinista de un proceso ARIMA, es decir el nivel local del proceso, se podría atribuir a una intervención, de acuerdo con las ideas de Box y Tiao (1975). Si dicha intervención es justificable y se conoce el momento de su ocurrencia, entonces el modelo ARIMA original puede extenderse para incluir la función dinámica D_t^τ, que se supondrá es de la forma

$$(1 - \delta B) D_t^\tau = \omega S_t^\tau \qquad (2.2)$$

en donde B denota al operador de retraso tal que $BZ_t = Z_{t-1}$, mientras que δ y ω son los parámetros de la intervención. Esta ecuación incluye como casos especiales a: un cambio inmediato de nivel de tamaño ω (si $\delta = 0$), un cambio gradual de nivel con ganancia eventual $\omega/(1-\delta)$ (si $|\delta| < 1$) y un cambio no acotado (si $|\delta| \geq 1$). La aplicabilidad de esta función en trabajos empíricos ha sido demostrada ya por varios autores (e.g. Box y Tiao, 1975 o Guerrero, 1986). La expresión (2.2) incluye a la función de "escalón" S_t^τ que toma el valor 1 cuando $t \geq \tau$ y es cero en otro caso, en donde el valor τ denota el momento de la intervención.

Por otro lado, es bien conocido (véase Granger y Morris, 1976) que al sumar un proceso de Ruido Blanco a un proceso ARIMA, se produce otro proceso ARIMA con diferente estructura

estocástica. Entonces, si se piensa que es la estructura estocástica del modelo ARIMA original la que se verá afectada por el cambio estructural, se considerará aquí la adición de un proceso de Ruido Blanco $\{v_t\}$ a dicho modelo para tener en cuenta el cambio. Así, dado Z_0, $\{v_t\}$ se supondrá que es un proceso de Ruido Blanco Gaussiano, independiente de $\{a_t\}$, con media cero y varianza σ_v^2.

Un modelo general que incluye tanto efectos deterministas, denotados por D, como estocásticos (V), está dado por

$$Z_{t,D,V} = Z_t + \left(\frac{\omega}{1-\delta B} + v_t \right) S_t^{\tau} \quad , \quad t = 1, \ldots, N+H \tag{2.3}$$

con H el horizonte de pronóstico. Sin pérdida de generalidad, supóngase que la intervención ocurre al momento $\tau = N+1$ (de otra manera se podrían incluir en Z_0 los pronósticos ARIMA convencionales $\hat{Z}_N(1), \ldots, \hat{Z}_N(\tau - N - 1)$, en notación de Box-Jenkins, y el origen de los pronósticos que surjan del modelo (2.3) sería $\tau - 1$). Entonces, si se define el vector de valores futuros de la serie como $Z_F = (Z_{N+1}, \ldots, Z_{N+H})'$ y se dan definiciones similares para D_F, V_F y $Z_{F,D,V}$, se pueden expresar los valores obtenidos con (2.3) para $t = N+1, \ldots, N+H$, como

$$Z_{F,D,V} = Z_F + D_F + V_F \tag{2.4}$$

Así pues, para obtener el pronóstico con ECMM dada la formulación (2.3), simplemente se aplicaría el operador esperanza condicional en (2.4), dado Z_O. Esto produce el pronóstico

$$E(Z_{F,D,V} \mid Z_O) = E(Z_F \mid Z_O) + D_F \tag{2.5}$$

cuyo error tiene matriz de covarianza

$$Cov[Z_{F,D,V} - E(Z_{F,D,V} \mid Z_O) \mid Z_O] = \sigma_a^2 \psi \psi' + \sigma_v^2 I \tag{2.6}$$

con ψ la matriz triangular inferior de dimensión $H \times H$, que tiene los valores $1, \psi_1, \ldots, \psi_{H-1}$ en la primera columna, los valores $0, 1, \psi_1, \ldots, \psi_{H-2}$ en la segunda columna, y así sucesivamente.

Es claro que si ω, δ y σ_v^2 fuesen conocidos, las expresiones (2.5)-(2.6) dan la solución al problema de obtener pronósticos que consideran un cambio estructural durante el horizonte de pronóstico. No obstante, en la práctica esos parámetros son desconocidos y se debe contar con información adicional a Z_O para poder determinar sus valores. El interés de este trabajo se centra entonces en obtener el vector de pronósticos, cuando la información adicional acerca de los valores futuros de la serie está dada en forma de restricciones lineales, o sea

$$Y = CZ_{F,D,V} \tag{2.7}$$

donde Y representa a un vector de dimensión m con valores conocidos y C es una matriz de constantes de dimensión $m \times H$.

Así, nótese que (2.1), (2.4) y (2.5) implican

$$Z_{F,D,V} = E(Z_{F,D,V} \mid Z_O) + \psi a_F + V_F \tag{2.8}$$

con $a_F = (a_{N+1}, \ldots, a_{N+H})'$, por lo tanto, como $Cov(\psi a_F + V_F \mid Z_O) = \sigma_a^2 \psi\psi' + \sigma_v^2 I$, se plantea un problema de minimización Lagrangiana de $Z_{F,D,V} - E(Z_{F,D,V} \mid Z_O)$ sujeto a la restricción (2.7). Para resolver este problema, considérese la siguiente función, que involucra tanto a Z_O como a Y

$$Q = [Z_{F,D,V} - E(Z_{F,D,V} \mid Z_O)]'(\sigma_a^2 \psi\psi' + \sigma_v^2 I)^{-1}[Z_{F,D,V} - E(Z_{F,D,V} \mid Z_O)]$$
$$+ 2L'(Y - CZ_{F,D,V}) \tag{2.9}$$

en la cual L es un vector de multiplicadores de Lagrange. Al resolver la ecuación $O = \partial Q / \partial Z_{F,D,V}$ se obtiene

$$\hat{Z}_{F,D,V} = E(Z_{F,D,V} \mid Z_O, Y)$$
$$= E(Z_{F,D,V} \mid Z_O) + A[Y - CE(Z_{F,D,V} \mid Z_O)] \tag{2.10}$$

con

$$A = (\sigma_a^2 \psi\psi' C' + \sigma_v^2 C')(\sigma_a^2 C\psi\psi' C' + \sigma_v^2 CC')^{-1} \tag{2.11}$$

Además, la matriz de covarianza de los errores del pronóstico restringido se puede verificar que es

$$Cov[(Z_{F,D,V} - \hat{Z}_{F,D,V}) \mid Z_O, Y] = (I - AC)(\sigma_a^2 \psi\psi' + \sigma_v^2 I) \tag{2.12}$$

Por consiguiente, el pronóstico $\hat{Z}_{F,D,V}$ dado por (2.10)-(2.11), satisface las restricciones impuestas por (2.7), pero aún se requiere obtener los valores de ω, δ y σ_v^2. Ya que (2.7) se supone que impone a lo más dos restricciones lineales, no es posible determinar a partir de ellas los valores de tres parámetros, así que a continuación se particularizan los resultados previos a cada uno de los casos considerados, por separado. De esta forma en (2.4) se tendrá $Z_{F,D} = Z_F + D_F$ cuando el cambio sea determinista, mientras que será $Z_{F,V} = Z_F + V_F$ cuando el cambio sea estocástico. Una manera de elegir entre estos dos modelos es con base en el conocimiento teórico que se tenga del fenómeno en estudio.

2.1 Cambio en la Estructura Determinista

Si es de esperar que ocurra un cambio en la estructura determinista del modelo, se hace $\sigma_v^2 = 0$ en la formulación previa y se obtiene así el pronóstico restringido correspondiente, que será denotado por $\hat{Z}_{F,D}$. Para determinar ω y δ se hará uso de la forma explícita de D_t^T (véase (2.2)) que se sabe es 0 si $t \leq N$ y es $\omega(1-\delta^{t-N})/(1-\delta)$ si $t > N$. Entonces, ya que toda la información sobre el cambio estructural se encuentra en los valores futuros de la serie, se pueden encontrar los valores de ω y δ al resolver el sistema de ecuaciones

$$C\hat{D}_F = Y - CE(Z_F \mid Z_O) \tag{2.13}$$

que se supondrá es consistente, o sea que cualquier relación lineal que exista entre los renglones de C también existe entre los correspondientes elementos de $Y - CE(Z_F \mid Z_O)$. La solución de dicho sistema es

$$\hat{D}_F = C^-[Y - CE(Z_F \mid Z_O)] + (I - C^-C)w \tag{2.14}$$

con C^- una inversa generalizada de C y w un vector arbitrario de constantes cuya dimensión es H. Entonces se obtiene

$$\hat{Z}_{F,D} = E(Z_F \mid Z_O) + \hat{D}_F \tag{2.15}$$

con

$$Cov[(\mathbf{Z}_{F,D} - \hat{\mathbf{Z}}_{F,D}) \mid \mathbf{Z}_O, \mathbf{Y}] = \sigma_a^2 (I - A_D C)\psi\psi' \qquad (2.16)$$

y

$$A_D = \psi\psi' C'(C\psi\psi' C')^{-1} \qquad (2.17)$$

Puesto que la matriz de covarianza de los pronósticos irrestrictos, que está dada por $\sigma_a^2 \psi\psi'$, excede a la matriz en (2.16) por la matriz $\sigma_a^2 A_D C\psi\psi'$, que es semidefinida positiva, se sigue que el pronóstico restringido $\hat{\mathbf{Z}}_{F,D}$ es al menos tan preciso como el irrestricto. De hecho, la matriz de covarianza (2.16) es idéntica a la que aparece en Guerrero (1989) asociada con el error de pronóstico $\mathbf{Z}_F - E(\mathbf{Z}_F \mid \mathbf{Z}_O, \mathbf{Y})$ y que surge del supuesto de que no hay cambio estructural.

2.2 Cambio en la Estructura Estocástica

Cuando sólo la estructura estocástica se espera que cambie debido a la intervención, se hace $\omega = O$ en el modelo general y se usa el hecho de que, dado $\mathbf{Z}_O$

$$\mathbf{Y} - CE(\mathbf{Z}_F \mid \mathbf{Z}_O) = C\psi \mathbf{a}_F + C\mathbf{V}_F \sim N(\mathbf{O}, \sigma_a^2 C\psi\psi' C' + \sigma_v^2 CC') \qquad (2.18)$$

Entonces se propone el uso del estadístico

$$K_V = [\mathbf{Y} - CE(\mathbf{Z}_F \mid \mathbf{Z}_O)]'(\sigma_a^2 C\psi\psi' C' + \sigma_v^2 CC')^{-1}[\mathbf{Y} - CE(\mathbf{Z}_F \mid \mathbf{Z}_O)] \qquad (2.19)$$

el cual, para N suficientemente grande, se distribuye aproximadamente como una variable Ji-cuadrada con m grados de libertad. Este estadístico es útil para verificar la validez de la restricción $\mathbf{Y} = C\mathbf{Z}_{F,V}$ que es equivalente a $\mathbf{Y} = CE(\mathbf{Z}_F \mid \mathbf{Z}_O) + C\psi \mathbf{a}_F + C\mathbf{V}_F$. la cual será válida cuando σ_v^2 sea elegida razonablemente. De esta forma, aquí se considera ahora la restricción como una hipótesis nula, la cual no será rechazada si σ_v^2 se elige adecuadamente. Además, dicho criterio de adecuación será satisfecho de acuerdo con un nivel de significación predeterminado. Esto es, no se rechaza la hipótesis cuando $K_V < \chi_m^2(\alpha)$, donde $\chi_m^2(\alpha)$ denota al punto porcentual superior α de la distribución Ji-cuadrada correspondiente. De aquí que se sugiera seleccionar el valor $\hat{\sigma}_v^2$ por ensayo y error hasta

que K_V alcance un valor menor que el punto porcentual especificado. Evidentemente este procedimiento no produce un valor único de $\hat{\sigma}_v^2$, pero ya que se sabe de (2.12) que mientras más grande sea este valor, menor será la precisión del pronóstico restringido, se deberá seleccionar dicho valor como el mínimo (o cercano a dicho mínimo) para el cual se cumple que $K_V < \chi_m^2(\alpha)$, con el nivel α dado de antemano.

Por lo tanto se obtiene

$$\hat{Z}_{F,V} = E(Z_F \mid Z_O) + A_V[Y - CE(Z_F \mid Z_O)] \qquad (2.20)$$

con

$$A_V = (\sigma_a^2 \psi\psi' C' + \hat{\sigma}_v^2 C')(\sigma_a^2 C\psi\psi' C' + \hat{\sigma}_v^2 CC')^{-1} \qquad (2.21)$$

y

$$Cov[(Z_{F,V} - \hat{Z}_{F,V}) \mid Z_O, Y] = (I - A_v C)(\sigma_a^2 \psi\psi' + \hat{\sigma}_v^2 I) \qquad (2.22)$$

Así, en este caso la precisión del pronóstico restringido puede ser mayor o menor que la del pronóstico irrestricto, dependiendo de si $A_V C(\sigma_a^2 \psi\psi' + \hat{\sigma}_v^2 I) - \hat{\sigma}_v^2 I$ es positiva o negativa semidefinida.

3. EJEMPLOS

Para apreciar cómo es que funcionan las soluciones propuestas, considérese un ejemplo simple, en el cual el modelo ARIMA original es un AR(1). Supóngase que el cambio estructural es determinista y que existen dos restricciones a ser satisfechas por los pronósticos, dichas restricciones se especifican mediante

$$C = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{y} \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

En este caso se tiene $E(Z_F \mid Z_O) = Z_N(\phi, \phi^2, \phi^3, \phi^4)'$, por lo cual $Y - CE(Z_F \mid Z_O) = (Y_1 - \phi^2 Z_N, Y_2 - \phi^4 Z_N)'$. Al usar la inversa generalizada $C^- = C'$ se deducen los siguientes valores para los parámetros de la intervención

$$\hat{\omega} = (Y_1 - \phi^2 Z_N)/(1 + \hat{\delta}) \quad , \quad \hat{\delta} = \pm[(Y_2 - \phi^4 Z_N)/(Y_1 - \phi^2 Z_N) - 1]^{1/2}$$

de donde se sigue que

$$\hat{Z}_{F,D} = (\phi Z_N + (Y_1 - \phi^2 Z_N)/(1+\hat{\delta}), Y_1, \phi^3 Z_N + [Y_2 - \phi^4 Z_N - \hat{\delta}(Y_1 - \phi^2 Z_N)]/(1+\hat{\delta}), Y_2)'$$

Además, como el modelo se puede escribir $Z_t = \phi Z_{t-1} + a_t$ (con Z_t medida en desviaciones respecto a la media) se tiene que $\psi_j = \phi^j$ para $j = 1, 2, \ldots$, así que

$$A_D = \begin{pmatrix} \phi(1+\phi^2)^{-1} & 0 \\ 1 & 0 \\ \phi(1+\phi^2)^{-3} & \phi(1+\phi^2)^{-1} \\ 0 & 1 \end{pmatrix}$$

y la matriz de covarianza de los errores de pronóstico, cuya parte triangular inferior se omite debido a la simetría, viene a ser

$$Cov[(Z_{F,D} - \hat{Z}_{F,D}) | Z_O, Y] = \sigma_a^2 \begin{pmatrix} 1-\phi^2(1+\phi^2)^{-1} & 0 & \phi^2(1-\phi^2) & 0 \\ \cdots & 0 & 0 & 0 \\ \cdots & \cdots & (1+\phi^2+\phi^4)[1-\phi^2(1+\phi^2)^{-1}]-\phi^2 & 0 \\ \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

Supóngase ahora que sólo existe una restricción, en la cual $C = (0,0,0,1)$ y $Y = Y$, de manera que $Y - CE(Z_F | Z_O) = Y - \phi^4 Z_N$. Entonces, si el cambio estructural es puramente estocástico y el proceso de Ruido Blanco contaminante tiene varianza σ_v^2, un valor razonable para este parámetro puede obtenerse mediante el uso del estadístico K_V; esto es, se debe seleccionar $\hat{\sigma}_v^2$ de tal forma que

$$(Y - \phi^4 Z_N)/\chi_1^2(\alpha) - (1+\phi^2+\phi^4+\phi^6)\sigma_a^2 < \hat{\sigma}_v^2$$

para un nivel α dado. Entonces, si se hace $\lambda = 1+\phi^2+\phi^4+\phi^6+\hat{\sigma}_v^2/\sigma_a^2$, se obtiene

$$A_V = \lambda^{-1}(\phi^3, \phi^2(1+\phi^2), \phi(1+\phi^2+\phi^4), \lambda)'$$

y el pronóstico restringido resulta ser

$$\hat{Z}_{F,V} = \lambda^{-1} \begin{pmatrix} (1+\phi^2+\phi^4+\hat{\sigma}_v^2/\sigma_a^2)\phi Z_N + \phi^3 Y \\ (1+\phi^2+\hat{\sigma}_v^2/\sigma_a^2)\phi^2 Z_N + \phi^2(1+\phi^2) Y \\ (1+\hat{\sigma}_v^2/\sigma_a^2)\phi^3 Z_N + \phi(1+\phi^2+\phi^4) Y \\ \lambda Y \end{pmatrix}$$

En estos dos ejemplos pueden observarse claramente dos puntos: primero, que las restricciones sobre los valores futuros de la serie son satisfechas exactamente por los respectivos pronósticos restringidos y segundo, que la varianza del valor restringido es cero, al igual que las covarianzas de los restantes pronósticos con este valor.

4. CONCLUSIONES

En este documento se presentan dos modelos básicos que sirven para tener en cuenta un cambio estructural que se espera ocurra dentro del horizonte de pronóstico de la serie en estudio. La información acerca de este cambio estructural se supone que es proporcionada exclusivamente por algunas restricciones lineales impuestas sobre los valores futuros de la serie. Estos modelos son la base a partir de la cual se derivan los procedimientos A_D y A_V que sirven para obtener los pronósticos restringidos.

Los ejemplos presentados llevan como finalidad el mostrar la utilidad potencial y la aplicabilidad de los métodos sugeridos. Aun cuando en aplicaciones prácticas, debe uno apoyarse en consideraciones de carácter teórico relacionadas con el fenómeno en estudio para discriminar entre los procedimientos A_D o A_V, una extensión de este trabajo debería considerar la aplicación del modelo general, en el cual se permitan cambios estructurales tanto deterministas como estocásticos en el modelo ARIMA. Desde luego, esto requeriría de mayor información que sea provista por las restricciones lineales que la que aquí se supuso.

REFERENCIAS

Bell, W. (1984) "Signal Extraction for Nonstationary Time Series". The Annals of Statistics 12, 646-664.

Box, G.E.P. y Tiao, G.C. (1975) "Intervention analysis with applications to economic and environmental problems". Journal of the American Statistical Association 70, 70-79.

Granger, C.W.J. y Morris, M.J. (1976) "Time Series Modelling and Interpretation". Journal of the Royal Statistical Society A-139, 246-257.

Guerrero, V.M. (1983) Análisis Estadístico de Series de Tiempo Económicas; Libro no-publicado, disponible en fotocopia. México.

Guerrero, V.M. (1986) "Un Modelo Estadístico útil para Pronosticar y Evaluar la Inflación durante el año de 1983". Investigación y Desarrollo Aplicados I-1, (Revista del Centro Científico de IBM-México) 73-85.

Guerrero, V.M. (1989) "Optimal Conditional ARIMA Forecasts". Journal of Forecasting 8, 215-229.

LA FUNCION DE AUTOCORRELACION EXTENDIDA Y SU EMPLEO EN LA CONSTRUCCION DE MODELOS PARA SERIES DE TIEMPO

Alejandro Islas Camargo, Depto. Matemáticas - UAM-Iztapalapa

RESUMEN

En este trabajo se presenta un método útil para la identificación del orden de un modelo de series de tiempo mixto ARMA(p,q), ya sea estacionario o no estacionario, propuesto por Tsay y Tiao en 1984. Se desarrollan básicamente las siguientes dos ideas de estos autores:

i) Proponer estimadores de mínimos cuadrados que sean consistentes para los parámetros autorregresivos.

ii) Presentar un procedimiento para la especificación del orden de un modelo dentro de la clase de modelos ARMA(p,q) estacionarios y no estacionarios.

En el desarrollo de estas ideas, se define una extensión de la Función de Autocorrelación, la cual se basa en estimadores consistentes de los parámetros autorregresivos.

INTRODUCCION

Los modelos autorregresivos y de promedios móviles de orden (p,q), usualmente abreviados modelos ARMA(p,q), son con frecuencia utilizados para describir y predecir observaciones en series de tiempo. Estos modelos comúnmente se representan mediante

$$\Phi_p(B)Z_t = C + \Theta_q(B)a_t \qquad (1)$$

en donde

$$\Phi_p(B) = U_d(B)\phi_{p-d}(B) = 1 - \sum_{i=1}^{p} \Phi_i B^i \ , \quad U_d(B) = 1 - \sum_{i=1}^{d} U_i B^i$$

$$\phi_{p-d}(B) = 1 - \sum_{i=1}^{p-d} \phi_i B^i \qquad y \qquad \theta_q(B) = 1 - \sum_{i=1}^{q} \theta_i B^i$$

son polinomios de retraso de orden p, d, p-d y q respectivamente B es el

operador de retraso tal que $B^k Z_t = Z_{t-k}$ para k = 0,1, ... y toda t, C al igual que las Φ's, θ's y U's son constantes y $\{a_t : t \in \mathbb{Z}\}$ es una sucesión de v.a.i.i.d. con distribución $N(0,\sigma_a^2)$. En la práctica, comúnmente se requiere que todas las raíces de $U_d(B)$ estén sobre el círculo unitario, además de que $\Phi_p(B)$ y $\theta_q(B)$ no tengan factores comunes. Las Z_t's representan la d-ésima diferencia o cualquier otra transformación apropiada para volver estacionaria alguna serie de tiempo originalmente no estacionaria; cuando $U_d(B) = (1-B)^d$ el modelo (1) se conoce como un modelo **ARIMA(p-d,d,q).**

El problema de estimar los parámetros autorregresivos de un modelo ARMA(p,q) mediante un ajuste de mínimos cuadrados ordinarios (MCO) ha sido tratado muy ampliamente. Despues de que Mann y Wald (1943) probaron que se obtenian estimadores consistentes en el caso autorregresivo (q=d=0) sólo avances marginale se habían logrado hasta 1983 con el trabajo de Tiao y Tsay.

Para procesos mixtos, aún estacionarios , los estimadores de MCO de los parámetros autorregresivos son sesgados e inconsistentes. La obtención de estimadores de MCO consistentes para procesos ARMA(p,q) en general aparece en Tsay y Tiao (1984) mediante un proceso de iteración de los ajustes autorregresivos. La idea del proceso iterativo les conduce además a introducir la Función de Autocorrelación Muestral Extendida (FAME) como una herramienta para la identificacioón del orden (p,q) de un modelo ARMA.

El símbolo $\doteq$ significará equivalencia asintótica y $\xrightarrow{p}$ convergencia en probabilidad.

Cabe hacer notar que en este trabajo se prersentan algunos lemas y teoremas sin demostración, las cuales se pueden consultar en Tsay y Tiao (1984) o para más detalles en Islas (1989).

1 REGRESIONES ITERADAS

El objetivo principal es encontrar estimadores consistentes por el método de MCO para los parámetros autorregresivos Φ_1's. La idea

desarrollada por Tsay y Tiao para solventar la inconsistencia en los parámetros Φ_i's mediante un ajuste AR(p), es la de incluir los residuos de dicho ajuste como nuevo regresor, reestimando el modelo y repitiendo el proceso; es decir, partiendo del modelo

$$Z_t = \sum_{i=1}^{p} \Phi_i Z_{t-i} - \sum_{j=1}^{q} \Theta_j a_{t-j} + a_t \qquad (1.1)$$

primero se hace un ajuste AR(p), dicha regresión puede escribirse como

$$Z_t = \sum_{i=1}^{p} \Phi^{(0)}_{i(p)} Z_{t-i} + e^{(0)}_{p,t} \qquad \text{para } t=p+1, \ldots, n \qquad (1.2)$$

en donde el supraíndice (0) indica la autorregresión ordinaria, mientras que el subíndice (p) indica el orden del ajuste AR y $e^{(0)}_{p,t}$ es el error en dicho ajuste. Se sabe que los estimadores de MCO $\hat{\Phi}^{(0)}_{i(p)}$ son consistentes para los Φ_i's, esto es

$$\hat{\Phi}^{(0)}_{i(p)} \xrightarrow{\quad p \quad} \Phi_i \qquad \text{para } i= 1,2, \ldots, p$$

si el proceso en cuestión sigue un modelo puramente AR(p) o un modelo puramente no estacionario ARMA(p,q) ($\phi(B)=1$). Estos resultados fueron probados por Mann y Wald (1943) para el caso estacionario, mientras que para el caso no estacionario se puede encontrar en Tiao y Tsay (1983).

Ahora considérese el caso en que los $\hat{\Phi}^{(0)}_{i(p)}$ son sesgados. Ya que los residuales de (1.2)

$$\hat{e}^{(0)}_{p,t} = Z_t - \sum_{i=1}^{p} \hat{\Phi}^{(0)}_{i,p} Z_{t-i}$$

no son ruido blanco para n suficientemente grande, entonces los valores retrasados de $\hat{e}^{(0)}_{p,t-j}$ para j>0, contienen algúna información acerca del

proceso $\{Z_t\}$. Esto motivó la definición de la primera regresión iterada AR(p) como

$$Z_t = \sum_{i=1}^{p} \Phi_{i(p)}^{(1)} Z_{t-i} + \beta_{1(p)}^{(1)} \hat{e}_{p,t-1}^{(0)} + e_{p,t}^{(1)} \quad \text{para } t = p+2, \ldots, n \qquad (1.3)$$

en donde el supraíndice (1) indica la primera regresión iterada y $e_{p,t}^{(1)}$ es el error correspondiente a esta primera regresión iterada. Se mostrará posteriormente que los estimadores por MCO $\hat{\Phi}_{i(p)}^{(1)}$ de esta regresión son consistentes, esto es

$$\hat{\Phi}_{i(p)}^{(1)} \xrightarrow{\quad p \quad} \Phi_i \quad \text{para } i = 1,2,\ldots, p$$

si $1 \geq q$ o $\phi(B) = 1$.

En la práctica el orden real (p,q) de un proceso ARMA usualmente es desconocido. Esta consideración conduce a plantear el proceso de regresiones iteradas a partir de un ajuste AR(m), con $m > 0$, con lo que se trata simultáneamente el problema de estimación y el de identificación del modelo. Específicamente, la j-ésima regresión iterada AR(m) de una serie de tiempo $\{Z_t\}$ se define como

$$Z_t = \sum_{i=1}^{m} \Phi_{i(m)}^{(j)} Z_{t-j} + \sum_{l=1}^{j} \hat{\beta}_{l(m)}^{(j)} \hat{e}_{m,t-j}^{(j-l)} + e_{m,t}^{(j)} \quad \text{para } t = m+j+1, \ldots, n ; \qquad (1.4)$$

$j = 0,1, \ldots$ y $m = 1,2, \ldots$

en donde

$$\hat{e}_{m,t}^{(j)} = Z_t - \sum^{m} \hat{\Phi}_{i(m)}^{(j)} Z_{t-i} + \sum^{j} \hat{\beta}_{l(m)}^{(j)} \hat{e}_{m,t-l}^{(j-l)} \qquad (1.5)$$

es el residual de la j-ésima regresión iterada AR(m) y los $\hat{\Phi}_{i(m)}^{(j)}$'s, $\hat{\beta}_{l(m)}^{(j)}$'s son los estimadores por el método de MCO de los $\Phi_{i(m)}^{(j)}$'s y $\beta_{l(m)}^{(j)}$'s

respectivamente.

2 PROPIEDADES DE LOS ESTIMADORES DE LAS REGRESIONES ITERADAS.

En esta sección se mostrará cómo los estimadores de la j-ésima iteración $\hat{\Phi}^{(j)}_{i(m)}$ pueden ser calculados recursivamente mediante mínimos cuadrados por etapas. Para esto, sea

$$\hat{\Phi}^{(j)}_{m} = 1 - \hat{\Phi}^{(j)}_{1(m)}B - \ldots - \hat{\Phi}^{(j)}_{m(m)}B^{m} \qquad (2.1)$$

el estimador del polinomio autorregresivo en la j-ésima regresión iterada AR(m) (1.4).

Los residuales de las regresiones iteradas, $\hat{e}^{(j)}_{m,t}$, verifican una relación recurrente de gran trascendencia para las propiedades de los estimadores $\hat{\Phi}^{(j)}_{i(m)}$, expresada en el siguiente lema.

Lema 2.1 Para cualesquiera enteros positivos m y j se verifica que

$$\hat{e}^{(j)}_{m,t} = \hat{e}^{(j-1)}_{m+1,t}$$

Como una consecuencia inmediata del lema (2.1) se tiene el siguiente resultado

Lema 2.2 Para cualesquiera enteros positivos m y j se verifica que

$$\hat{\Phi}^{(0)}_{m+j}(B) = \hat{\Phi}^{(j)}_{m}(B) - \sum_{l=1}^{n} \hat{\beta}^{(j)}_{l(m)} \hat{\Phi}^{(0)}_{m+j-l}(B)B^{l} \qquad (2.2)$$

Otra importante consecuencia del lema (2.1) es una expresión recurrente para los estimadores $\hat{\Phi}^{(j)}_{i(m)}$ que permite en último extremo calcularlos todos en función de estimadores $\hat{\Phi}^{(0)}_{i(m)}$, es decir, a partir de estimadores obtenidos en ajustes autorregresivos ordinarios.

Lema 2.3 Para cualesquiera enteros positivos m y j se verifica que

$$\hat{\Phi}_m^{(j)}(B) = \hat{\Phi}_{m+1}^{(j-1)}(B) + \hat{\alpha}_m^{(j-1)}\hat{\Phi}_m^{(j-1)}(B)B,$$

en donde

$$\hat{\alpha}_m^{(j-1)} = -\hat{\Phi}_{m+1(m+1)}^{(j-1)} \quad \hat{\Phi}_{m(m)}^{(j-1)}$$

Corolario 2.1 Se verifica que

$$\hat{\Phi}_{i(m)}^{(j)} = \hat{\Phi}_{i-1(m)}^{(j-1)} - \hat{\Phi}_{i(m)}^{(j-1)} \quad \hat{\Phi}_{m+1(m+1)}^{(j-1)} \quad \hat{\Phi}_{m(m)}^{(j-1)}$$

para $i = 1,2, \ldots, m$; $m \geq 1$ y $j \geq 2$ en donde $\hat{\Phi}_{0(m)}^{(j-1)} = -1$.

3 RELACION CON LAS SOLUCIONES DE LAS ECUACIONES DE MOMENTOS

Supóngase que el proceso $\{Z_t\}$ es estacionario. Se estudiará ahora la relación entre los estimadores de parámetros autorregresivos de un ajuste AR(m) en su j-ésima iteración y las soluciones de las ecuaciones de momentos muestrales (las ecuaciones de Yule-Walker estimadas).

En lo que sigue el símbolo $\sum_t$ indica que la suma es tomada desde m+j+1 hasta n a menos que se indique otra cosa.

También recuérdese que, si $\{X_n\}$ es una sucesión de variables aleatorias y $\{g_n\}$ una sucesión de números reales positivos, se dice que $\{X_n\}$ es a lo más de orden (g_n) en probabilidad y se denota como

$$X_n = O_p(g_n)$$

si, para toda $\varepsilon > 0$, exixte un número real positivo M_ε tal que

$$P\{|X_n| \geq M_\varepsilon\, g_n\} \leq \varepsilon \qquad \text{para toda } n$$

Lema 3.1 Si $j \geq 0$ y $m \geq 1$ se verifica

$$\sum_t Z_{t-h}\, \hat{e}_{m,t}^{(j)} = 0 \qquad \text{para} \quad h = 1,2, \ldots, m+j \tag{3.1}$$

Lema 3.2 Sea $\{Z_t\}$ una serie estacionaria. Se verifica que para $1 \leq i \leq j$

$$\sum_t Z_{t-h}\, \hat{e}^{(j-1)}_{m,t-i} = O_p(1) \quad \text{para } h=i+1, \ldots, m+j \qquad (3.2)$$

Por otra parte, los estimadores $\hat{\beta}^{(j)}_{l(m)}$ son asintóticamente funciones lineales del vector $\underline{V}_j \; \hat{\Phi}^{(j)}_m(B)\hat{\rho}_h \; : h=1,2,\ldots,m)$. en consecuencia, también dichos estimadores son $O_p(1)$.

A partir de la j-ésima iteración del ajuste AR(m) para Z_t, se obtiene el siguiente sistema de ecuaciones

$$\sum_t Z_{t-h} W_{j,t} = \sum_{l=1}^{j} \hat{\beta}^{(j)}_{l(m)} \left\{ \sum_t Z_{t-h}\, \hat{e}^{(j-1)}_{m,t-l} \right\} \; ; \; h=j+1, \ldots, j+m \qquad (3.3)$$

en donde

$$W_{j,t} = Z_t - \sum_{i=1}^{m} \Phi^{(j)}_{i(m)}\, Z_{t-i}$$

En virtud del lema (3.2) y de lo dicho para los estimadores $\hat{\beta}^{(j)}_{l(m)}$ se sigue que el miembro de la derecha de (3.3) es $O_p(1)$. Así pues, dividiendo por $\sum_t Z_t^2$, el sistema (3.3) se transforma en

$$\hat{\Phi}^{(j)}_m(B)\hat{\rho}_h = O_p(n^{-1}) \; ; \quad h= j+1, \ldots, j+m \qquad (3.4)$$

La relación (3.4) se traduce en el siguiente resultado

Lema 3.3 Si $\{Z_t\}$ es una serie estacionaria, los estimadores de los parametros autorregresivos de un ajuste AR(m) en su j-ésima iteración, $m \geq 1$ y $j \geq 0$, son asintóticamente equivalentes a la solución de las ecuaciones generalizadas de Yule-Walker

$$\rho_h = \sum_{i=1}^{m} \Phi_i \rho_{h-i} \qquad \text{para} \quad j+1 \leq h \leq j+m \qquad (3.5)$$

Por otra parte, es bien sabido que $\Phi_1, \ldots, \Phi_p$ satisfacen (3.5) para $j \geq q$ y $m=p$. De hecho, haciendo $\Phi_i = 0$ para $i>p$, los coeficientes $\{\Phi_i\}_{1 \leq i \leq k}$ verifican (3.5) para $j \geq q$ y $m \geq p$.

4 CONSISTENCIA DE LOS PARAMETROS AUTORREGRESIVOS ESTIMADOS MEDIANTE UN AJUSTE AR(m) EN SU J-ESIMA ITERACION

En esta sección se consideran las propiedades de consistencia de los parámetros autorregresivos estimados mediante un ajuste AR(m) en su j-ésima iteración; $m \geq 1$ y $j \geq 0$.

Recuérdese que, si X_n es una sucesión de variables aleatorias y C una constante, entonces

$$X_n = C + O_p(n^{-\delta}) \; ; \; \delta > 0, \quad \text{implica} \quad X_n \xrightarrow{p} C$$

Como una consecuencia del lema (3.3), se tiene que la consistencia de los parámetros autorregresivos estaría probada si la solución de las ecuaciones generalizadas de Yule-Walker fuese única. El siguiente lema dá condiciones sobre j y m que aseguran dicha unicidad y, por ende, la consistencia.

Lema 4.1 Sea $A_m^{(j)}$ la matriz del sistema (3.5), entonces se verifica que a) $|A_p^{(j)}| \neq 0$ para $j \geq q$ b) $|A_m^{(q)}| \neq 0$ para $m \geq p$

Como consecuencia inmediata del lema (4.1) se tiene que para un modelo estacionario ARMA(p,q), la solución de las ecuaciones

$$\hat{\rho}_h - \sum_{i=1}^{m} \hat{\phi}_i \hat{\rho}_{h-i} = 0 \quad ; \quad h = j+1, \ \ldots \ , j+m$$

para a) $m \geq p$ y $j = q$ o b) $m = p$ y $j > q$
también proporcionan estimadores consistentes de los parámetros autorregresivos. Entonces, usando el resultado (3.4) y el lema (3.3), se demuestra fácilmente el siguiente teorema.

Teorema 4.1 Supóngase que $\{Z_t\}$ sigue un proceso estacionario ARMA(p,q). Entonces

$$\hat{\phi}_{i(m)}^{(j)} = \phi_i + O_p(n^{-1}) \quad \text{para} \quad i = 1, 2, \ \ldots \ , m$$

si

i) m≥p y j=q o *ii*) m=p y j>q

en donde se entiende que ϕ_i=0 para *i*>p.

El siguiente resultado es una generalización del anterior, que permite observar cómo el proceso de regresiones iteradas se puede manejar, tanto para procesos ARMA estacionarios como los no estacionarios.

Teorema 4.2 Supóngase que $\{Z_t\}$ sigue un proceso ARMA(p,q) ya sea estacionario o no estacionario. Entonces

$$\hat{\Phi}^{(j)}_{i(m)} = \Phi_i + O_p(n^{-1/2}) \quad \text{para } i=1,2, \ldots ,m$$

si

i) m≥p y j=q o *ii*) m=p y j>q,

en donde se entiende que Φ_i=0 para *i*>p. También, $O_p(n^{-1/2})$ se transforma en $O_p(n^{-1})$ si $\phi_{m-d}(B) = 1$, m=p y j≥q.

5 FUNCION DE AUTOCORRELACION MUESTRAL EXTENDIDA (FAME).

La conexión que se ha puesto de manifiesto entre el orden (p,q) de un proceso ARMA y la consistencia de los estimadores de los parámetros autorregresivos, es una herramienta idónea para la identificación del modelo. Esto se logra mediante la FAME, que se define mediante la igualdad

$$\hat{\rho}_{j(m)} = \hat{\rho}_j \left[W^{(j)}_{m,t} \right] \qquad (5.1)$$

en donde

$$W^{(j)}_{m,t} = Z_t - \sum_{i=1}^{m} \hat{\Phi}^{(j)}_{i(m)} Z_{t-i}$$

nótese que para m=0, $\hat{\rho}_{j(0)}$ se reduce a la FAC usual.

Como se verá, la FAME tiene una propiedad asintótica de "punto de corte" justo en el orden (p,q) del modelo, que generaliza a la FAC de los procesos de promedios móviles y a la FACP de los procesos autorregresivos.

Una primera explicación a la utilización de la serie transformada $W^{(j)}_{m,t}$ en la definición (5.1) se tiene en el hecho de que para m=p y j≥q se sigue un proceso MA(q), como consecuencia inmediata de los teoremas de

consistencia. Por lo tanto, se verifica que

$$\hat{\rho}_{j(p)} \doteq 0 \ , \quad \text{para } j \geq q+1$$

La estructura que sigue la serie $W_{m,t}^{(j)}$ para $m \geq p$, viene dada por el siguiente lema

Lema 5.1 Para $m \geq p$ y $j \geq q$, la serie $W_{m,t}^{(j)}$ sigue un proceso MA(q+h), siendo h=mín {m-p, j-q}.

Como consecuencia se verifica la siguiente propiedad de punto de corte asintótico.

Teorema 5.1 $\hat{\rho}_{j(m)} \doteq 0$, para $m \geq p$ y $j > q-p+m$

6. LA IDENTIFICACION DE LOS MODELOS ARMA MEDIANTE LA FAME.

Las propiedades asintóticas de la FAME mostradas en el teoreama (5.1), pueden ser utilizadas en la práctica para identificar modelos ARMA(p,q). Para este propósito resulta útil arreglar las autocorrelaciones muestrales como se muestra en la siguiente tabla

TABLA 6.1

LA FUNCION DE AUTOCORRELACION MUESTRAL EXTENDIDA

AR \ MA	0	1	2	3	...
0	$\hat{\rho}_{1(0)}$	$\hat{\rho}_{2(0)}$	$\hat{\rho}_{3(0)}$	$\hat{\rho}_{4(0)}$	...
1	$\hat{\rho}_{1(1)}$	$\hat{\rho}_{2(1)}$	$\hat{\rho}_{3(1)}$	$\hat{\rho}_{4(1)}$	...
2	$\hat{\rho}_{2(2)}$	$\hat{\rho}_{2(2)}$	$\hat{\rho}_{3(2)}$	$\hat{\rho}_{4(2)}$	...
...			...		

En el primer renglón de esta tabla se localiza la FAC, mientras que el segundo renglón muestra la primera FAME, el tercero la segunda FAME y así sucesivamente. Obsérvese que los renglones estan numerados del 0,1,2,... lo cual indica el orden AR, de manera similar se numeran las columnas para indicar el orden de promedios móviles MA.

Para ilustrar el uso de la tabla (6.1), supóngase que el modelo real es un ARMA(p,q). De las propiedades de la FAC es bien sabido que $\hat{\rho}_{j(0)} \neq 0$

para $j \geq q+1$. Ahora del teoreama(5.1) se puede observar que

i) cuando $m=p$ $\quad \hat{\rho}_{j(p)} \doteq 0 \quad$ para $j \geq q+1$

ii) cuando $m=p+1$ $\quad \hat{\rho}_{j(p+1)} \doteq 0 \quad$ para $j > q+2$

y así sucesivamente

Para identificar el orden (p,q) de un modelo ARMA mediante la FAME se busca en la tabla (6.1) el "punto de corte triangular", es decir el punto (c_1, c_2) tal que las rectas $m = c_1$ y $m-j=c_2$ delimiten un triángulo de ceros, tal como el de la tabla (6.2), en donde X denota un valor diferente de cero y 0 es un valor igual a cero. El renglón y columna coordenadas del vértice de este triángulo, corresponden precisamente con el orden AR y MA respectivamente.

TABLA 6.2

COMPORTAMIENTO ASINTOTICO DE LA FAME EN UN MODELO ARMA(p,q)

AR \ MA	...	q-1	q	q+1	q+2	q+3	...
...				...			
p-1		X	X	X	X	X	...
P		X	0	0	0	0	...
P+1		X	X	0	0	0	...
P+3		X	X	X	0	0	...
...			...				

En la práctica, para muestras finitas, los elementos $\hat{\rho}_{j(m)}$ no serán iguales a cero. Por lo tanto, para detectar el orden de un modelo ARMA se utiliza una tabla "simbólica" similar a la tabla (6.1) en cuyo lugar (m,j) se coloca un cero si $\hat{\rho}_{m,j+1}$ es asintóticamente nulo (al nivel de significación del 5%), es decir, si $|\hat{\rho}_{m,j+1}| \leq 2\sigma_{m,j+1}$, siendo $\sigma^2_{m,j+1}$ la varianza asintótica de $\hat{\rho}_{m,j+1}$. En caso contrario, se coloca, por ejemplo, una X. A estos efectos, se precisa una estimación de la varianza asintótica de los coeficientes $\hat{\rho}_{m,j+1}$. Tiao y Tsay proponen como estimación a $(N-m-j)^{-1}$, aproximación que resulta de la fórmula de Bartlett, bajo la hipótesis de que $\{W^{(j)}_{m,t}\}$ es ruido blanco, aunque

reconocen que posiblemente sea una estimación sesgada a la baja.

El problema que presenta este método es la falta de unicidad, pues la teoría de la FAME sólo determina el comportamiento asintótico para $j \geq q$ y $m \geq p$, pero no impide la existencia de puntos de corte para $m < p$.

7 CONCLUSIONES

La ventaja que presenta este método con respecto al de Box-Jenkins, es que la FAME se aplica directamente a series de tiempo estacionarias y no estacionarias, evitándose con esto una transformación de los datos de la serie para volverla estacionaria antes de especificar el orden del modelo. Una desventaja que presenta este método, es el problema de falta de unicidad, pues la teoría de la FAME sólo determina el comportamiento asintótico para $j \geq q$ y $m \geq p$, pero no impide la existencia de puntos de corte con $m < p$. Sin embargo, la discriminación entre varios puntos de corte posibles, puede lograrse mediante la etapa de estimación de los parámetros autorregresivos y las propiedades de convergencia que presentan (teorema 4.2).

BIBLIOGRAFIA

Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forescasting and Control.* San Francisco: Holden Day.

Islas, A. (1989). "Métodos para determinar el orden de un modelo autorregresivo y de promedios móviles para series de tiempo". Tesis de Maestria en matemáticas UAM-I.

Mann, H.B., and Wald, A. (1943). "On the Statistical Treatmen of Linear Stochastic Equations," *Econometrica,* 11,173.

Tiao, G. C., and Tsay, R.S. (1983). "Consistency Properties of Least Squares Estimates of Autoregressive Parameters in ARMA Models. *Annals of Statistics,* 11, 856-871.

Tsay, R.S. & Tiao, G.C. (1984). "Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models". *j. Am. Statist. Assoc.* 79, 84-96.

ESTUDIO DE LA RELACIÓN MIGRACIÓN-SALUD MENTAL, DESDE UN PUNTO DE VISTA DE LA PSIQUIATRÍA SOCIAL.

I. Rafael Madrid Rios, UACPYP-IIMAS-UNAM.
Susana Cuevas Córdova, Instituto Mexicano de Psiquiatría.
Francisco Javier Aranda Ordaz, IIMAS-UNAM.

Resumen

El trabajo que se presenta es el producto de la planeación de una investigación en proceso. El diseño de la investigación es de tipo comparativo observacional, que permitirá hacer la primera aproximación sobre la relación de causalidad entre la migración y la salud mental. En el estudio se toman en cuenta otras variables que pudieran actuar como factores de confusión (sexo, otros eventos estresantes, rasgos de personalidad y redes sociales de apoyo). Se propone el empleo de modelos logísticos para respuesta politómica ordenada, que se ejemplifica en un conjunto de datos que no son los del estudio.

1. Introducción.

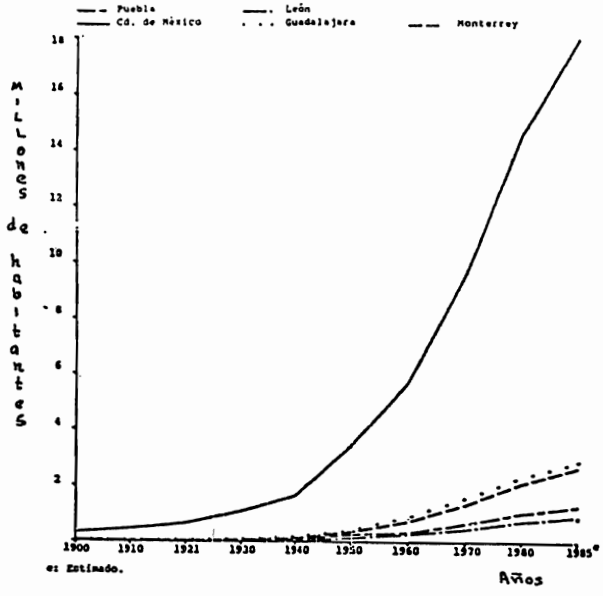
El estudio de la migración rural-urbana en relación a la salud mental es importante por dos razones:

1.- La magnitud de los movimientos migratorios (Oliveira y García , 1984 ; Unikel, 1976).

2.- Su trascendencia hacia todos los ámbitos de la vida individual y socio-económica (Arizpe, 1978 ; Muñoz , et al, 1972).

En relación a la magnitud, se sabe que la Ciudad de México es una de las ciudades más grandes del mundo y que presenta tasas de crecimiento poblacional muy altas a nivel internacional . Por ejemplo, la tasa de crecimiento medio anual de 4.7% presentada para la década 70-80, se ubica muy por encima de las tenidas en las principales ciudades europeas y estado unidenses que presentan tasas de a los más el 2% anual(Ruíz de Chavez, 1978; Oliveira y García ,1984); dicha tasa ha dado lugar a perfiles de crecimiento urbano que en el interior del país contrastan en la forma que se presenta en la gráfica A. Componen este crecimiento demográfico tan intenso el crecimiento natural (resultante de la diferencia de la natalidad y mortalidad) y, primordialmente, el crecimiento social o derivado de los movimientos migratorios.

Gráfica A. POBLACION DE CINCO CIUDADES MAYORES. 1900-1985*



FUENTE: México Social 1985 - 1986.

Para varios investigadores sociales, los movimientos migratorios de la Ciudad de México contribuyen en más del 50% a las tasa de crecimiento y en una proporción mayor cuando se considera la tasa de natalidad entre los migrantes (Muñoz, et al, 1972).

Con respecto a la trascendencia de estos movimientos en los diversos niveles de la vida individual y social, ésta se deriva de las características de la población que migra y de los efectos que los movimientos masivos producen tanto en el lugar de origen como en el de destino. La movilización masiva de población rural depauperada, joven, con bajos niveles sociales y culturales, genera desequilibrio en la estructura demográfica, ocupacional y productiva en ambos polos del movimiento y múltiples y diversos problemas sociales, económicos y de salud en la áreas de asentamiento (Arizpe, 1978).

Una observación frecuente que se hace en la literatura médica, se refiere, a que el individuo en proceso de transición física y cultural presenta mayor riesgo de enfermar . El individuo que migra del campo a la ciudad presenta una serie de cambios, pérdidas y ganancias que afectan su vida en aspectos como los hábitos alimenticios, la vestimenta, la pérdida de lazo afectivos y amistosos, al tiempo que puede obtener ganancias y realizaciones en los aspectos pecuniarios y laborales.

La serie de cambios impuestos al migrante exigen de él una respuesta adaptiva que le permita desenvolverse en las nuevas condiciones de vida para las cuales, el migrante rural, proveniente de las capas socio-económicas más bajas del campesinado, no está capacitado.

La forma como la migración influye en la salud mental es compleja y ambigua . La revisión bibliográfica señala una evolución en la concepción y metodologías que ha superado enfoques reduccionistas de la migración como un evento dicotómico (migrante / no migrante), hasta llegar a enfoques teórico-conceptuales capaces de integrar la multidemensionalidad de un proceso al que concurren factores de orden económico, cultural, psicológico, etc (Cuevas, S, 1989). Sin embargo, el enfoque con respecto a la salud mental aún presenta dos problemas que limitan la evaluación del conocimiento en psiquiatría social. El primero consiste en el tratamiento siempre patologizante que se dá a la variable salud mental, sin tener en cuenta que la migración conlleva una serie de mejoras y ganancias socioeconómicas (sobre todo en la experiencia mexicana), las cuales pueden constituir estímulos positivos y realizaciones para el individuo, de tal manera que inciden positivamente en la salud del migrante. El segundo surge de la reducción que se hace del complejo evento de la salud mental a la variable dicotómica enfermo-no enfermo. Esta crítica es válida sobre todo desde el punto de vista de la presente investigación. El conocimiento psiquiatrico social se encuentra aún en estadíos iniciales en los que la investigación descriptiva no debe de tener cortapisas ni prejuicios impuestos por la tradición y el conocimiento alcanzado por los expertos en el nivel clínico y y biológico; dicha escotomización representa una pérdida de

información que impide conocer con mayor precisión interacciones probables entre eventos sociales y psicológicos y enriquecer el conocimiento sobre la causalidad social de la enfermedad mental.

Otros dos elementos del problema son las diferencias metodológicas y conceptuales, que dan poca consistencia al conocimiento que pudiera ser derivado e impiden su utilización como material empirico antecedente y la carencia aparentemente absoluta de investigaciones sobre el tema en este país.

En conclusión, la investigación sobre salud mental y migración rural urbana implica una problemática diversa y compleja, sobre la cual aún pesan varias insuficiencias. Dadas sus características de magnitud y trascendencia en el caso de nuestro país, el problema puede ser considerado como uno de la salud pública sobre el cuál nada se ha investigado.

2. Planeación.

En la Psiquiatría social se considera la experiencia migratoria como un proceso o experiencia psicosocial potencialmente generadora de estres (American Psychiatric Association, 1980). Bajo esta consideración es que se pretende estudiar el efecto de la migración, medida a través de los eventos estresantes que produce, en la calidad de la salud mental, tanto en su aspecto patológico como de bienestar mental.

En particular el trabajo tiene como objetivo general "Realizar un primer estudio de aproximación a la relación de causalidad entre el proceso de migración y la calidad de salud mental". Por lo cual se ha diseñado un estudio de tipo "Encuesta Comparativa" que tiene como caracteristicas ser observacional, comparativo y prospectivo (Méndez et al, 1984) y en donde las hipótesis propuestas son:

Hipótesis conceptual : La migración incide sobre la salud mental.

Hipótesis operacionales:

(1) La Migración se relaciona con la patología mental a través de la generación de estres que la experiencia migratoria produce.

(2) Los logros y las realizaciones condicionados por la migración producirán niveles positivos en salud mental.

(3) El tipo de personalidad de los individuos y el apoyo de sus redes sociales, son variables intermedias que modifican la relación migración-trastorno mental. De esta forma tenemos que:

(3.1) Ante un alto nivel de estres generado por la migración, cuando las redes sociales de apoyo sean "inefectivas" y la personalidad "inapropiada" se observarán altos niveles de trastorno mental.

(3.2) Ante un alto nivel de estres pero con redes sociales de apoyo "efectivas" y personalidad "apropiada" el nivel de trastorno mental será comparativamente mas bajo que en 3.1.

En el cuadro A se presentan las variables por considerar en el estudio y sus definiciones. La información del proceso de migración será medida como un conjunto de factores estresantes en tres momentos (antes, durante y despues del asentamiento) mediante un cuestionario diseñado ex professo. En cuanto al diagnóstico de trastorno mental (ansiedad, depresión y somatizaciones) será medido también en los mismos momentos que la migración utilizando el cuestionario estructurado DIS (Diagnostic Interview Schedule). La información sobre redes sociales de apoyo se colectará tambien a través de un cuestionario diseñdo ex professo y la referente a personalidad y eventos estresantes mediante cuestionarios ya existentes.

Se definirá como MIGRANTE a los individuos de cualquier sexo, nacidos en cualquier entidad de la República Mexicana (a excepción del Edo de México y el D.F), con tiempo de haber migrado igual o menor a cinco años, que permanezcan más de un año o con miras a establecerse en la Ciudad de México o área conurbada y que migren por causas económicas. Como NO MIGRANTE a los nacidos en el D.F o Edo de México. En ambos grupos el rango de edad de los individuos en estudio será de 18-60 años.

En cuanto a la forma de tomar la muestra, se considerará como población en estudio a los habitantes de una área conurbada del municipio de Chalco del Estado de México, en donde se sabe que existen intensos flujos de población rural inmigrante. La información con la que se cuenta es:

1. Mapa de linea con información de AGEBS (áreas geoestadísticas básicas).

CUADRO A INFORMACIÓN POR OBTENER PARA ESTUDIAR LA RELACIÓN MIGRACIÓN-SALUD MENTAL

VARIABLES EN ESTUDIO	DEFINICION CONCEPTUAL	DEFINICION OPERACIONAL
+ SALUD MENTAL* . Número de sintomas. . Tipo de sintoma. . Momento de aparición. . Severidad.	Proceso de óptimo desarrollo de las potencialidades y capacidades del individuo ; parte indisoluble de la salud global en interrelación dinámica con el medio.	Proceso de salud-enfermedad que se desarrolla a lo largo de un eje o continuo, el cual contiene en sus polos la parte de salud positiva y negativa.
+ MIGRACIÓN*, ** . número de eventos estresantes. . Tipo de eventos.	Movilidad geográfica: Determinantes: Económicos (nivel macro estructural), Políticos y Culturales (nivel regional), Psicológicos y Familiares(nivel personal).	Poceso de movilidad geográfica que representa una experiencia Psicosocial,donde se presentan situaciones de cambios, pérdidas y ganancias potencialmente generadoras de estres.

a.- Aspectos nosológicos:
Somatización, Ansiedad, Depresión.
b.- Indices de satisfacción.- Aspectos de interés que informan de manera indirecta sobre la salud mental positiva de un individuo.

VARIABLES INTERMEDIAS (ATENUANTES O CONTRIBUYENTES AL ESTRES):

VARIABLES EN ESTUDIO	DEFINICION CONCEPTUAL	DEFINICION OPERACIONAL
+ REDES SOCIALES DE APOYO . Calidad. . Efectividad.	Conjunto de vínculos y conductas sociales, cuyo tamaño , estructura y efectividad tienen un papel importante en la salud, tanto física como mental.	Ausencia o presencia relativa del apoyo Psico-Social proviniente de personas significativas con el propósito de mantener el bienestar de las partes.

+ PERFIL O RASGO DE PERSONALIDAD.

Extro-introversión , Neuroticismo, Psicocitismo, Agradabilidad social.

+ OTROS EVENTOS ESTRESANTES QUE CONDICIONAN UN TRASTORNO MENTAL:

Número de eventos, Tipo de eventos, Momento de aparición.

* Referidas : antes, durante y en el asentamiento.
** Solo se mide en migrantes.

2. Porcentajes de prevalencia de las categorías nosológicas de trastorno mental que se pretenden estudiar: Somatización, Ansiedad y Depresión (Cuevas, 1990; Burnam, et al, 1987; Karno, et al, 1987 ; Burnam, et al.1987 ;Canino, et al, 1987 ; American Psychiatric Association, 1980).

El área de estudio se eligió en base a la información proporcionada por las autoridades de salud del estado y subsecuentemente por estudios de observación y exploración directa en campo.

El tipo de muestreo que se llevará a cabo es estratificado en los AGEBS y polietápico. La estratifiacación se hará de acuerdo al tamaño de los AGEBS y en cada uno de estos se seleccionarán manzanas (unidades de muestreo de primera etapa) con igual probabilidad, de las manzanas muestra se tomarán las viviendas (unidades de muestreo de segunda etapa) aleatoriamente y finalmente en las viviendas se seleccionará también de manera aleatoria a un miembro de la familia (unidad última de muestreo).
La recolección de los datos se realizará a través de cuestionarios en entrevista directa por encuestadores entrenados.

En la obtención de la información se tendrán los cuidados necesarios para que los datos que se recaben para el análisis permitan obtener buenas estimaciones en términos de calidad y validez de las observaciones, ya que a partir de ellos se procedera a determinar sí la asociación migración-salud mental es significativa, y finalmente si la asociación resulta significativa considerar diversos elementos tanto estadísticos como de la Psiquiatría Social para establecer si la asociación encontrada pudiera validarse como una relación de tipo causal.

Es de importancia anotar que los siguientes aspectos del proceso de planeación en la investigación, se contemplan, lo cual permitirá contar con información de calidad:

- La utilización de instrumentos validados en México (Gonzáles, 1982; Eysenck y Lara, 1989).
- El piloteo y ajuste de las secciones del cuestionario diseñado ex professo, teniendo como escenario las hipótesis operacionales mencionadas.

- La selección de encuestadores con experiencia y formación sobre el área de estudio , así como su capacitación con medios audiovisuales y practicas.
- La supervisión y coordinación permanente del trabajo de campo.
- Al control y ajuste periódico de trabajo de campo.

3. Análisis de la Información.

A continuación se describe el tipo de modelos que se pretende emplear para el estudio de la relación migración-salud mental. Debido a que el proyecto de investigación esta en desarrollo, se ha optado por utilizar a manera de ilustración los datos de un estudio sobre efecto de drogas-postoperación en la disminución de dolor.

Como la investigación que se lleva a cabo es comparativa - observacional las categorías del factor en estudio (migración) no son asignadas por el investigador, aleatoriamente a los individuos como cuando el estudio es experimental, por lo cual se le designa factor de riesgo o de exposicion. Así el factor migración (de riesgo) se refiere al agente a que estan expuestos los individuos y cuyo efecto en la salud mental (factor respuesta) quiere determinarse.

MODELOS LOGÍSTICOS GENERALIZADOS (Respuesta politómica)

Se refiere a un modelo de regresión particular para estudiar una variable respuesta (salud mental) con más categorías que una respuesta binaria y considerando para su explicación ciertas variables (eventos estresantes producidos por la migración: en número e intensidad=X_1, otros eventos estresantes=X_2, rasgos de personalidad =X_3, redes sociales de apoyo=X_4 y sexo=X_5) que se supone estan relacionadas con la respuesta.

En la investigación la variable explicativa (factor de riesgo o de exposición) en estudio es la migración (que se puede expresar como una conbinación de número de eventos estresantes, tipo e intensidad), sin embargo como es posible que la salud mental pueda ser explicada o modificada por otros factores (X_2,X_3,X_4,X_5) es conveniente estudiar si estas variables en su efecto interactuan

con el factor de riesgo en cuestión y explican la relación del factor de riesgo con la respuesta. Así entonces, al inicio de la investigación las variables explicativas X_2, X_3, X_4 y X_5 se consideran como factores de confusión de la relación migración-salud mental.

Cuando la interacción entre el factor de riesgo y los que se proponen como de confusión no existe se tendrán elementos para validar la relación propuesta. Si la interacción existe será necesario replantear el estudio considerando realmente como otros factores de riesgo a los que en principio se consideraban como de confusión y tomarlos en cuenta para planear la recolección de la información en un próximo estudio.

El tipo de modelo logístico politómico por emplear depende del número de factores de riesgo, de confusión y de su naturaleza, asi por ejemplo para modelos logístico (para respuesta binaria) se tiene una diversidad de situaciones que pueden analizarse (Ducoing,A, 1988) por ejemplo.

+ Un factor de riesgo dicotómico (X_1) y respuesta dicotómica(Y)
+ Dos factores de riesgo dicotómicos y respuesta dicotómica.
+ Dos factores de riesgo un dicot ó mico, uno politómico y respuesta dicotómica.
+ Un factor de riesgo dicotómico (X_1),uno de confusión politómico (X_4) y la respuesta dicotómica (Y).
+ Un factor de riesgo politómico de 3 categorías ordenadas (X_1: menos de 5 eventos estresantes (ee), ó $\leq$ee $\leq$15, más de 16 ee) y la respuesta dicotómica (Y).
+ Un factor de riesgo continuo, factor de confusión con I niveles y respuesta dicotómica.

Las situaciones que interesa estudiar con los modelos logísticos politómicos o generalizados, son similares a las que se análizan con los modelos logisticos para respuesta binaria, con la única variante de que la respuesta politómica puede ser ordenada o no.

En el estudio si se considera por ejemplo el factor de exposición "condición migratoria" con cuatro niveles (uno para no migrante y tres para migrantes: menos de cinco eventos estresantes (ee) , ó $\leq$ ee $\leq$ 15, más de 16 ee) y el factor con respuesta politómica

"trastorno mental" con cuatro categorías (sin trastorno mental, con trastorno leve, medio y alto). Se tiene que la respuesta, para cada nivel del factor de exposicion, constituye una muestra aleatoria de una distribución multinomial.
Para el análisis de esta situación cuando la respuesta es ordenada existen diversos modelos logísticos de respuesta politómica que pueden emplearse, los cuales quedan determinados por la elección de la función logit que se considere para las probabilidades de la variable respuesta en los niveles del factor de exposición; algunos de estos modelos presentados por Cox y Chuang 1984, son:
M1.-Modelo logit para la JI-cuadrada particionada.
M2.-Modelo logit con nivel base respuesta la categoría más importante.
M3.-Modelo logit de momios acumulativos:por renglón o por columna.

Con el propósito de mostrar algunos de los análisis que utilizaremos para el estudio de la relación migración-salud mental cuando no se consideran factores de confusión, presentamos en un ejemplo con datos ajenos al proyecto (pues el estudio se encuentra en desarrollo) el tipo de conclusiones que pueden obtenerse con los modelos M1 Y M2.

EJEMPLO.

Este ejemplo se tomó de Cox y Chuang 1984, donde el estudio que se presenta es experimental. Sin embargo, estos modelos pueden emplearse en nuestro caso para estudiar la relación migración -salud mental aunque el posible apoyo a una relación de causalidad es mucho menor, por ser un estudio observacional.
Se pretendia estudiar el efecto de 4 drogas-postoperación en la reacción de los pacientes a la disminución de dolor. Se parte del supuesto de que los pacientes asignados a las cuatro drogas son comparables, es decir el efecto de la droga no esta confundido con diferencias poblacionales. La escala ordinal empleada para evaluar la disminución del dolor de manera global es:disminución (dism) pobre= C1 , dism moderada=C2, dism buena=C3 y dism de muy buena a excelente=C4.

M1.- JI CUADRADA PARTICIONADA.

El procedimiento de análisis que se sigue en la tabla total 4x4(RxC) es dividirla en 3 subtablas de dimensiones 4x2. Ya que la estadística de razón de verosimilitud (ERV = G^2) para la tabla completa consiste de 3(C-1) componentes aditivos, los cuales son respectivamente las ERV de las C-1 tablas. Esto sugiere entonces estudiar por separado la contribución de la hetereogeneidad observada en la tabla completa (Chuang , 1982) a través de las subtablas. Denotemos el número de subtabla con s, para formar la subtabla s (1,2,3) se procede como sigue:

la primer categoría de la subtabla s corresponderá con las frecuencias asociadas a la categoría de la respuesta de la tabla completa que sea igual con s en todos los niveles del factor de exposición y para la segunda categoría de la subtabla s se suman las frecuencias de las categorías de la respuesta de orden mayor que s para cada nivel del factor de exposición.

En este modelo los logíts en la C-1 categorías en cada uno de los niveles de exposición (Fienberg, 1980 ; Mc Cullagh's, 1980) son:

$$Lij = Ln\left(Pij / \sum_{t>j} Pit\right)$$

donde Pij =probabilidad para la respuesta j del nivel i del factor de exposición de la tabla completa.

Para el ejemplo que se plantea se obtiene lo siguiente:

TABLA TOTAL
frec obs y % de resp.

	C1	C2	C3	C4
R1	5	1	10	14
	.167	.033	.333	.467
R2	5	3	3	20
	.161	.097	.097	.645
R3	10	6	12	3
	.323	.193	.387	.097
R4	7	8	12	2
	.241	.414	.276	.069

χ^2= 43.89 , G^2= 46.74
gl= 9 , (p < .001)

SUBTABLA 1 frecuencias observadas

	C1	(C2+C3+C4)
R1	5	25
R2	5	26
R3	10	21
R4	7	22

G^2=3(gl=3,p=.40)
R1 vs R2 ⟶
R3 vs R4 ⟶
R1,R2 vs R3,R4⟶

SUBTABLA 2

	C2	(C3+C4)
R1	1	24
R2	3	23
R3	6	15
R4	12	10

G^2=19.86***
G^2= 1.05
G^2= 3.02□
G^2=15.79***

SUBTABLA 3

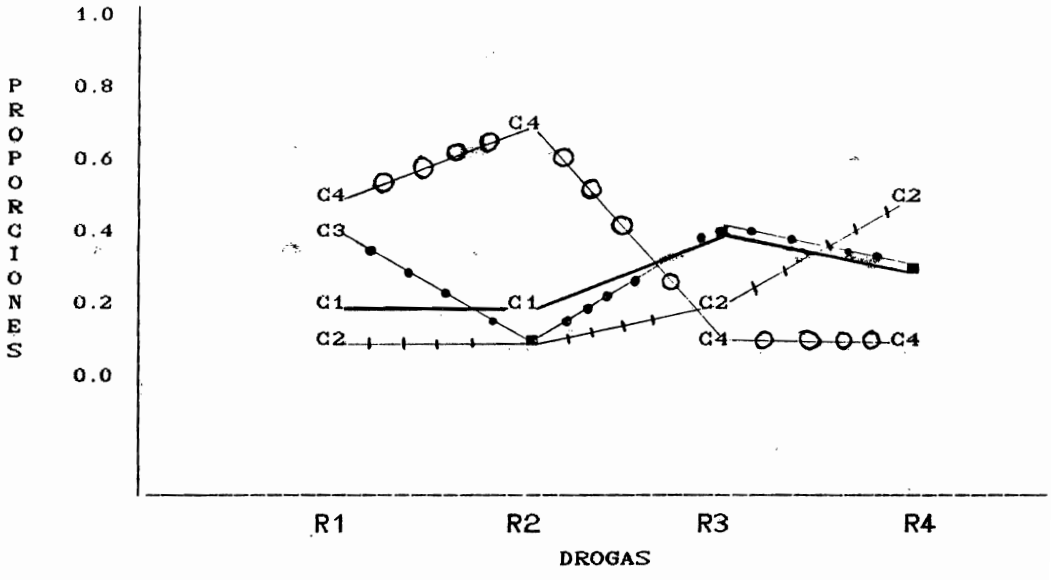
	C3	(C4)
R1	10	14
R2	3	20
R3	12	3
R4	8	2

G^2=23.88***
G^2= 5.02*
G^2= 0.00
G^2=18.86***

NOTA: □ (p=0.10), * (p<0.05), ** (p<0.01), *** (p<0.001).

De la tabla total puede observarse que los pacientes responden de manera distinta a las drogas. En particular, en la gráfica B pueden apreciarse las diferencias de las drogas en la respuesta.

GRAFICA B: EFECTO DE CUATRO DROGAS EN LA DISMINUCION DEL DOLOR



De la división hecha en la tabla total se puede concluir:

Subtabla 1: Las proporcion en la respuesta menos favorable (C1) a la disminución del dolor tiende a ser la misma en las 4 drogas.

Subtabla 1,2 y 3: Las proporciones observadas suguieren que R1-R2 y R3-R4 son similares y que los dos grupos difieren de manera sustancial uno de otro.

Subtabla 2 y 3: La diferencia sustancial entre los grupos se verifica. En dos hay cierta evidencia de que R3 sea más favorable que R4 (p=0.10) si excluimos la respuesta menos favorable. En 3 hay evidencia de que R2 es más favorable que R1 si restringuimos la respuesta a las dos categorías más favorables.

M2.-MODELO LOGÍT CON NIVEL BASE RESPUESTA LA CATEGORÍA MÁS IMPORTANTE

Este tipo de modelo considera la categoría más favorable y la compara con las restantes categorías, dando lugar al concepto

de razones de continuacion (cambio relativo de una categoría particular contra la más favorable). El logit empleado es:

$$Lij=Ln\ (\ Pij\ /\ Pi4\ )\qquad i=1,2,3,4;\quad j=1,2,3$$

y el modelo correspondiente es:

$$Lij = \beta j + \theta i\ \delta j\qquad \text{con}\qquad \theta_4=0\ ,\ \delta_3=1$$

donde el orden por la magnitud de los valores de las θi expresan diferencias entre drogas. Cuando $\delta j>0$ para toda j, valores pequeños de θi corresponden a logits pequeños, lo cual significa una mayor efectividad en la disminución de dolor para la droga i; βj especifica la frecuencia de la respuesta en las diversas categorías, las cuales son independientes del efecto de la droga; Lij es una función lineal de los parámetros para las drogas en cada categoría de la respuesta.

1.- En el análisis al gráficar Lij vs j=1,2,3 (Gráfica C) se observa que R1 y R2 vs R3 y R4 se diferencian perfectamente, además R1 y R2 se diferencian mejor por la categoría C3(buena disminución de dolor:L13,L23) mientras R3 y R4 con la categoría C2(disminución de dolor moderado).

2.- Estimación del modelo completo y elaboración de la gráfica D ($\hat{L}ij$ Vs θj) para este modelo. Esta gráfica junto con los parámetros estimados permite proponer modelos más simples para la descripción de los datos. Así, en este caso, los valores estimados de θ en el modelo completo suguieren igualar θ_1 con θ_2 y θ_3 con θ_4 por otro lado la gráfica suguiere agrupamientos de categorías de la respuesta, en particular puede observarse que C1 y C3 casi se sobreponen. Ello indica la reconsideración del modelo tomando en cuenta $\beta 1=\beta 3$ y $\delta 1=\delta 3=1$. Aunque estas restricciones no estan relacionadas directamente con las diferencias entre las drogas, permiten simplificar la descripción de los datos (véase Tabla 3).

3.- Proporciones estimadas (Tablas 4 y 5).- Con el mejor modelo ajustado (modelo final:Tabla 3) se estiman las proporciones en la respuesta para cada droga; las proporciones estimadas muestran una buena aproximación con la tabla total y los parámetros son todos significativos, al juzgar por los errores estandard respectivos. En base a las proporciones estimadas podemos decir

MODELO LOGIT CONSIDERANDO COMO CATEGORIA DE COMPARACION LA MAS IMPORTANTE

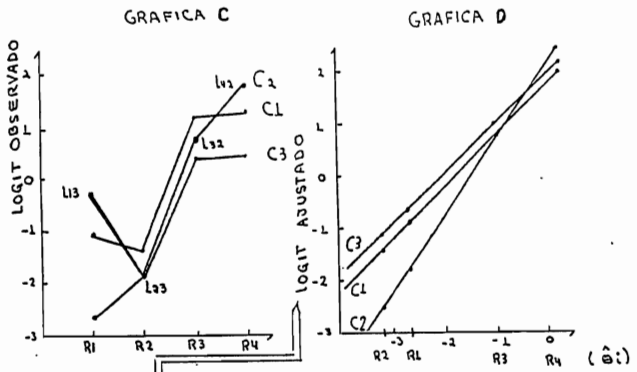


TABLA 3 RESUMEN DE AJUSTES SUCESIVOS PARA EL MODELO LOGIT CON CATEGORIA BASE DE COMPARACION.

G^2	GL	DESCRIPCION DEL MODELO
6.69	4	modelo completo. $\theta 1=-2.66, \hat{\theta 2}=-3.17$ $\theta 3=-0.93, \theta 4=0$
8.37	5	$\theta 3=\theta 4=0$
9.58	6	$\theta 1=\theta 2, \theta 3=\theta 4=0$
10.22	8	$\theta 1=\theta 2, \theta 3=\theta 4=0$, $\delta 1=\delta 3=1, \beta 1=\beta 3$ mod final.
46.74		homogeneidad.

TABLA 4 ESTIMADORES MV DE LOS PARAMETROS PARA EL MODELO FINAL.

PARAM	EST	ERROR EST
$\beta 1(\beta 3)$	1.308	0.476
$\beta 2$	1.281	0.506
$\theta 1(\theta 2)$	-2.392	0.548
$\delta 2$	1.430	0.293

TABLA 5 ESTIMADORES MV PARA LAS PROPORCIONES USANDO EL MODELO FINAL.

	C1	C2	C3	C4	PROM*
R1	0.189	0.066	0.189	0.557	3.12
R2	0.189	0.066	0.189	0.557	3.12
R3	0.308	0.300	0.308	0.083	2.17
R4	0.308	0.300	0.308	0.083	2.17

* de la distribución de la respuesta estimada usando los valores 1,2,3,4 para las cuatro categorías.

que las diferencias entre los dos grupos de drogas ocurre principalmente en C2(dism moderada) y C4(dism muy buena a excelente). Esto explica parcialmente la falta de significancia de la subtabla 1, pues la diferencia con respecto a esas dos categorías se cancela cuando sumamos las tres categorías más favorables ya que las diferencias en C2 y C4 difieren pero en dirección contraria.

Es importante anotar que en el análisis con modelos logísticos politómicos es usual utilizar varios de ellos con el propósito de

complementar las conclusiones en el estudio de la relación de los factores de exposición y la respuesta.

4. Comentario final

Si bien es cierto que con los modelos politómicos de respuesta ordenada presentados (M1 y M2) se ilustró el caso simple correspondiente al estudio de la relación migración-salud mental cuando no esta confundida por diferencias poblacionales, tambien lo es, que en la investigación que se presenta para estudiar dicha relación se requiere de una modelación donde las diferencias poblacionales se espera que ocurran por la presencia de factores de confusión potenciales (otros eventos estresantes=X_2, rasgos de personalidad=X_3, redes sociales de apoyo=X_4 y sexo=X_5) que modifican la relación migración-salud mental.

Las propuestas de modelos para el análisis de datos politómicos abundan (Cox and Chuang, 1984 ; Cox, 1988 ; McCullagh, 1980 ; Ruíz Velasco, 1984) entre ellos se encuentran los que consideran uno o dos factores en estudio y además posibles factores de confusión. Sin embargo dichos modelos no se han popularizado todavia debido a las dificultades que se tienen para su ajuste, que requiere de "software" más especializado, aunado a que la manera de interpretar los parámetros de los modelos no es directa. Así, por ejemplo, para el modelo M2 ejemplificado se utiliza el paquete de cómputo estadístico BMDP (Biomedical Statistical Package) en su módulo de regresión no lineal, el cual requiere además de una subrutina que en cada iteración calcula las probabilidades multinomiales, las derivadas $\partial\Pi_J/\partial\theta$ y los pesos $1/\Pi_J$ correspondientes a la inversa generalizada de la matriz de covarianza de la multinomial (Cox,C ; 1985). En cambio en el paquete GLIM (Generalized Linear Interactive Medelling) para analizar el mismo modelo se requiere del empleo de una función de liga compuesta (Thompson and Baker, 1981).

Referencias

(1) Cox, C and Chuang, C. "A comparison of chi-square partitioning, and two logit analyses of ordinal pain data from a pharmaceutical study". Statistics in Medicine,Vol 3.273-285,1984.

(2) Cox, C. "Computa t ion of maximum likelihood estimates by interactively reweighted least squares :a spectrum of examples . BMDP Statistical Software, Technical Report No 82,1985.

(3) Cox, C. "Multinomial regression models based on continuation ratios". Statistics in Medicine,Vol 7,435-441,1988.
(4) McCullagh, P. " Regression models for ordinal data.Journal of the Royal Statistical Society. Series B.42,109-127,1980.
(5) Ducoing, A. "Descripción de diversos Métodos de Análisis Estadístico para los estudios de Casos y Controles ", Tesis Maestro en Ciencias, IIMAS-UNAM, 1988.
(6) Ruiz Velasco, S "Análisis y Modelos para datos Politómicos y Categóricos Multivariados", Tesis de Maestro en Ciencias, IIMAS-UNAM,1984.
(7) Thompson, R and Baker, R.J. "Composite Link funtions generalized linear models". Applied Statistics.30.125-131,1981.
(8) Oliveira, O y García, G . "Migración a grandes ciudades del tercer mundo:algunas implicaciones sociodemográficas" .Estudios sociológicos II:1,1984.
(9) Unikel, L."El desarrollo urbano de México". Colegio de México, 1976.
(10) Arizpe, L."Migración etnicismo y cambio económico". Colegio de México, 1978.
(11) Ruiz de Chavez, L . "Marginalidad y conducta antisocial en menores".Cuadernos del Instituto Nacional de Ciencias Penales 1. México,1978.
(12) Muñoz, H; Oliveira, O; Singer, P y Stern, C . "Migración y desarrollo.Consideraciones Teóricas". Consejo Latinoamericano de Ciencias Sociales. Serie Población. 1972.
(13) Cuevas, C, S. "El estudio de la salud mental en relación con los procesos migratorios. Esbozo de un modelo".Salud Mental 12:1. México, 1989.
(14) Cuevas, C, S. "Trastornos mentales en una comunidad de bajo nivel socioeconómico". En prensa : Salud Mental,1990.
(15) Burnam, M, A; Hough, R, L; Karno, M. "Acculturation and lifetime prevalence in psychiatric disorders among Mexican-Americans in los Angeles". Journal of Health and Social Behavior 28, 1987.
(16) Karno, M ; Hough, R, L ;Burnam, A, M."Lifetime prevalences of specific psychiatric disorders among Mexican-Americans and non-hispanic whites in los Angeles". Archives General Psychiatric 1987:44.PP 695-761.
(17) Burnam, A, M; Hough, R, L; Escobar, J . "Six-month prevalence of specific psychiatric disorders among Mexican-Americans and non-hispanic whites in los Angeles". Archives General Psychiatric 1987:44,12.PP 727-735.
(18) Canino, G, J; Bird, H, R; Shrout, P, E . "The prevalence of specific psychiatric disorders in Puertorico". Archives General Psychiatric:1987:44 .PP 687-694.
(19) American Psychiatric Association: Diagnostical and Statistical manual of Mental Disorders. Tercera Ed.Washington, D.C,1980.
(20) Méndez, R, I; Nahimira, D; Moreno, L y Sosa, C." El protocolo de Investigación: Lineamientos para su elaboración y Análisis". Trillas. México, 1984.
(21) Gonzáles, F, C."Estudio de la validez del DIS". Manuscrito no publicado. Instituto Mexicano de Psiquitría, 1982.
(22) Eysenck, S, VG y Lara, C, MA. "Un estudio transcultural de la perzonalidad en adultos Mexicanos e Ingleses". Salud Mental 12:3. México,1989.

LA ECONOMIA MEXICANA EN EL PERIODO 1939 - 1979.

-una aplicación de los métodos multivariados-

Hernando Enrique Mutis Gaitán.

I.I.M.A.S. - U.N.A.M.

RESUMEN

Utilizando el análisis de componentes principales con el enfoque de la escuela francesa del análisis de datos, se aborda el comportamiento de tres grupos de variables de la economía mexicana entre 1939 y 1979. El centro de este estudio es interpretar los resultados de la aplicacación mencionada.

PRESENTACION.

El trabajo que se presenta forma parte de un examen más amplio sobre la economía mexicana. El período de investigación en sus inicios arrancaba con la década de los veinte. No obstante, por problemas de información -física inexistencia en algunos casos, poca fidelidad en otros e insuficiencia en las más- se corrió el estudio aproximadamente 20 años. De esta manera el período cobijado comienza, años más, años menos, en los cuarentas y pretende cubrir hasta el decenio de los ochenta. Para este IV Foro de Estadística se esbozan los principales resultados encontrados en el lapso entre 1939 y 1979, enfatizando en la etapa que va de 1940 a 1960.

OBJETIVO.

El interés es buscar ciclos, tendencias y características de la economía nacional, utilizando para ello las técnicas estadísticas multivariadas. Este objetivo central se parte en un doble aspecto: por un lado, se busca dar respaldo empírico a planteamientos teóricos y, por otro lado, se busca escudriñar explicaciones teóricas sugeridas por los datos mismos. En los dos casos, el espíritu que anima esta investigación es la confrontación a través de lo "real", entendiendo lo "real" como la particular apreciación de la dinámica económica sintetizada por la matemática del álgebra lineal.

EL PROCEDIMIENTO ESTADISTICO.

El párrafo anterior se aclara cuando precisamos que el procedimiento utilizado forma parte del enfoque del Análisis de Datos. Entre los los diversos aspectos que caracterizan a este enfoque, quiero resaltar dos de ellos:

El primero, se refiere al hecho de que en el análisis se privilegian los procedimientos algebraicos y numéricos de la información antes que a su modelado. Desde este punto de vista no se presume ningún modelo subyacente explicativo de la información, al menos en una primera etapa. Esto significa, por consiguiente, que no es necesario definir supuestos especiales -salvo la linealidad- sobre el comportamiento de los datos y al mismo tiempo, no se tiene el problema de la estimación de parámetros. Así, el propósito es el destacar las características básicas de la información y no de expresarla con un modelo específico. La famosa frase de Benzécri sintetiza el espíritu de la idea en que se quiere insistir: "El modelo debe ajustarse a los datos y no al contrario."

El segundo aspecto se refiere a que la información se concibe como un conjunto de puntos en el espacio multidimensional e interesa una específica representación geométrica de esa información. En términos de Greenacre: "... es una técnica geométrica, antes que estadística". A pesar de esta afirmación no se debe soslayar que los procedimientos algebraicos y numéricos recuperan o, expresándolo mejor, reconstruyen resultados de la estadística clásica, sólo que a través de un camino diferente y, más aún, con un enfoque también diferente. Además, los procedimientos que se invocan son mucho más versátiles que los obtenidos por la estadística clásica. [1]

Sobre el álgebra de la técnica, remito a los autores que se citan en la nota 1. Unicamente se insistirá en que el método

---

1. Se pueden consultar los siguientes autores: Benzécri (1973), L'Analyse des Données, Dunod, Paris; Greenacre (1984), Theory and Aplications of Correspondence Analysis, Academic Press, London; Barnett -ed- (1981), Interpretting Multivariate Data, Wiley, Bath; Lebart et al. (1984), Multivariate Descriptive Statistical Analysis, Wiley, New York.

busca llevar la información original a un subespacio tal que disminuyendo la dimensionalidad recoja la mayor inercia posible de los datos iniciales. La pérdida relativa -susceptible de medir- de ciertos aspectos de la información se recupera con su mayor simplicidad. Para el caso que nos ocupa se utilizó el análisis de componentes principales, repito, con un enfoque diferente al clásico y sin necesidad de acudir a sus supuestos. De suyo se comprende, y tal vez se puede argumentar que, como limitación del enfoque, no son válidas las inferencias sobre la información, lo cual sí es posible en el enfoque clásico.

LAS VARIABLES.

En el estudio se consideraron 18 variables en tres grupos: uno, referido a la producción y a la productividad, otro, relacionado con ingreso y gasto y un tercero con variables que tienen que ver con el comercio exterior. Todas las variables están medidas en precios constantes de 1960 y cubren el período entre 1939 - 1979:

PIIN : productividad del sector industrial.
PIAG : productividad agrícola.
PINA : productividad no agrícola.
PIB : producto interno bruto.
MAN : PIB manufacturero.
PETR : PIB del sector petróleo.
CONS : PIB del sector construcción.
ELEC : PIB del sector electricidad.
IBFT : inversión bruta fija total.
PEA : población económicamente activa.
SIP : salario industrial promedio.
SEPI : proporción de las remuneraciones en el PIB.
PSS : relación productividad-salarios del sector industrial.
GAST : gasto social del gobierno.
PIME : producción interna de maquinaria y equipo.
MXME : importaciones de maquinaria y equipo.
EXPO : exportaciones.
IMPO : importaciones.

Utilizando las rutinas FACP, IDON y GRVV del paquete "chadoc", se seleccionaron los tres primeros componentes. Los resultados y su análisis se presentan a continuación.

LOS RESULTADOS Y EL ANALISIS:

En el cuadro No. 1, al final de este artículo, se encuentra la matriz de correlaciones. Se destaca la férrea estructura existente en el conjunto. Las variables menos correlacionadas se refieren, en su orden, a la relación productividad-salarios de la industria y al gasto social del gobierno. Estas dos variables están asociadas con el nivel de bienestar económico general. Con respecto a la primera, es factible afirmar que nos indica en qué medida el crecimiento económico favorece bien a los sectores asalariados, bien a los empresarios. La variable gasto social está construida de tal forma que nos ofrezca una pista sobre el proceso de redistribución de ingresos hacia los grupos de la población menos favorecidos.

Del cuadro No. 2 se observa que los tres primeros componentes explican el 97 % del conjunto analizado. Nótese de manera particular el gran peso explicativo que tiene el primer eje (87.5%). Del cuadro no. 3 se destacan las altas correlaciones negativas de las variables con el primer eje, exceptuando la relación productividad-salarios que se encuentra negativa y fuertemente correlacionada con el segundo componente. Por estas razones y por las magnitudes de los valores de los coeficientes de las variables en los vectores propios -véase cuadro no.4-, se "bautizó" el primer componente como un índice del crecimiento económico. El segundo eje, por razones que se explicarán cuando se analicen los cuadros nos. 8 y 9, se interpretó como una medida del bienestar social. Se quiere hacer hincapié en que el significado del primer eje se restringe al crecimiento en términos exclusivamente cuantitativos, es decir, se está diferenciando crecimiento de desarrollo.

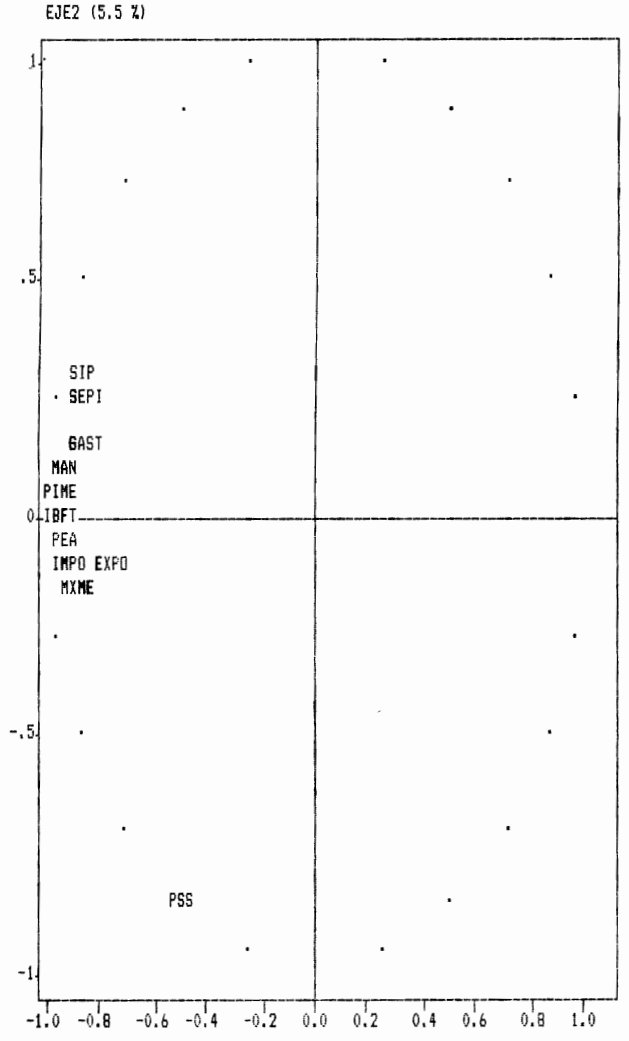
El segundo componente -bienestar- intenta cuantificar el papel de los salarios en el panorama nacional. Así, dada la ortogonalidad de los ejes, habría una independencia entre bienestar y crecimiento, al menos en cuanto a esta interpretación de la economía mexicana. Se entiende la dinámica de las variables que tienen que ver con la distribución de ingresos, como un índice de bienestar económico. En los cuadros nos. 7 y 8 se remarca el mayor peso específico que tienen estas variables con el segundo componente, disminuyendo clara y contundentemente su importancia con el primer eje.

En torno a la validez de la representación de las variables y los años con respecto a los componentes, nos atenemos al criterio del cos2 y de la calidad global de la representación

Gráfica No. 1

México: 1939 - 1979.

REPRESENTACION DE LAS VARIABLES CON RESPECTO A LOS NUEVOS EJES.



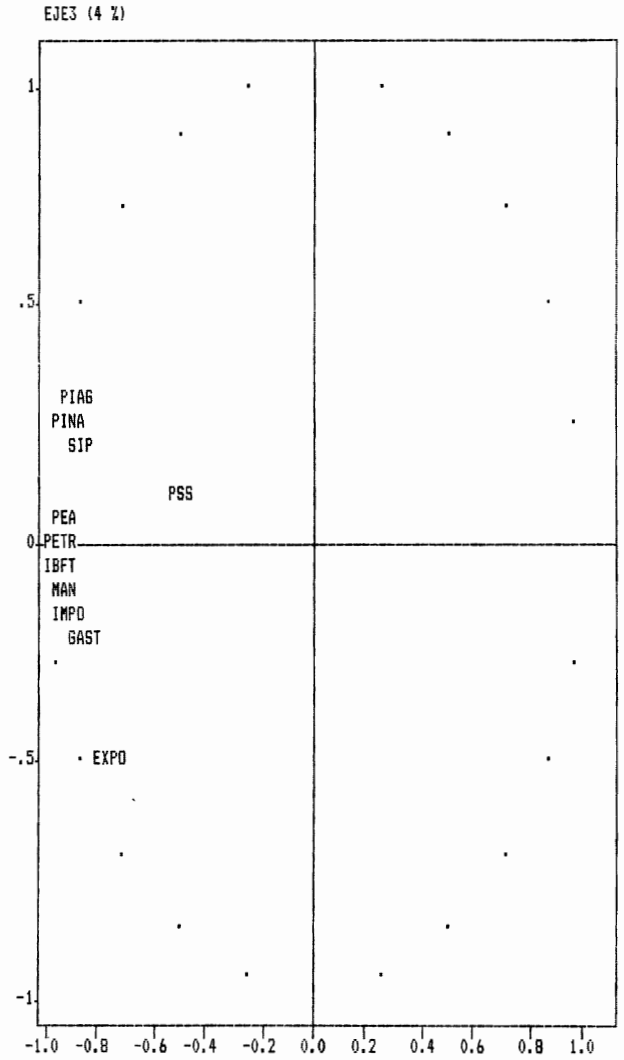
Puntos superpuestos:

PIIN: -0.98, -0.04
PIAG: -0.94, -0.06
PINA: -0.96, -0.01
PIB : -0.98, 0.06
PETR: -0.99, 0.01
CONS: -0.97, 0.08
ELEC -0.98, 0.08

Gráfica No. 2

México: 1939 - 1979.

REPRESENTACION DE LAS VARIABLES CON RESPECTO A LOS NUEVOS EJES.



Puntos superpuestos:

SEPI: -0.90, 0.22
PIIN: -0.98, 0.21
CONS: -0.97, -0.12
PIB : -0.98, -0.08
ELEC: -0.98, -0.12
PIME: -0.99, -0.11
MXME: -0.94, 0.19

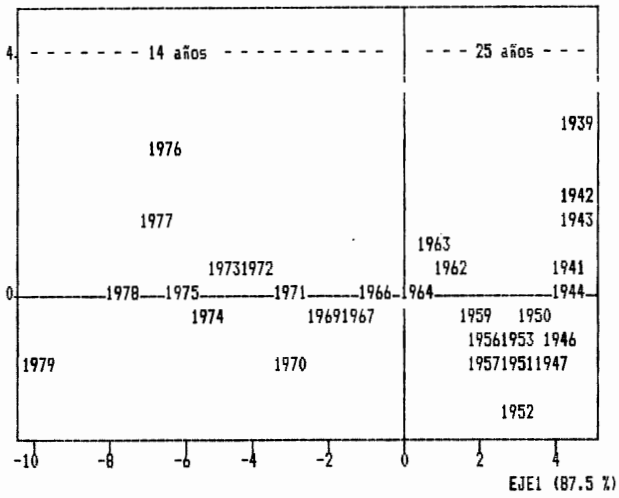
(columnas COS1..COS3 y QLT3) los cuales nos dicen qué tan bien están graficados los diferentes puntos (variables y años). Hay que resaltar que tales indicadores nos informan que, en general, la calidad de la representación es bastante alta. Habría que excluir de la última afirmación el año de 1964, con una representación muy pobre cuando se considera el lapso 1939-1979 y, en cambio, notablemente mejorada cuando nos referimos al período 39 - 65. Los períodos se comportan casi al contrario al determinar gráficamente el año 1954. La representación global es, repito, sumamente buena.

En la gráfica no. 1 se encuentra la posición de cada una de las variables con respecto a los dos primeros componentes principales: el eje 1 y el eje 2, respectivamente. Entre paréntesis en la gráfica se incluye la proporción de la inercia que le corresponde a cada uno de ellos. Entre los dos acumulan el 93 % de la inercia total.

Así mismo, en la gráfica no. 1 se encuentran expresados los vectores asociados a las variables respecto a los ejes principales: forman un haz muy cerrado, lo cual no es sino expresión de la fuerte estructura de correlaciones. No obstante, las variables asociadas con la distribución del ingreso tienden a tener un comportamiento ligeramente diferente -SIP, SEPI,GASTO, PSS- al resto, destacándose la posición de la relación productividad-salarios, esta última abiertamente separada del conjunto y vinculada de manera directa con el segundo componente. Tal diferenciación es muchísimo más marcada al hacer el análisis para el subperíodo 1939 - 1965, como se puede confirmar en la gráfica no. 5, en la cual, adicionalmente, las variables referidas al comercio exterior van a su vez a diferenciarse del racimo inicial, tendiendo a comportarse como un grupo con características propias.

Se subraya que en el lapso 1939-1965 las variables que se refieren a ingresos, como son el salario industrial promedio (SIP),la proporción de remuneraciones en el PIB (SEPI) y la relación productividad sobre salarios (PSS), se encuentran vinculadas con el segundo componente, además, con la particularidad de que la PSS está negativamente correlacionada con las dos anteriores y su conducta es indicativa precisamente de la relación inversa que guarda con las anteriores: los aumentos de la productividad del trabajo no se reflejan en el peso que tiene el salario en el concierto económico. El viejo enunciado de la teoría económica de que la remuneración del factor trabajo debe concordar con su productividad tiene para el caso mexicano un nítido contraejemplo: el incremento de la productividad del

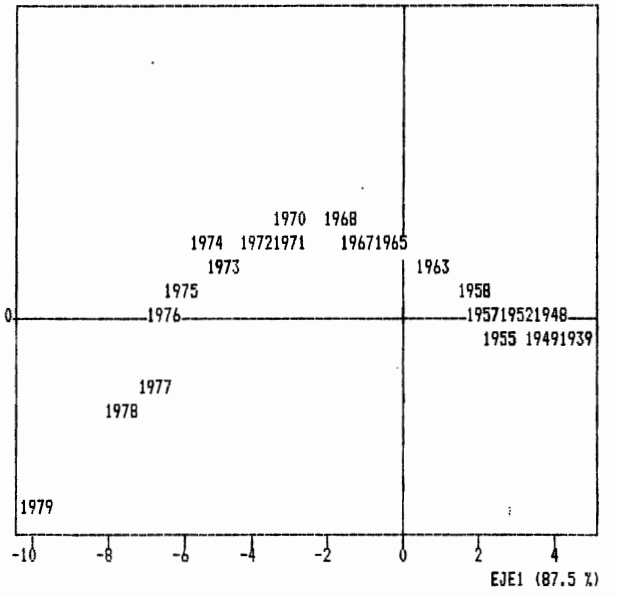
Gráfica No. 3
México: 1939 - 1979.
REPRESENTACION DE LOS AÑOS CON RESPECTO A LOS NUEVOS EJES.
EJE2 (5.5 %)



Puntos superpuestos:

1940: 4.21, 2.63
1945: 3.99, 0.19
1948: 3.45, -0.86
1949: 3.36, -0.03
1954: 2.39, -0.63
1955: 2.09, -0.82
1958: 1.55, -0.54
1960: 1.11, -0.30
1961: 1.04, -0.38
1965: -0.70, 0.00
1968: -2.19, -0.25

Gráfica No. 4
México: 1939 - 1979.
REPRESENTACION DE LOS AÑOS CON RESPECTO A LOS NUEVOS EJES.
EJE3 (4 %)



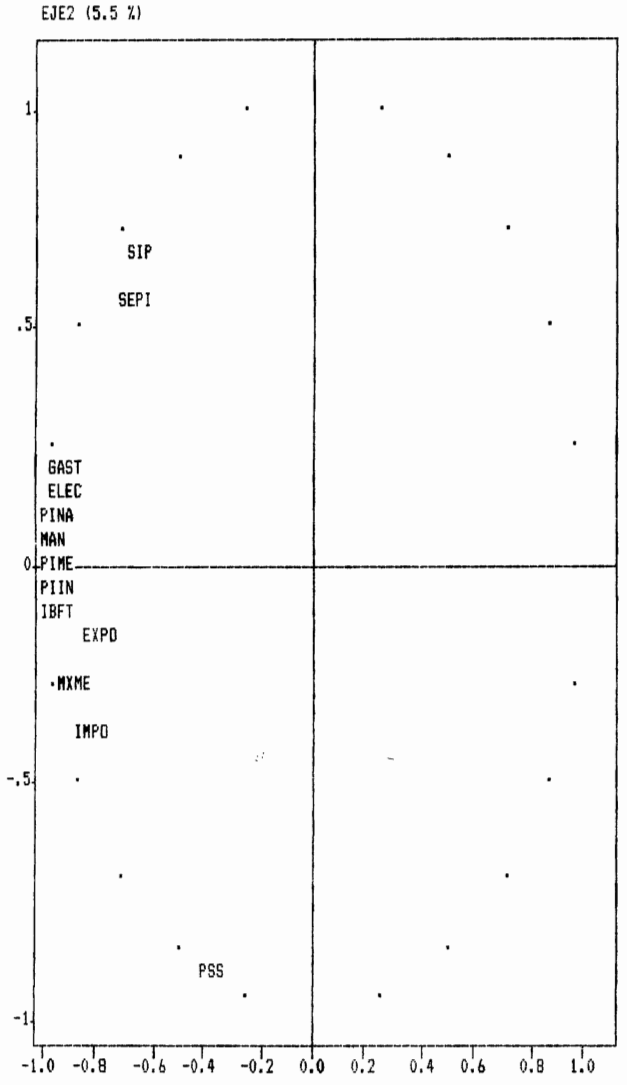
Puntos superpuestos:

1940: 4.21, -0.38
1941: 4.11, -0.36
1942: 4.20, -0.42
1943: 4.20, -0.42
1944: 4.09, -0.37
1945: 3.99, -0.35
1946: 3.76, -0.40
1947: 3.54, -0.35
1950: 3.02, -0.43
1951: 2.74, -0.43
1953: 2.74, -0.31
1954: 2.39, -0.32
1956: 1.77, -0.37
1959: 1.51, 0.10
1960: 1.11, 0.28
1961: 1.04, 0.22
1962: 0.79, 0.34
1964: -0.07, 0.50
1966: -1.08, 1.00
1969: -2.62, 1.25

Gráfica No. 5

México: 1939 - 1965.

REPRESENTACION DE LAS VARIABLES CON RESPECTO A LOS NUEVOS EJES.



Puntos superpuestos:

PEA : -0.97, -0.13
PIAG: -0.98, -0.08
PETR: -0.99, 0.09
CONS: -0.99, -0.04
PIB : -1.00, 0.03

no va acompañado de su correspondiente aumento salarial, situación que, desde luego, favorece a las ganancias empresariales.

En la gráfica no. 2 están presentes las variables contra los componentes primero y tercero. En lo fundamental, se resalta la diferente posición de las variables exportaciones (EXPO) y la relación productividad a salarios (PSS). A pesar de la pequeña inercia concentrada en el eje no. 3, el conjunto representa el 91.5 % de la variabilidad.

La gráfica no. 3 expresa la situación de los períodos anuales en referencia a los dos primeros componentes. Recordemos que el primer eje lo habíamos interpretado como crecimiento económico.

El segundo componente, lo traducíamos en términos del bienestar económico, debido al importante papel desempeñado por las variables relacionadas con el salario. Respecto al primer componente, se observa que el crecimiento económico aumenta hacia la izquierda -hay correlaciones negativas entre las variables y el primer eje, como se muestra en los cuadros nos. 3 y 9-; en cuanto al componente 2, las correlaciones de las variables que pesan en este eje con él mismo son positivas, exceptuando a la variable PSS, pero por la forma como se ha interpretado a esta última, correlaciones negativas indican pérdida relativa del salario con respecto a la productividad, luego es factible explicarnos que, en términos del salario, la relación es directa.

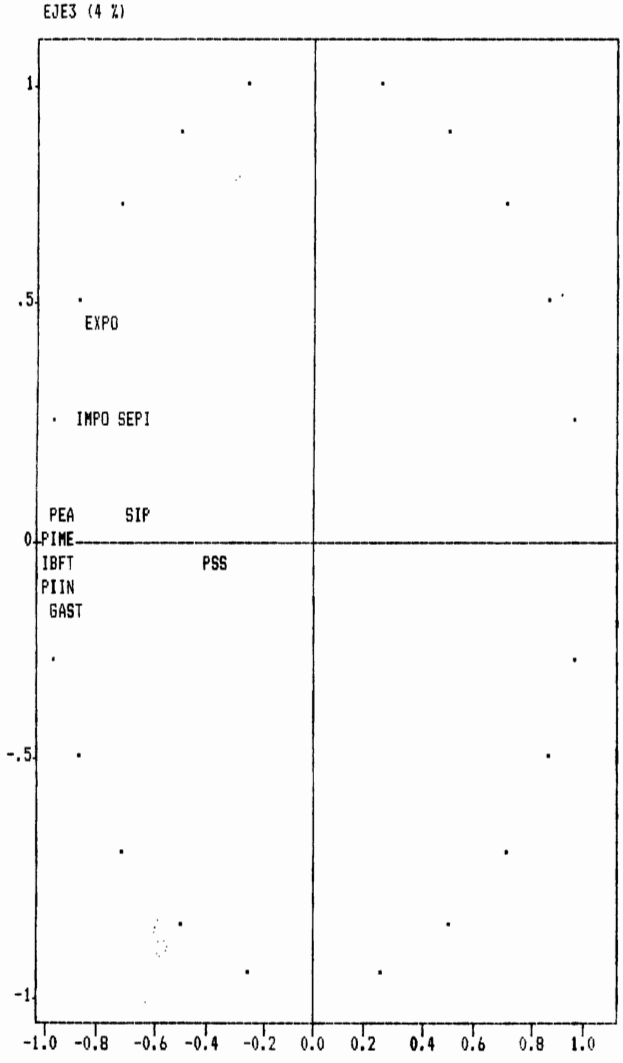
Esta gráfica no. 3 sugirió la necesidad de estudiar el comportamiento del lapso de 25 años -arriba en el cuadro de la derecha- para discernir de una manera más precisa qué estaba sucediendo en su interior. Esto dio lugar a la gráfica no. 7 en la cual -como si se hubiese tomado una ampliación- se ilustra con mayor nitidez la senda del crecimiento económico en el mencionado lapso.

Concentremos ahora nuestra atención sobre el primer eje. La proyección de cada uno de los años sobre el eje no. 1 nos indica la posición de ese año con respecto al crecimiento. Las distancias horizontales de las mencionadas proyecciones nos miden la intensidad del crecimiento. De esta manera estamos obteniendo una medida del crecimiento que aprovecha la información de las 18 variables: estamos cuantificando la velocidad del crecimiento sobre un solo eje, es decir, en una sola dimensión concentramos la información de esas 18 variables. Lo que de por sí ya es una ventaja puesto que se

Gráfica No. 6

México: 1939 - 1965.

REPRESENTACION DE LAS VARIABLES CON RESPECTO A LOS NUEVOS EJES.



Puntos superpuestos:

PIAG: -0.98, -0.05
PINA: -0.99, -0.05
MAN : -0.99, -0.08
PETR: -0.99, -0.05
CONS: -0.99, -0.07
ELEC: -0.98, -0.12
MXME: -0.94, -0.10
PIB : -1.00, -0.05

está dilucidando el crecimiento no por una sola variable -el PIB, por ejemplo-, sino que estamos sintetizando la variabilidad de todo el grupo de variables -y no, repito, de una sola de ellas- en un único eje. Además, se está obteniendo una medida del crecimiento un tanto distinta a la que estamos acostumbrados en los análisis económicos, pero es una medida que tiene un gran potencial de información.

Un primer vistazo de la gráfica insinuaría una subdivisión del lapso considerado en dos períodos: el primero, entre el 39 y el 65; el segundo, desde 1965 hasta 1979. No obstante si fijamos la atención en lo que hemos denominado la velocidad del crecimiento, la periodización es distinta: un primer lapso que empieza en el año de 1939 y culmina en 1959-1960; el siguiente arranca desde el 60.Uno de los criterios de periodización se fundamenta en las proyecciones de los años sobre el primer componente: se define un ciclo que comienza en el 40 y termina en el 48, alcanzando un máximo en el año de 1946; a continuación se recupera en los dos años siguientes pero cae drásticamente en 1953, año en que se observa la mayor sima del lapso considerado. Después se presenta un ciclo muy suave que finaliza en 1959. Prosiguiendo con la trayectoria, los sesentas inauguran un nuevo período de crecimiento, con características diferentes a los años iniciales, que cae un tanto en 1966, vuelve a hacerlo tres años después, pero de una manera más marcada, repite en el 71 y posteriormente en 1976-1977, para experimentar, cuando menos hasta 1979, un aumento considerable. Es posible concluir que se cuentan seis ciclos agrupados en dos subperíodos o etapas: 1940 - 1959 y 1960 - 1977; estas etapas están directamente asociadas a las fases de la sustitución de importaciones que surgen en los cuarenta y se refuerzan en los sesenta.

Analicemos los cambios que se presentan entre los dos períodos involucrados; el criterio que se usa es el tipo de variables asociadas con los ejes 2 y 3. Para los años 40-60, las variables que nos explican los derroteros de las gráficas 3 y 7 (ateniéndonos al segundo componente) son aquellas relacionadas con el salario. En cambio, del 60 en adelante, el papel anterior lo desempeñan las variables relacionadas con las productividades -PIAG y PIIN- y con el comercio exterior -EXPO y MXME-. Esta diferenciación corresponde claramente a una distinta caracterización de los dos períodos analizados, en concreto, el segundo período corresponde al

denominado "desarrollo estabilizador". Si se centra la mirada en los elementos constitutivos del tercer eje, se descubre que hay una suerte de trastocamiento en lo que corresponde a los dos períodos: aquellas variables tildadas de cruciales en el segundo eje se convierten en las decisivas en el tercero y viceversa. Visto históricamente, significa que el pasado avizora su futuro y este último no olvida sus orígenes.

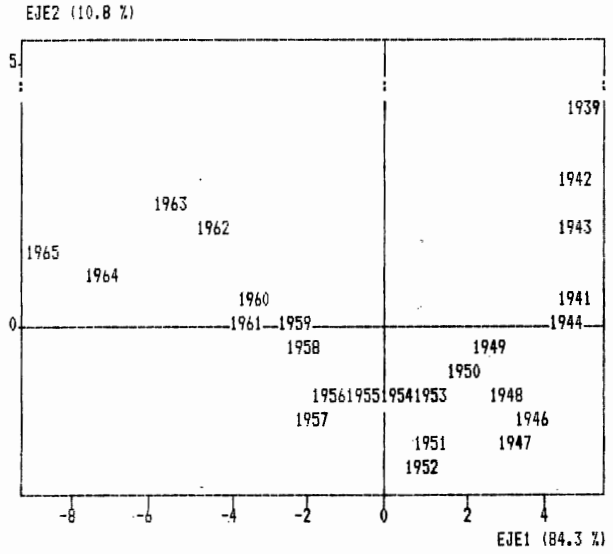
Se ha planteado que la economía mexicana tiene una suerte de asociación con los regímenes gubernamentales. De los resultados de nuestro estudio, la relación de los ciclos económicos inducidos por los ciclos políticos no es tan diáfana, podemos acercarnos a una visualización de ello si graficamos las distancias horizontales de las proyecciones de los años con respecto a cada uno de los períodos presidenciales. De todos modos las gráficas nos. 3 y 7 no son tan transparentes para este propósito, especialmente porque ciertos puntos superpuestos oscurecen dicha interpretación.

En la gráfica No. 4 se ve el llamado "efecto de herradura" o de "media luna", sintomático de la existencia de un cierto orden en la información. Aquí el orden tiene que ver con el crecimiento económico, el cual aumenta con mayor rapidez a medida que en la curva hay desplazamiento hacia la izquierda. De los 41 puntos, los 25 primeros están aglutinados en la tercera parte -a la derecha- de la gráfica. En la gráfica no. 7 se descubre también el efecto de "media luna" si nos detenemos en el año de 1960. A partir de este óltimo el efecto se voltea: es como si este año -y su inmediato alrededor- se convirtiese en un punto de inflexión, denotando el cambio sustantivo en el trayecto del crecimiento. Hay, entonces, una buena correspon- dencia entre las dos grandes etapas ya señaladas. Obsérvese que desde el 39 hasta ya entrada la década de los cincuenta se experimenta una fuerte caída de lo que hemos denominado "bienestar económico", es decir, de las variables asociadas con el salario. El decenio de los 40 marca expresamente una sostenida y aguda reducción del poder de compra de los salarios reales, situación que sólo se mejora en la segunda parte de los cincuenta y presenta un crecimiento suavizado -y con altibajos- que alcanza su máximo en 1963. De la gráfica no. 3 se nota que a partir de este año disminuye el "bienestar" de una manera muy consistente hasta el mínimo relativo de 1970. Después, y de nuevo con altibajos, hay una

recuperación que alcanza su cima en 1976. Por esta trayectoria de las variables relacionadas con el salario se han interpretado las gráficas nos. 3 y 7, como los "mapas" del crecimiento económico financiado por el salario: nos bosquejan el crecimiento a costa de las restricciones salariales.

Detengamos la atención en la gráfica no. 8. El eje 3 está íntimamente vinculado al comportamiento del comercio exterior, conclusión que se obtiene al ojear el cuadro no. 8. en el que las exportaciones y las importaciones constituyen sus elementos más connotados, lo cual se corrobora parcialmente en la gráfica no. 2 y en la representación visual que se hiciese del eje 1 contra el eje 3 -que por razones de espacio no se incluye en esta síntesis, aunque sí se presentó en la exposición oral- con respecto a las variables. De esta manera se nos va configurando otro "mapa", en este caso el del déficit externo en relación al crecimiento -de hecho, se puede trasladar esta afirmación a la gráfica no. 4-. De esta forma, se han reconstruido, para las gráficas mencionadas en los dos últimos párrafos, distintas rutas o trayectorias de crecimiento en función bien de los salarios o del bienestar económico (gráfcs. 3 y 7), bien del déficit externo (gráfcs. 4 y 8); gracias a este enfoque gráfico multivariado se ha alcanzado una gran capacidad de síntesis que ha permitido concentrar -en relativamente pocas gráficas- el camino recorrido por la economía mexicana al considerar ese específico grupo de variables en el transcurso de ocho lustros.

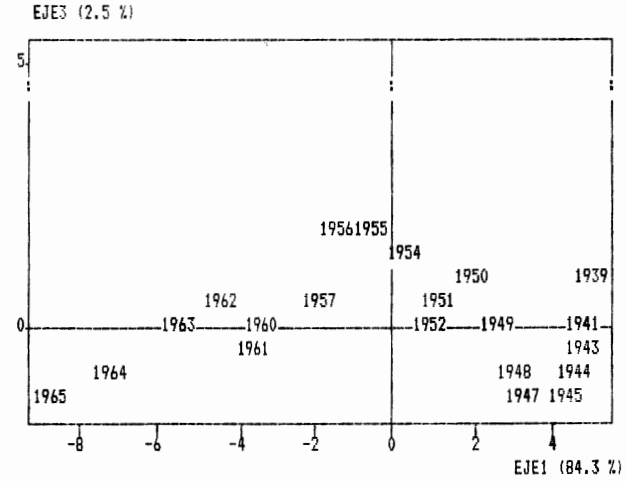
Gráfica No. 7
México: 1939 - 1965.
REPRESENTACION DE LOS AÑOS CON RESPECTO A LOS NUEVOS EJES.



Puntos superpustos:

1940: 4.41, 3.08
1945: 3.84, -0.07

Gráfica No. 8
México: 1939 - 1965.
REPRESENTACION DE LOS AÑOS CON RESPECTO A LOS NUEVOS EJES.



Puntos superpuestos:

1940: 4.41, 0.60
1942: 4.36, 0.03
1946: 3.30, -1.09
1953: 0.79, 0.23
1958: -2.38, 0.40
1959:- 2.61, 0.19

CUADRO No. 1

México: 1939 - 1979.

Matriz de correlaciones.

	IBFT	SIP	SEPI	GAST	PEA	PSS	PIIN	PIAG	PINA
IBFT	1.000	0.902	0.874	0.920	0.964	0.517	0.958	0.916	0.939
SIP	0.902	1.000	0.942	0.841	0.886	0.247	0.930	0.915	0.932
SEPI	0.874	0.942	1.000	0.814	0.892	0.263	0.903	0.893	0.926
GAST	0.920	0.841	0.814	1.000	0.883	0.359	0.843	0.797	0.818
PEA	0.964	0.886	0.892	0.883	1.000	0.573	0.961	0.941	0.957
PSS	0.517	0.247	0.263	0.359	0.573	1.000	0.561	0.570	0.530
PIIN	0.958	0.930	0.903	0.843	0.961	0.561	1.000	0.990	0.993
PIAG	0.916	0.915	0.893	0.797	0.941	0.570	0.990	1.000	0.988
PINA	0.939	0.932	0.926	0.818	0.957	0.530	0.993	0.988	1.000
MAN	0.981	0.892	0.852	0.922	0.929	0.436	0.919	0.869	0.895
PETR	0.992	0.912	0.900	0.916	0.972	0.508	0.967	0.934	0.952
CONS	0.979	0.889	0.846	0.915	0.927	0.441	0.916	0.865	0.893
ELEC	0.992	0.905	0.868	0.936	0.940	0.446	0.935	0.888	0.910
PIME	0.995	0.894	0.866	0.931	0.949	0.485	0.940	0.893	0.916
MXME	0.929	0.850	0.829	0.767	0.917	0.630	0.959	0.941	0.954
EXPO	0.843	0.621	0.644	0.805	0.806	0.470	0.712	0.649	0.681
IMPO	0.968	0.803	0.806	0.867	0.953	0.586	0.899	0.851	0.880
PIB	0.986	0.902	0.869	0.920	0.940	0.467	0.939	0.896	0.917

	MAN	PETR	CONS	ELEC	PIME	MXME	EXPO	IMPO	PIB
IBFT	0.981	0.992	0.979	0.992	0.995	0.929	0.843	0.968	0.986
SIP	0.892	0.912	0.889	0.905	0.894	0.850	0.621	0.803	0.902
SEPI	0.852	0.900	0.846	0.868	0.866	0.829	0.644	0.806	0.869
GAST	0.922	0.916	0.915	0.936	0.931	0.767	0.805	0.867	0.920
PEA	0.929	0.972	0.927	0.940	0.949	0.917	0.806	0.953	0.940
PSS	0.436	0.508	0.441	0.446	0.485	0.630	0.470	0.586	0.467
PIIN	0.919	0.967	0.916	0.935	0.940	0.959	0.712	0.899	0.939
PIAG	0.869	0.934	0.865	0.888	0.893	0.941	0.649	0.851	0.896
PINA	0.895	0.952	0.893	0.910	0.916	0.954	0.681	0.880	0.917
MAN	1.000	0.960	0.999	0.991	0.990	0.867	0.817	0.944	0.994
PETR	0.960	1.000	0.954	0.982	0.984	0.924	0.841	0.953	0.971
CONS	0.999	0.954	1.000	0.987	0.987	0.873	0.807	0.944	0.992
ELEC	0.991	0.982	0.987	1.000	0.998	0.881	0.835	0.950	0.992
PIME	0.990	0.984	0.987	0.998	1.000	0.895	0.844	0.962	0.993
MXME	0.867	0.924	0.873	0.881	0.895	1.000	0.715	0.895	0.885
EXPO	0.817	0.841	0.807	0.835	0.844	0.715	1.000	0.899	0.805
IMPO	0.944	0.953	0.944	0.950	0.962	0.895	0.899	1.000	0.944
PIB	0.994	0.971	0.992	0.992	0.993	0.885	0.805	0.944	1.000

CUADRO No. 2

México: 1939 - 1979.

Valores propios en orden descendente:

Num.	Val. propios	% traza	% Acum.
1	15.75	87.481	87.481
2	0.99	5.477	92.957
3	0.71	3.963	96.920
4	0.21	1.143	98.063
5	0.13	0.745	98.808
6	0.08	0.425	99.233
7	0.05	0.258	99.491

CUADRO No. 3

México: 1939 - 1979.

Correlaciones entre las variables
y los componentes principales:

	EJE1	EJE2	EJE3
IBFT	-0.996	-0.002	-0.065
SIP	-0.919	0.300	0.216
SEPI	-0.902	0.271	0.217
GAST	-0.915	0.155	-0.209
PEA	-0.978	-0.069	0.041
PSS	-0.525	-0.838	0.075
PIIN	-0.975	-0.043	0.206
PIAG	-0.943	-0.064	0.303
PINA	-0.961	-0.014	0.262
MAN	-0.974	0.086	-0.130
PETR	-0.994	0.010	-0.021
CONS	-0.971	0.078	-0.124
ELEC	-0.984	0.080	-0.116
PIME	-0.989	0.035	-0.113
MXME	-0.936	-0.174	0.192
EXPO	-0.822	-0.117	-0.476
IMPO	-0.960	-0.127	-0.181
PIB	-0.983	0.062	-0.079

CUADRO No. 4

México: 1939 - 1979.

Vectores propios (se incluyen sólo los 6 primeros)

	EJE1	EJE2	EJE3	EJE4	EJE5	EJE6
IBFT	-0.251	-0.002	-0.077	0.052	-0.094	0.114
SIP	-0.232	0.302	0.256	0.015	-0.018	0.187
SEPI	-0.227	0.273	0.257	-0.388	0.150	-0.503
GAST	-0.231	0.156	-0.247	0.181	0.751	0.306
PEA	-0.246	-0.070	0.049	-0.183	0.208	-0.313
PSS	-0.132	-0.844	0.089	0.165	0.225	-0.155
PIIN	-0.246	-0.044	0.244	0.006	0.002	0.141
PIAG	-0.238	-0.065	0.359	-0.040	0.120	0.124
PINA	-0.242	-0.014	0.310	-0.108	-0.005	0.003
MAN	-0.245	0.087	-0.154	0.300	-0.161	-0.156
PETR	-0.250	0.010	-0.024	-0.119	0.061	0.118
CONS	-0.245	0.079	-0.147	0.334	-0.221	-0.158
ELEC	-0.248	0.080	-0.137	0.172	-0.051	0.036
PIME	-0.249	0.036	-0.134	0.148	-0.064	-0.024
MXME	-0.236	-0.175	0.227	-0.107	-0.387	0.487
EXPO	-0.207	-0.118	-0.564	-0.603	-0.058	0.201
IMPO	-0.242	-0.128	-0.214	-0.146	-0.219	-0.282
PIB	-0.248	0.062	-0.093	0.274	-0.100	-0.159

CUADRO No. 5

México: 1939 - 1979.

Componentes principales y calidad de la representación.

	EJE1	EJE2	EJE3	COS1	COS2	COS3	QLT3
1939	4.22	2.57	-0.34	71.485	26.419	0.467	98.371
1940	4.21	2.63	-0.38	70.763	27.618	0.573	98.954
1941	4.11	0.38	-0.36	95.039	0.800	0.741	96.580
1942	4.20	1.76	-0.42	84.007	14.745	0.839	99.590
1943	4.20	1.31	-0.42	89.500	8.775	0.913	99.187
1944	4.09	0.25	-0.37	96.999	0.354	0.785	98.138
1945	3.99	0.19	-0.35	95.956	0.211	0.747	96.914
1946	3.76	-0.77	-0.40	90.904	3.825	1.045	95.774
1947	3.54	-1.14	-0.35	86.250	8.946	0.829	96.025
1948	3.45	-0.86	-0.32	90.844	5.604	0.790	97.239
1949	3.36	-0.03	-0.44	96.461	0.010	1.620	98.090
1950	3.02	-0.45	-0.43	95.362	2.127	1.913	99.402
1951	2.74	-1.07	-0.43	83.282	12.744	2.032	98.058
1952	2.61	-1.79	-0.33	65.115	30.741	1.045	96.901
1953	2.74	-0.67	-0.31	91.587	5.458	1.211	98.256
1954	2.39	-0.63	-0.32	89.367	6.184	1.622	97.173
1955	2.09	-0.82	-0.36	81.341	12.631	2.449	96.421
1956	1.77	-0.82	-0.37	75.411	15.968	3.309	94.687
1957	1.64	-1.13	-0.10	65.392	30.871	0.265	96.528
1958	1.55	-0.54	0.08	81.354	9.793	0.229	91.376
1959	1.51	-0.25	0.10	85.705	2.422	0.369	88.497
1960	1.11	-0.30	0.28	66.367	4.846	4.221	75.434
1961	1.04	-0.38	0.22	68.609	8.873	3.006	80.488
1962	0.79	0.61	0.34	36.377	22.020	6.949	65.346
1963	0.46	0.68	0.43	16.077	35.232	14.335	65.644
1964	-0.07	0.23	0.50	0.511	4.859	23.522	28.893
1965	-0.70	0.00	0.86	29.067	0.001	44.099	73.167
1966	-1.08	0.04	1.00	46.132	0.077	39.586	85.794
1967	-1.55	-0.29	1.16	58.300	2.012	32.772	93.084
1968	-2.19	-0.25	1.36	67.366	0.841	25.969	94.177
1969	-2.62	-0.43	1.25	75.665	2.022	17.052	94.739
1970	-3.37	-0.87	1.49	76.152	5.074	14.925	96.152
1971	-3.53	-0.01	1.11	87.051	0.000	8.649	95.700
1972	-4.29	0.36	0.91	92.817	0.665	4.204	97.686
1973	-5.17	0.28	0.62	96.180	0.282	1.396	97.858
1974	-5.59	-0.34	0.81	86.812	0.318	1.833	88.963
1975	-6.45	0.02	0.31	96.068	0.001	0.224	96.293
1976	-6.75	2.39	-0.03	87.674	10.949	0.001	98.625
1977	-7.00	1.11	-1.25	92.162	2.308	2.958	97.427
1978	-8.04	0.17	-1.60	95.543	0.043	3.807	99.393
1979	-10.20	-1.16	-3.15	88.921	1.147	8.490	98.558

CUADRO No. 6

México: 1939 - 1965.

Matriz de correlaciones.

	IBFT	SIP	SEPI	GAST	PEA	PSS	PIIN	PIAG	PINA
IBFT	1.000	0.612	0.639	0.939	0.970	0.487	0.992	0.984	0.973
SIP	0.612	1.000	0.831	0.785	0.581	-0.324	0.632	0.626	0.713
SEPI	0.639	0.831	1.000	0.786	0.642	-0.200	0.631	0.617	0.747
GAST	0.939	0.785	0.786	1.000	0.894	0.204	0.947	0.929	0.972
PEA	0.970	0.581	0.642	0.894	1.000	0.522	0.963	0.959	0.951
PSS	0.487	-0.324	-0.200	0.204	0.522	1.000	0.461	0.462	0.337
PIIN	0.992	0.632	0.631	0.947	0.963	0.461	1.000	0.988	0.973
PIAG	0.984	0.626	0.617	0.929	0.959	0.462	0.988	1.000	0.959
PINA	0.973	0.713	0.747	0.972	0.951	0.337	0.973	0.959	1.000
MAN	0.984	0.710	0.727	0.980	0.958	0.360	0.988	0.977	0.991
PETR	0.977	0.732	0.760	0.981	0.957	0.333	0.978	0.965	0.985
CONS	0.995	0.643	0.672	0.953	0.974	0.441	0.992	0.981	0.984
ELEC	0.968	0.750	0.750	0.989	0.921	0.280	0.973	0.964	0.980
PIME	0.988	0.659	0.702	0.946	0.961	0.433	0.980	0.965	0.971
MXME	0.966	0.469	0.479	0.850	0.917	0.608	0.950	0.954	0.899
EXPO	0.823	0.494	0.572	0.689	0.859	0.464	0.810	0.836	0.780
IMPO	0.888	0.371	0.492	0.717	0.918	0.654	0.865	0.878	0.836
PIB	0.988	0.690	0.718	0.969	0.973	0.391	0.988	0.980	0.991

	MAN	PETR	CONS	ELEC	PIME	MXME	EXPO	IMPO	PIB
IBFT	0.984	0.977	0.995	0.968	0.988	0.966	0.823	0.888	0.988
SIP	0.710	0.732	0.643	0.750	0.659	0.469	0.494	0.371	0.690
SEPI	0.727	0.760	0.672	0.750	0.702	0.479	0.572	0.492	0.718
GAST	0.980	0.981	0.953	0.989	0.946	0.850	0.689	0.717	0.969
PEA	0.958	0.957	0.974	0.921	0.961	0.917	0.859	0.918	0.973
PSS	0.360	0.333	0.441	0.280	0.433	0.608	0.464	0.654	0.391
PIIN	0.988	0.978	0.992	0.973	0.980	0.950	0.810	0.865	0.988
PIAG	0.977	0.965	0.981	0.964	0.965	0.954	0.836	0.878	0.980
PINA	0.991	0.985	0.984	0.980	0.971	0.899	0.780	0.836	0.991
MAN	1.000	0.996	0.991	0.992	0.982	0.919	0.788	0.833	0.998
PETR	0.996	1.000	0.985	0.989	0.984	0.900	0.793	0.822	0.994
CONS	0.991	0.985	1.000	0.976	0.984	0.940	0.808	0.871	0.994
ELEC	0.992	0.989	0.976	1.000	0.970	0.897	0.750	0.779	0.985
PIME	0.982	0.984	0.984	0.970	1.000	0.934	0.831	0.868	0.984
MXME	0.919	0.900	0.940	0.897	0.934	1.000	0.803	0.892	0.923
EXPO	0.788	0.793	0.808	0.750	0.831	0.803	1.000	0.921	0.808
IMPO	0.833	0.822	0.871	0.779	0.868	0.892	0.921	1.000	0.857
PIB	0.998	0.994	0.994	0.985	0.984	0.923	0.808	0.857	1.000

CUADRO No. 7
México: 1939 - 1965.
Valores propios en orden descendente

Num.	Val. propios	% traza	% Acum.
1	15.17	84.259	84.259
2	1.94	10.777	95.036
3	0.45	2.514	97.550
4	0.19	1.067	98.617
5	0.07	0.416	99.033
6	0.06	0.319	99.351
7	0.05	0.252	99.603
8	0.02	0.127	99.730
9	0.02	0.094	99.824
10	0.01	0.062	99.886
11	0.01	0.042	99.928
12	0.01	0.030	99.958
13	0.00	0.017	99.975
14	0.00	0.013	99.989
15	0.00	0.008	99.997
16	0.00	0.003	100.000
17	0.00	0.000	100.000

CUADRO No. 9
México: 1939 - 1965.
Correlaciones entre las variables y los componentes principales.

	EJE1	EJE2	EJE3
IBFT	-0.992	-0.090	-0.065
SIP	-0.682	0.679	0.042
SEPI	-0.714	0.581	0.244
GAST	-0.957	0.224	-0.161
PEA	-0.975	-0.127	0.058
PSS	-0.411	-0.890	-0.048
PIIN	-0.989	-0.064	-0.095
PIAG	-0.983	-0.079	-0.050
PINA	-0.986	0.080	-0.054
MAN	-0.994	0.059	-0.081
PETR	-0.991	0.094	-0.048
CONS	-0.993	-0.037	-0.069
ELEC	-0.980	0.138	-0.121
PIME	-0.990	-0.019	-0.015
MXME	-0.936	-0.262	-0.097
EXPO	-0.843	-0.178	0.474
IMPO	-0.881	-0.352	0.277
PIB	-0.997	0.025	-0.049

CUADRO No. 8
México: 1939 - 1965
Vectores propios. (se incluyen sólo los 6 primeros)

	EJE1	EJE2	EJE3	EJE4	EJE5	EJE6
IBFT	-0.255	-0.065	-0.096	-0.022	0.112	-0.067
SIP	-0.175	0.487	0.062	-0.468	-0.501	-0.350
SEPI	-0.183	0.417	0.362	0.669	0.156	-0.255
GAST	-0.246	0.161	-0.239	0.091	0.073	0.020
PEA	-0.250	-0.091	0.087	0.115	-0.485	0.231
PSS	-0.105	-0.639	-0.072	0.247	-0.386	-0.445
PIIN	-0.254	-0.046	-0.141	-0.106	0.013	0.100
PIAG	-0.253	-0.057	-0.074	-0.228	0.116	0.158
PINA	-0.253	0.057	-0.080	0.115	-0.025	0.290
MAN	-0.255	0.042	-0.121	0.029	-0.011	0.082
PETR	-0.255	0.067	-0.071	0.081	-0.101	-0.054
CONS	-0.255	-0.027	-0.103	0.034	-0.022	0.192
ELEC	-0.252	0.099	-0.179	-0.026	0.191	-0.003
PIME	-0.254	-0.014	-0.022	0.062	0.026	-0.352
MXME	-0.240	-0.188	-0.144	-0.208	0.480	-0.378
EXPO	-0.216	-0.128	0.704	-0.338	0.134	-0.062
IMPO	-0.226	-0.253	0.411	0.032	0.040	0.309
PIB	-0.256	0.018	-0.072	0.051	-0.100	0.153

CUADRO No. 10

México: 1939 - 1965.

Componentes principales y calidad de la representación.

	EJE1	EJE2	EJE3	COS1	COS2	COS3	QLT3
1939	4.57	3.04	0.76	67.569	29.810	1.873	99.253
1940	4.41	3.08	0.60	66.038	32.095	1.241	99.375
1941	4.43	0.38	-0.17	90.787	0.658	0.133	91.579
1942	4.36	1.95	0.03	82.968	16.541	0.005	99.515
1943	4.36	1.36	-0.24	89.791	8.757	0.282	98.829
1944	4.23	0.04	-0.79	94.725	0.008	3.343	98.076
1945	3.84	-0.07	-1.07	91.848	0.031	7.106	98.985
1946	3.30	-1.29	-1.09	78.715	11.979	8.639	99.334
1947	2.82	-1.66	-1.04	66.858	23.162	8.988	99.007
1948	2.59	-1.22	-0.74	75.348	16.680	6.078	98.105
1949	2.24	-0.29	0.05	86.686	1.450	0.037	88.173
1950	1.51	-0.77	0.56	66.511	17.304	9.078	92.894
1951	0.73	-1.67	0.30	13.546	71.838	2.382	87.767
1952	0.59	-2.13	0.14	6.305	83.313	0.377	89.995
1953	0.79	-0.96	0.23	30.492	45.611	2.618	78.721
1954	-0.09	-0.93	0.95	0.458	44.075	46.284	90.816
1955	-0.94	-1.23	1.26	20.679	35.412	36.807	92.898
1956	-1.88	-1.18	1.23	52.224	20.347	22.280	94.851
1957	-2.16	-1.42	0.44	65.019	28.015	2.740	95.774
1958	-2.38	-0.50	0.40	88.085	3.795	2.443	94.323
1959	-2.61	0.11	0.19	93.422	0.161	0.477	94.059
1960	-3.65	0.18	0.01	95.021	0.224	0.001	95.246
1961	-3.98	0.15	-0.24	96.884	0.133	0.338	97.354
1962	-4.74	1.50	0.18	89.821	9.039	0.130	98.991
1963	-5.81	1.70	-0.17	91.004	7.782	0.082	98.868
1964	-7.52	0.69	-0.82	95.976	0.797	1.153	97.926
1965	-9.00	1.16	-0.97	96.009	1.585	1.110	98.704

EL USO DE TRANSFORMACIONES EN MODELOS DE REGRESION

Miguel Nakamura
Centro de Investigación en Matemáticas, A.C.
Apartado Postal 402, Guanajuato, Gto., 36000, México.

RESUMEN

Las transformaciones se han propuesto para evitar las consecuencias de la violación de los supuestos comunes sobre los errores en modelos de regresión, como lo es por ejemplo la pérdida de eficiencia de los estimadores clásicos como mínimos cuadrados. Se presenta una breve reseña de modelos que incorporan una transformación así como varios estimadores que se han propuesto. Se hace énfasis en algunos de los métodos más recientes, los cuales son semi-paramétricos en el sentido de que sus propiedades asintóticas dependen únicamente del supuesto de que los errores provienen de una distribución simétrica. En particular, con mayor detalle se describe un estimador que minimiza la distancia de Cramer-von Mises entre la distribución empírica de los residuales y su reflexión alrededor de cero. Las propiedades de dicho estimador se derivan de una teoría de M-estimadores generalizados.

1. INTRODUCCION

En la búsqueda de relaciones entre una variable dependiente Y y un vector de variables independientes X, frecuentemente se adopta el modelo básico de regresión no-lineal dado por

$$Y_i = f(X_i,\beta) + \sigma\varepsilon_i \tag{1.1}$$

para observaciones $(X_1,Y_1),\ldots (X_n,Y_n)$. El objetivo principal de los métodos de regresión es estimar el valor del parámetro desconocido β, con el objeto de construir intervalos de predicción, o simplemente porque el valor de β tiene alguna interpretación física de interés.

Existen varias características que pueden suponerse sobre los errores ε_i, por ejemplo, que sean independientes, que tengan esperanza cero, y que cada ε_i tenga la misma varianza. Es muy común que se suponga que ε posee una distribución normal. En este caso, el método de mínimos cuadrados, el cual consiste en encontrar el valor de $\hat{\beta}$ que minimiza la suma de cuadrados de residuales dada por

$$\sum_{i=1}^{n} \{Y_i - f(X_i,\beta)\}^2 \tag{1.2}$$

sobre valores de β, es un método consistente y asintóticamente eficiente

(consúltese, por ejemplo, Jennrich (1969)). Sin embargo, si la distribución de ε no se asemeja a una distribución normal, las características óptimas del estimador por mínimos cuadrados pueden verse seriamente afectadas; en particular, si los errores ε_i tienen una distribución sesgada, o si no tienen varianza constante para cada i.

Existen varios métodos para contrarrestar los efectos de errores que no sean simétricos y de varianza constante. Si la varianza de ε no parece ser constante para cada observación, es posible utilizar mínimos cuadrados ponderados, o bien otros modelos propuestos explícitamente para manejar dicha condición. Entre ellos, se encuentran los modelos que involucran funciones de varianza, en los cuales se asume que la desviación estándar de cada observación depende de X a través de una función g con cierto parámetro:

(1.3) $$Y_i = f(X_i,\beta) + g(X_i,\alpha)\varepsilon_i$$

(ver Carroll y Ruppert (1988)). Por otra parte, si la distribución de los errores deja de ser normal, uno de los enfoques que se han propuesto para modificar la distribución de los errores, es el empleo de transformaciones. El objetivo del presente artículo es presentar brevemente la metodología existente para transformaciones y resaltar la bibliografía básica. Se describirán algunos estimadores recientes de transformaciones en regresión cuyas propiedades dependen sólo de la simetría de los errores.

2. TRANSFORMACIONES

Una transformación h(y) puede tener efecto sobre la distribución de una colección de variables aleatorias $Y_1,\ldots,Y_n$ en dos formas distintas. Por ejemplo, si la distribución de Y_i es sesgada positivamente, la distribución de $h(Y_i)$ puede resultar ser aproximadamente normal si h es una función estrictamente cóncava. Por otra parte, si las desviaciones estándar de cada

Y_i son distintas (por ejemplo funciones crecientes de sus valores esperados), es posible que las variables aleatorias $h(Y_1),\ldots,h(Y_n)$ tengan desviación estándar constante. Ocurre con frecuencia en la práctica que un juego de datos muestre simultáneamente sesgo positivo y una varianza que no es constante para cada X. En estos casos, el empleo de una sola transformación puede convertir los datos a una distribución simétrica y de varianza homogénea. El problema consiste, entonces, en la selección de una transformación apropiada. En regresión, como consecuencias de que se consigan errores simétricos y de varianza homogénea, los estimadores para el parámetro β tendrán mejores propiedades y los intervalos de predicción serán más precisos que si se emplearan otros métodos.

El uso de una transformación en regresión fue sugerido inicialmente por Box y Cox (1964). En lugar de suponer que la distribución de Y sea simétrica directamente, se supone que existe una tranformación h(y) tal que h(Y) es simétricamente distribuida. La transformación h(y) puede seleccionarse a través de un parámetro λ. Para una variable estrictamente positiva, Box y Cox proponen la familia de transformaciones dada por

$$h(y,\lambda) = (y^{\lambda}-1)/\lambda \quad \text{si } \lambda \neq 0$$
$$= \ln(y) \quad \text{si } \lambda = 0 , \tag{2.1}$$

y consideran el modelo

$$h(Y_i,\lambda) = X^T\beta + \sigma\varepsilon_i . \tag{2.2}$$

El modelo de Box y Cox dado por (2.2) presupone que existe una transformación (o sea, un valor de λ) tal que se obtienen simultáneamente tres características: (i) normalidad de errores, (ii) varianza constante para cada i, y (iii) una estructura sencilla para la esperanza condicional de Y dado X (nótese la elección de la forma $X^T\beta$).

La introducción de un parámetro adicional (λ) en el modelo implica la

necesidad de estimarlo adecuadamente. Box y Cox proponen estimar (λ,β,σ) por el método de máxima verosimilitud bajo normalidad de los errores. Si los errores no son normales, el método produce estimadores que no son necesariamente consistentes (Hernández y Johnson (1980), Bickel y Doksum (1981)). Es por este motivo que varios autores han propuesto estimadores alternos al de Box y Cox. En Hinkley (1975, 1977), y Taylor (1985), pueden encontrarse ejemplos de métodos que producen estimadores consistentes bajo la suposición de errores simétricos (no necesariamente normales) cuando el modelo es

$$h(Y_i,\lambda) = \theta + \sigma\varepsilon_i \ . \tag{2.3}$$

Estos últimos métodos se basan en el empleo de una estadística que mide simetría.

Cabe mencionar que a pesar de que la transformación (2.1) depende de un solo parámetro λ, el modelo de transformación podría definirse para parámetros de transformación de mayores dimensiones, aunque los procedimientos para estimarlos se complicarían. Algunos de los métodos propuestos no son aplicables a parámetros de este tipo.

Con respecto a la transformación de Box y Cox, Hernández y Johnson (1980) estudian el comportamiento del estimador de Box y Cox cuando los errores no son normales y caracterizan el límite del mismo. Bickel y Doksum (1981) tratan el modelo del Box y Cox con una densidad general g para los ε_i en lugar de la normal estándar y estudian la consistencia y propiedades robustas del estimador. Propiedades robustas también son analizadas en Carroll (1980). El problema de estimar la mediana condicional de Y dado X es tratado en Carroll y Ruppert (1981) y el de estimar la media condicional de Y dado X, en Taylor (1986).

3. EL MODELO DE TRANSFORMACION A AMBOS LADOS

Carroll y Ruppert (1984) proponen un modelo distinto para transformaciones en regresión, el cual se describe a continuación. En aplicaciones, es frecuente que se conozca la relación teórica entre X y Y, salvo por el valor desconocido de un parámetro. Por ejemplo, es común que en física o química se sepa de antemano que

(3.1) $$Y = f(X,\beta) .$$

Sin embargo, al obtenerse mediciones, habrá la influencia de errores de medición y variabilidad natural en el problema, lo cual causa que la relación (3.1) no sea observada exactamente y por tanto, sea natural considerar un término aleatorio en (3.1). Sin embargo, en la práctica la forma precisa en que el error se manifiesta no se conoce. La manera más común y sencilla de incluir el error, es en forma aditiva, o sea, obteniéndose el modelo (1.1). Como se mencionó anteriormente en la sección 1, si el error no se distribuye en forma normal, los métodos usuales para estimar β carecen de buenas propiedades. Mediante la consideración de una transformación a la manera de Box y Cox se podrían resolver algunos de los problemas relacionados con la pérdida de eficiencia en la estimación del parámetro β, pero se destruiría entonces la relación teórica (3.1) que existe entre X y Y. El modelo de transformación a ambos lados de Carroll y Ruppert (1984) fue propuesto con el objeto de evitar esta dificultad, y está dado por

(3.2) $$h(Y_i,\lambda) = h(f(X_i,\beta),\lambda) + \sigma\varepsilon_i .$$

En la propuesta original, se asume que cada ε_i tiene una distribución normal, aunque, como se verá más adelante, es interesante apartarse de esta condición y trabajar simplemente con errores simétricos. Bajo el modelo (3.2) y si la familia de transformaciones $h(y,\lambda)$ es monótona, la ausencia de error de medición implica (3.1), o sea, que se conserva la relación teórica entre X y

Y. Por otra parte, este enfoque da lugar a que la variabilidad del sistema afecte la relación (3.1) en forma no necesariamente aditiva. Es importante mencionar que el modelo (3.2) no constituye una generalización del modelo (2.2), pues cada modelo pretende cumplir con diferentes objetivos. Mientras que el modelo (2.2) trata de obtener los objetivos (i), (ii) y (iii) mencionados en la sección 1, el modelo (3.2) trata de modelar la forma en que la variabilidad ingresa a una relación previamente determinada. Se han desarrollado métodos asintóticos para probar la hipótesis nula H_0: $\lambda = \lambda_0$, con lo cual es posible probar si los datos justifican una transformación dada, como podría ser si la transformación es logarítimica (tómese $\lambda_0 = 0$), o la validez del modelo (1.1) (tómese $\lambda_0 = 1$).

Para estimar el modelo (3.2), Carroll y Ruppert (1984) estudian las propiedades del estimador de máxima verosimilitud, $\hat{\lambda}_{MV}$, usando la transformación (2.1) y considerando errores con distribución normal. Al igual que en el caso del modelo de Box y Cox, el estimador de máxima verosimilitud no es consistente bajo desviaciones del supuesto de normalidad, y tampoco es robusto en presencia de observaciones aberrantes. De lo anterior se desprende la necesidad de estudiar otros estimadores cuyas propiedades no dependan de la normalidad de los errores. Como ejemplo, se podría pensar en estimadores consistentes bajo supuestos de simetría en los errores, sin que se tenga necesariamente normalidad. En este sentido, los métodos alternativos serían semi-paramétricos, pues no dependerían del conocimiento exacto de la distribución de los errores.

Ejemplos de aplicaciones de transformación a ambos lados pueden encontrarse en diversas disciplinas en Snee (1986); Ruppert y Carroll (1985); Bates, Wolf y Watts (1986); y Ruppert, Cressie, y Carroll (1989), entre otras. El estudio del estimador de máxima verosimilitud se lleva a cabo en Carroll y

Ruppert (1988) (capítulo 4). Para el desarrollo de técnicas para la detección de observaciones aberrantes y dignósticos de influencia, véase Carroll y Ruppert (1987).

Ruppert y Aldershof (1989) consideran el modelo de transformación a ambos lados y la transformación (2.1) y proponen estimadores nuevos para los parámetros. El primero de ellos será descrito a continuación, y está basado en simetría de los errores. Es crucial definir primero los residuales que resultan del modelo (3.2), es decir,

$$r_i(\lambda,\beta) = h(Y_i,\lambda) - h(f(X_i,\beta),\lambda) \ . \tag{3.3}$$

Con $\hat{\beta}(\lambda)$ y $\hat{\sigma}(\lambda)$ denotamos los estimadores de mínimos cuadrados de β y σ, para cada valor fijo de λ, es decir,

$$\hat{\beta}(\lambda) \quad \text{minimiza} \sum_{i=1}^{n} \{r_i(\lambda,\beta)\}^2 \ , \ \text{y} \tag{3.4}$$

$$\hat{\sigma}^2(\lambda) = (1/n) \sum_{i=1}^{n} \{r_i(\lambda,\hat{\beta}(\lambda))\}^2 \ . \tag{3.5}$$

Se define el *estimador basado en sesgo* $\hat{\lambda}_S$ como aquel valor de λ que es solución de la ecuación

$$C_n(\lambda) = (1/n) \sum_{i=1}^{n} \{r_i(\lambda,\hat{\beta}(\lambda))/\hat{\sigma}(\lambda)\}^3 = 0 \ . \tag{3.6}$$

Intuitivamente, el estimador $\hat{\lambda}_S$ es aquel valor de λ que produce que cierto criterio para simetría $C_n(\lambda)$ aplicado a los residuales sea igual a cero. Dicho criterio es el coeficiente de simetría (o sesgo), basado en el tercer momento de una variable aleatoria W con valor esperado μ y desviación estándar σ, el cual se define por $E(W-\mu)^3/\sigma^3$. Es posible demostrar la consistencia del estimador $\hat{\lambda}_S$, bajo la suposición de simetría de los errores y sin suponer normalidad de ellos. En el caso de que los errores sean efectivamente normales, la pérdida de eficiencia con respecto al estimador de máxima verosimilitud no es considerable, y por otra parte, Ruppert y Aldershof

han demostrado que la elección $u(x) = x^3$ es asintóticamente óptima entre los criterios de simetría que son de la forma

$$C_n(\lambda) = (1/n) \sum_{i=1}^{n} u(r_i(\lambda,\hat{\beta}(\lambda))/\hat{\sigma}(\lambda)) \tag{3.7}$$

en donde $u(x)$ es una función impar.

4. ESTIMADOR DE MINIMA DISTANCIA

Un procedimiento alternativo para estimar el parámetro λ en el modelo de transformación a ambos lados, es considerado en Nakamura (1989), basado en un criterio para medir asimetría que tiene sus orígenes en problemas de estimación por mínima distancia. Con el objeto de describir el estimador, se consideran los residuales $r_i(\lambda,\beta)$ definidos por (3.3). Asimismo, se emplea la notación $\hat{\beta}(\lambda)$ y $\hat{\sigma}(\lambda)$ para los estimadores por mínimos cuadrados de β y σ definidos por las relaciones (3.4) y (3.5), respectivamente.

Para cada valor fijo de λ, se denota la distribución empírica de los residuales normalizados por $F_n(\lambda,x)$, es decir, para $-\infty < x < \infty$,

$$F_n(\lambda,x) = (1/n) \sum_{i=1}^{n} 1\{r_i(\lambda,\hat{\beta}(\lambda))/\hat{\sigma}(\lambda) \le x\}. \tag{4.1}$$

La cantidad

$$V_n(\lambda) = \int \{F_n(\lambda,x) + F_n(\lambda,-x)-1\}^2 \, w(x)dx \tag{4.2}$$

es una medida de asimetría de la distribución $F_n(\lambda,x)$. (Nótese que si una distribución G es simétrica, entonces $G(x) + G(-x) - 1 = 0$ para toda x). En (4.2), la función $w(x)$ es una función positiva y simétrica, fija, que se introduce con el objeto de tener influencia sobre la varianza del estimador resultante. Existen motivos para concluir que se obtienen mejores estimadores de λ si la función $w(x)$ es convexa; como ejemplo, para la familia de transformaciones dada por (2.1) y errores normales, una elección óptima es $w(x) = \exp(x^2/2)$ (ver Nakamura (1989)).

El *estimador de mínima distancia* se define como aquel valor de λ que minimiza $V_n(\lambda)$ como función de λ, y es denotado por $\hat{\lambda}_{MD}$. Su interpretación es que es el valor de λ que fuerza a que los residuales sean lo más simétrico posible, de acuerdo al criterio $V_n(\lambda)$.

Bajo algunas condiciones usuales de regularidad sobre la familia de transformaciones $h(y,\lambda)$ y la función de regresión $f(x,\beta)$, es posible demostrar que el estimador $\hat{\lambda}_{MD}$ es fuertemente consistente y asintóticamente normal, suponiendo tan solo que el error ε es simétrico. El análisis teórico del estimador de mínima distancia se deriva de una teoría que generaliza los M-estimadores de Huber (1967). Esto se debe a que escribiendo $Z_i=(X_i,Y_i)$, puede mostrarse que el estimador de $\theta = (\beta,\lambda,\sigma^2)$ es una solución de la ecuación

$$(1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} \Psi(Z_i,Z_j,\theta) = 0$$

para cierta función Ψ. Para los detalles, así como la expresión para la varianza asintótica del estimador, véase Nakamura (1989).

Las características asintóticas anteriores se cumplen cuando la familia de transformaciones está dada por (2.1), pero la ventaja primordial del estimador $\hat{\lambda}_{MD}$ es que produce estimadores consistentes de λ aún en el caso de que la dimensión de λ sea mayor que uno.

Un ejemplo importante de una familia de transformaciones con parámetro multidimensional, es la llamada *familia de potencias con translación*, definida por

$$h(y,\lambda_1,\lambda_2) = ((y + \lambda_1)^{\lambda_2} - 1)/\lambda_2 \quad \text{si } \lambda_2 \neq 0$$
$$= \ln(y + \lambda_1) \quad \text{si } \lambda_2 = 0 \,. \tag{4.3}$$

Con la familia (4.3), las observaciones de Y no tienen que ser necesariamente positivas, de modo que se extiende la familia (2.1) en una

dirección importante. Si se tienen observaciones negativas, se les suma primero una cantidad λ_1 antes de tomar la potencia. La elección adecuada de la cantidad λ_1 es importante, puesto que se ha encontrado que si se elige λ_1 arbitrariamente, las estimaciones de λ_2 son fuertemente dependientes de la elección de λ_1 (Carroll y Ruppert (1988)). De aquí que sea de interés el obtener un procedimiento de estimación para elegir λ_1 basado en los datos.

Atkinson (1985), presenta un capítulo en el cual se considera la transformación (4.3) con detalle. Ahí puede verse que el método de máxima verosimilitud falla para estimar (λ_1,λ_2) simultáneamente, aún en el caso de errores normales. La dificultad se debe a que la verosimilitud se vuelve infinita para algunos valores de λ_1. Por otra parte, otros estimadores como el estimador $\hat{\lambda}_S$ de Ruppert y Aldershof, no están definidos para parámetros de transformación con dimensión mayor que uno. El estimador de mínima distancia ofrece una solución al problema de estimar λ_1.

Si el modelo (3.2) se cumple y la distribución de los errores es simétrica, entonces la cantidad $V_n(\lambda)$ converge a cero cuando λ es el valor correcto del parámetro, y converge a algo estrictamente positivo en caso contrario. Por tanto, la cantidad minimizada $V_n(\hat{\lambda}_{MD})$ tiene el potencial para ser utilizada para probar la hipótesis nula de que existe una transformación que produce errores simétricos. Por supuesto, para encontrar el criterio para rechazar la hipótesis nula, es necesario determinar cuál es la distribución de $V_n(\hat{\lambda}_{MD})$. Esperamos investigar este problema en el futuro.

5. EJEMPLO

Consideramos un ejemplo en el campo de la biología (Carroll y Ruppert (1988), ejemplos 4.3 y 6.2). Los datos se refieren a un experimento para estudiar el número de bacterias presentes en los pulmones de ratones, después

de habérseles administrado ciertos antibióticos en una cámara de gas. La variable X es el tiempo en horas a que fué sacrificado cada animal después de haber sido expuesto a las bacterias; en el estudio, X fue 0, 4, o 24 horas. La variable Y es el cociente entre el número de bacterias que entran al pulmón inicialmente y el número de bacterias encontradas al momento del sacrificio, expresado como porcentaje.

A manera de ilustración, consideramos el grupo de 18 animales tomados como grupo control, con la función

$$f(X,\beta) = \exp(\beta_0 + \beta_1 X)$$

la cual tiene una interpretación en biología. El juego de datos es el siguiente:

X=0; Y=103.3, 92.5, 72.1, 41.4, 63.7, 74.2

X=4; Y=26.4, 52.7, 45.5, 63.1, 29.1, 27.8

X=24; Y=0, 1.5, 1.7, 0.6, 2.0, 1.5.

Carroll y Ruppert (1984) aplican el modelo de transformación a ambos lados con la transformación (4.3). La necesidad de sumar λ_1 a las observaciones obedece a que existen observaciones muy próximas a cero o cero dada la precisión del instrumento de medición. Para estudiar el comportamiento del estimador de maxima verosimilitud de λ_2 y para hacer comparaciones (y ante la imposibilidad de estimar λ_1), se seleccionan arbitrariamente los valores λ_1=0.05 y λ_1=1. Las estimaciones que se obtienen para λ_2 son -0.069 y -0.18 respectivamente. Como λ_2 es una potencia, estos valores representan transformaciones substancialmente distintas. Por otra parte, las técnicas de diagnóstico desarrolladas por Carroll y Ruppert (1987) también muestran que $\hat{\lambda}_2$ esta fuertemente influenciado por λ_1. Es deseable seleccionar entonces λ_1 con base en los datos. El valor λ_2=-0.069 es cercano a la transformación logarítmica y esta transformación es la que sostienen los autores en el

ejemplo, debido a que homogeniza la varianza y produce un modelo que explica observaciones extremas.

Una aplicación del método de mínima distancia descrito en la sección 4, resulta en las estimaciones $\hat{\lambda}_1$=3.78, $\hat{\lambda}_2$=-0.049, $\hat{\beta}_0$=4.29, y $\hat{\beta}_1$=-0.17, lo cual corresponde cercanamente al valor de λ_2 obtenido en Carroll y Ruppert (1984).

6. REFERENCIAS

Atkinson, A. C. (1985) *Plots, Transformations, and Regression*. Clarendon Press, Oxford.

Bates, D. M., Wolf, D. A., and Watts, D. G. (1986) "Nonlinear Least Squares and First-order Kinetics", in *Proceedings of Computer Science and Statistics: Seventeenth Symposium on the Interface,* ed. David Allen, New York: North Holland.

Bickel, P. J., and Doksum, K. A. (1981) "An Analysis of Transformations Revisited", *Journal of the American Statistical Association,* 76, 296-311.

Box, G. E. P. and Cox, D. R. (1964) "An Analysis of Transformations". *Journal of the Royal Statistical Society*. B 26, 211-252.

Carroll, R. J. (1980) "A Robust Method for Testing Transformations to Achieve Approximate Normality", *Journal of the American Statistical Asssociation,* 74, 674-679.

Carroll, R. J. and Ruppert, D. (1981) "On Prediction and the Power Transformation Family", *Biometrika,* 68, 609-615.

Carroll, R. J. and Ruppert, D. (1984) "Power transformations when Fitting Theoretical Models to Data", *Journal of the American Statistical Association,* 79, 321-328.

Carroll, R. J. and Ruppert, D. (1987) "Diagnostics and Robust Estimation When Transforming the Regression Model and the Response", *Technometrics,* 29, 287-299.

Carroll, R. J. and Ruppert, D. (1988) *Transformations and Weighting in Regression*, Chapman and Hall, New York and London.

Hernández, F. and Johnson, R. A. (1980) "The Large-Sample Behaviour of Transformations to Normality", *Journal of the American Statistical Association,* 75, 855-861.

Hinkley, D. V. (1975) "On Power Transformations to Symmetry", *Biometrika* 62, 101-111.

Hinkley, D. F. (1977) "On Quick Choice of Power Transformation", *Applied Statistics* 26, 67-68.

Huber, P. J. (1967) "Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions", *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* 1, 221-233.

Jennrich, R. I. (1969) "Asymptotic Properties of Nonlinear Least Squares Estimators", *Annals of Mathematical Statistics* 40, 633-643.

Nakamura, M. (1989) "Transformations to Symmetry in the Transform-Both-Sides Regression Model" Mimeo Series #2004, Department of Statistics, University of North Carolina at Chapel Hill.

Ruppert, D. and Aldershof, B. (1989) "Transformations to Symmetry and Homoscedasticity", *Journal of the American Statistical Association* 81, 437-446

Ruppert, D. and Carroll, R. J. (1985) "Data Transformations in Regression Analysis with Applications to Stock-Recruitment Relationships", in *Resource Management,* M. Mangel editor, *Lecture Notes in Biomathematics* 61, Springer Verlag, New York.

Ruppert, D., Cressie, N., and Carroll, R. J. (1989) "A Transformation/ Weighting Model for Estimating Michaelis-Menten Parameters", *Biometrics,* 45, 637-656.

Snee, R. D. (1986) "An Alternative Approach to Fitting Models when Re-expression of the Response is Useful", *Journal of Quality Technology,* 18, 211-225.

Taylor, J. M. G. (1985) "Power Transformation to Symmetry", *Biometrika* 72, 145-152.

Taylor, J. M. G. (1986) "The Retransformed Mean After a Fitted Power Transformation", *Journal of the American Statistical Association,* 81, 114-118.

SOBRE LA PROBLEMATICA DEL ANALISIS DE DATOS DE ENCUESTAS.

Mario Miguel Ojeda
Laboratorio de Investigación y Asesoría Estadística
Fac. de Estadística, Universidad Veracruzana
Av. Xalapa esq. A. Camacho, Xalapa, Ver. México.

RESUMEN

En este trabajo se hace un análisis preliminar de los aspectos filosóficos, teóricos y metodológicos que conforman la problemática del análisis de datos generados a través de encuestas. Se caracteriza a los tipos de problemas que se presentan en el contexto de la planeación y realización de una investigación que considera una encuesta. Se aborda el problema filosófico relacionado con el enfoque del diseño y del análisis a través de modelos superpoblacionales, considerando dos tipos de inferencia, descriptiva y analítica, en el marco de datos provenientes de muestras complejas. Se refiere la consideración del Análisis Exploratorio, del Análisis Inicial y de la postulación de modelos. Se considera, también, la bondad del uso de técnicas estadísticas multivariadas. A lo largo del trabajo recomendaciones generales son hechas.

INTRODUCCION

Una gran cantidad de aplicaciones de la Estadística tienen que ver con una encuesta y, sin embargo, la metodología para el análisis de datos provenientes de encuestas aún no se encuentra integrada y tratada de manera global. La experiencia del autor le indica que la problemática se aborda marginalmente en cuatro diferentes niveles, los cuales se enuncian a continuación.

(1) Planeación del levantamiento de la encuesta.

(2) Planeación de la verificación de la calidad.

(3) Planeación del análisis de los datos.

(4) Planeación para la elaboración del informe.

En teoría, la planeación de una investigación que involucre una encuesta debería considerar integralmente y de manera conjunta todos los niveles, determinando todas las actividades que, de manera global y particular, se abordarán en el estudio en cuestión. Obviamente todo esto deberá estar

integrado a los objetivos de la investigación, a su adecuado y bien definido marco teórico y a un esquema global de la realización y flujo del plan de acción. Pero cabe preguntarse: ¿qué sucede en la mayoría de los casos?. ¿Por qué quienes planean y realizan las investigaciones que involucran encuestas rara vez consideran la problemática aquí planteada? Múltiples razones saltan a la vista, y grandes tareas se adivinan como responsabilidad de los metodólogos (llámense Estadísticos, Matemáticos Aplicados o más especificamente Muestristas). ¿Por qué grandes responsabilidades para ellos? Por el simple hecho de que este tipo de profesionales tienen los elementos para lograr una visión más global, más completa en sus partes y más profunda. Sin embargo esto no quiere decir que el Estadístico o Muestrista sea el responsable único, o el profesional que tenga toda la carga que involucra vigilar la calidad de una encuesta. Desde luego que el Estadístico deberá propiciar el concurso de todo el equipo de investigación en los diferentes niveles de la realización del trabajo investigativo. En este sentido debe entenderse que este profesional comprenda mayormente el desarrollo metodológico de la investigación y en qué radica que esta provea resultados de alto valor por su adecuada conducción.

En este escrito planteamos un análisis general de la problemática reseñada. No abundamos en todos los aspectos señalados en esta introducción, pero tratamos de dar los puntos de partida para una reflexión y estudio más profundo y completo de este tema, particularmente en lo referente al análisis de datos.

I.- PROBLEMATICA DE UNA ENCUESTA

Los problemas que se presentan en una investigación que involucra una encuesta son múltiples. Todos se relacionan directamente con la metodología de la investigación y la aplicación del método científico, particularmente bajo la consideración de los métodos cuantitativos.

Jessen (1978) plantea que aunque los tipos de problemas que se presentan de encuesta a encuesta pueden variar, los más típicos se pueden agrupar en las siguientes categorías:

(1) En la determinación de los objetivos de la investigación.

(2) En las consideraciones para el diseño de la encuesta.

(3) En el diseño muestral.

(4) En los métodos de medición.

(5) En el procesamiento de los datos.

(6) En los procesos de estimación y determinación de la confiabilidad.

(7) En la determinación de la calidad.

(8) En la presentación de los resultados.

La determinación de los objetivos de la investigación, en cuyo contexto se considera la encuesta, es a decir de muchos de los que enfrentan frecuentemente esta tarea, la fase más difícil. Si la planeación global se hiciere, aquí el Estadístico podría incorporar recomendaciones de gran valor. Desgraciadamente a veces no se le considera en esta fase y la etapa de evaluación de los factores que determinan el diseño de la encuesta es para él más tediosa, porque involucra la comprensión clara de los objetivos de la investigación. De todas formas, a pesar de que los problemas que surgen en los puntos (1) y (2) son todos, digamos, de caracter teórico, determinan preponderantemente la calidad global de la investigación. El balance entre factores económicos, de restricciones de tiempo y de la precisión deseada en los resultados, es una tarea fundamental en el diseño, en la que el concurso del Estadístico es determinante.

El contexto del diseño muestral incluye una gran gama de problemas estadísticos, pero desde la perspectiva del arte de diseñar la forma de configurar y captar la muestra, la determinación de tamaños de muestra y forma de afijación son dos aspectos de carácter técnico a los que, en el enfoque clásico, se ha dedicado especial atención. Otros aspectos como la determinación de la información auxiliar a considerar y el procedimiento mismo de selección (esquema de muestreo) son cuestiones cotidianas a tratar por el Muestrista.

Debemos señalar que en los puntos (1), (2) y (3) se va desarrollando de manera implicita la implementación del proceso de la metodología de investigación. La definición de las hipótesis considera ya una clara definición de objetivos, y la definición operacional y estadística de variables implica la consideración de información específica para el diseño muestral. Ahora, una vez determinado el diseño muestral, o de manera paralela al proceso de determinación de éste, se debe plantear el diseño del cuestionario, generalmente el instrumento de medición en una encuesta. Sobre esto existe una gran cantidad de trabajos que se han realizado, los cuales se traducen frecuentemente en recomendaciones metodológicas.

Una tarea que no es abordada, con bastante frecuencia, con la seriedad que se debiera es la planeación del análisis de los datos que producirá la encuesta. Más bien, podría decirse que esto no se realiza en una gran mayoría de los casos. En las secciones posteriores discutiremos aspectos paralelos y concomitantes a tal cuestión.

Digamos que la tarea que abordan con mayor entrega y conciencia los Muestristas es la referente al punto (6); es más, grandes volúmenes culminan con la determinación de estrategias de estimación e inferencia vía datos obtenidos dado un diseño muestral. La literatura referente a este tema es sin duda la más abundante.

Se ha hecho frecuente, al planear una encuesta, la consideración del diseño de un sistema para verificar la calidad del levantamiento y de la información a procesar. En este contexto se encuentran problemas interesantes, incluso desde el punto de vista de la teoría y la metodología estadística. Aspectos de la no respuesta y análisis relacionado con la información faltante son también de interés.

El nivel del detalle y la forma en que se presentarán los resultados constituyen una tarea que tiene mucho de artística.

Debemos señalar que nuestro objetivo, en adelante, es discutir lo que se refiere al tipo de problemas que se presentan en el contexto del punto (5) y las relaciones directas e indirectas que esto guarda con el resto de los puntos.

II.- DISEÑO E INFERENCIA

El punto de vista clásico en el muestreo considera que la población de interés está constituida por un conjunto de individuos que una vez medidos proveen un conjunto de valores "fijos"; es decir se considera que cada individuo en la población posee un valor fijo para la característica o características de
interés. El objetivo en este enfoque consiste en que a partir de los datos obtenidos a través de una muestra (una parte de los individuos seleccionados mediante un mecanismo aleatorio manipulado por el investigador) obtener una estimación de alguna característica de la población, que generalmente es una función de los valores en toda la población y que se denomina parámetro. Comunmente se refieren la media, el total, alguna razón o proporción, etc. Con esta visión se ha desarrollado lo que se conoce como teoría y metodología clásica del muestreo. El libro de Cochran (1979) es una referencia bastante popular que presenta un tratamiento suficiente para la comprensión de las técnicas de selección de la muestra y procedimientos de inferencia en este contexto. Deben señalarse aquí los trabajos de Godambe (1955, 1960),los cuales mostraron la necesidad de realizar estudios más profundos con el objetivo de unificar la teoría del muestreo con la teoría de la inferencia estadística en general, lo que dio origen a nuevas tendencias tanto en

estudios teóricos como metodológicos.

En el enfoque clásico del muestreo los modelos se empezaron a usar de manera implicita en los procesos de estimación, particularmente en la estimación de razón y de regresión. Pero recientemente se aceptó su uso de manera explicita, lo que dió origen a un nuevo enfoque para tratar el problema de muestreo de una población de tamaño conocido (finita): el enfoque superpoblacional. Básicamente la diferencia con el esquema clásico radíca en que los valores de la población finita se asumen como generados por un proceso estocástico: el modelo superpoblacional. Así, la muestra final obtenida se podrá tratar como una muestra de una población infinita obtenida en dos etapas: la primera que es el proceso mediante el que se genera la población finita, y la segunda que es el proceso aleatorio, o no, mediante el que se obtiene la muestra final.

El uso de los modelos superpoblacionales ha permitido la consideración de nuevos propósitos en los análisis estadísticos para los datos de una encuesta obtenida por muestreo. Esto se justifica también desde el punto de vista práctico, ya que las encuestas son bastante costosas, en general, y no es posible, por lo tanto, que la información que proveen sea utilizada simplemente para estimar medias, totales, proporciones y razones, que es lo único permisible por la vía del enfoque clásico.

Cuando se asume un modelo superpoblacional se plantea una familia de relaciones que interesan al investigador como tendencias generales de patrones de comportamiento en la población finita, la cual se considera, no como un ente estático, sino quese le da movimiento en el tiempo y el espacio. Esta concepción es bastante realista, ya que las encuestas generalmente se refieren a poblaciones que cambian, pero que sin otra alternativa en el enfoque clásico se asumen fijas. Al decir cambian nos referimos a que en el momento de la toma de los datos se tiene una determinada población, pero a la hora de obtener las conclusiones, generalmente varios meses después, se tiene otra población.

La inferencia que se puede realizar vía el enfoque de población fija se refiere, como ya se señaló, solo a constantes poblacionales, bajo este enfoque fijas (totales, medias, etc.), pero no a parámetros en un modelo. Kish y Frankel (1974) discuten aspectos de inferencia bajo el enfoque de población fija, incluso definiendo parámetros poblacionales como funciones complejas de las observaciones en la población fija. Estos "parámetros" pueden referir implicitamente a un modelo, pero este no se asume de manera específica.

El enfoque superpoblacional ha permitido introducir un propósito más general en la inferencia a través de datos de encuestas; es decir, permite vía la consideración explicita

de modelos hacer directamente inferencia sobre los parámetros en los modelos postulados. A este tipo de inferencia se le llama analítica, para diferenciarla de la otra que tiene propósitos descriptivos.

Los resultados de la teoría de la inferencia estadística, en general, se pueden, con adaptaciones, en muchos casos simples, extender al tratamiento de los modelos superpoblacionales, e incluso una teoría de estimación de funciones de los valores de la población fija finita se puede construir. (ver Rodrigues y otros (1985)).

Bajo la visión de los modelos superpoblacionales se obtienen múltiples ganancias desde el punto de vista del análisis de datos. Es posible considerar estrategias generales para analizar encuestas una vez que los datos han sido recogidos. En la siguiente sección formalizaremos escuetamente aspectos de inferencia, análisis de datos y elementos de modelos superpoblacionales. Una discusión interesante respecto de los modelos superpoblacionales puede hallarse en Smith (1984).

III.- MODELOS SUPERPOBLACIONALES Y ANALISIS DE DATOS

Con bastante frecuencia en investigaciones sociales y económicas se utiliza a la encuesta como instrumento para colectar información relevante para algún propósito. Esta información se usa para estimar características de interés en una población finita $U = \{ U_1,\ldots,U_N\}$, las cuales son siempre promedios, totales ó razones en un conjunto de variables $\{X_1,\ldots, X_P, Y_1,\ldots, Y_q\}$. La encuesta se planea o diseña usando la información auxiliar z y de acuerdo a un esquema probabilístico p(s/z), al cual se asocia un procedimiento práctico para determinar las unidades a encuestar.

Mediante este proceso se obtiene la información que provee la encuesta, la cual está contenida en un conjunto de datos

$d_i = (x_i, y_i)$; $i = 1,2,\ldots,n$, donde n es el número de elementos diferentes en s, la muestra seleccionada de U bajo el esquema p(s/z), x_i y y_i son los vectores de observaciones asociados al individuo i-ésimo en la muestra.

El costo de una encuesta obliga a que los propósitos del análisis de los datos sean múltiples, y más aún una política racional en este sentido deberá indicar que es necesario aplicar todo tipo de análisis posible y que provea alguna información relevante tanto para la investigación en cuyo contexto se planea la encuesta como para propósitos

adicionales. Es por esto que en una investigación particular, dado un conjunto de variables $\{X_1,\ldots, X_p, Y_1,\ldots Y_q\}$ para los cuales se tienen los datos d_i; i=1, 2,..., n, se hace necesario la definición de estrategias y procedimientos de análisis que permitan cubrir una lista amplia de objetivos que pueden estar planteados en términos de hipótesis de trabajo específicas.

La idea central de este trabajo es que en el contexto de una encuesta el investigador o investigadores involucrados necesitan de todos los tipos de análisis estadísticos en los diferentes niveles que la metodología se los permita. Se necesita del análisis exploratorio (AE) para verificar la calidad de la información y para compenetrarse en un nivel adecuado de conocimiento de las relaciones dentro y entre las variables $\{X_1,\ldots,X_p,Y_1,\ldots,Y_q\}$ inmersas en el fenómeno y ocultas en la masa de datos; además el AE provee elementos para justificar el uso de modelos estocásticos y garantizar la razonabilidad de los supuestos que los acompañan, aunque tal vez esto podamos referirlo al Análisis Inicial de Datos (AI) (ver Chatfield (1985)). En el contexto inferencial el investigador necesita conducir procesos de inferencia sobre funciones de las observaciones en la población finita, por ejemplo realizar procesos de estimación sobre totales poblacionales, sobre medias, sobre razones y sobre sumas de cuadrados y funciones de éstas. En el caso de la inferencia descriptiva el investigador está interesado en parámetros explícitos en la población finita, y desde luego también tendrá interés en parámetros implícitos en esta población, los cuales se hacen explícitos a través de la postulación de un modelo sobre una población hipotética infinita llamado modelo superpoblacional.

Para realizar el AE el investigador podrá hacer uso de los más diversos procedimientos tanto aritméticos como descriptivos, gráficos y matemáticos, destacando las poderosas técnicas multivariadas como los componentes principales, la correlación canónica, el análisis de correspondencia y las técnicas de agrupación. Solo que en la forma estándar estas técnicas asumen que la muestra es una muestra aleatoria simple, lo cual no se justifica generalmante para el caso de datos de encuestas ya que por lo general dado el diseño o dada la naturaleza de la población muestreada, la muestra presenta una estructura grupal. Así los procedimientos de AE o AI a usar deberán considerar la naturaleza de la muestra y proveer alternativas de uso en diversos contextos y situaciones.

Debido a que la naturaleza de los estudios por muestreo es generalmente multivariada y dado que en la actualidad existen grandes facilidades de cómputo para el análisis de datos, podemos decir que cada vez se hace más frecuente el uso de modelos multivariados más que univariados. Binder

(1983) menciona que a partir de datos de encuestas se hacen con frecuencia análisis de regresión, análisis de discriminación, análisis de tipo probit y logit, así como análisis de tablas de contingencia a través de modelos loglineales. En lo que se refiere a el uso de la metodología de regresión para el análisis de datos de encuestas Pfefferman y Smith (1985) presentan una revisión de las aportaciones más importantes, destacando los trabajos teóricos y metodológicos en los que se demuestra que cuando se procede como si la muestra fuera aleatoria simple cuando en realidad no lo es, se incurre en la sobreestimación del error estándar de las estimaciones. En lo referente al análisis de tablas de contingencia en muestras complejas existen también avances significativos, como lo muestran los trabajos de Rao y Scott (1984, 1987), Hidiroglou y Rao (1987), y el propio Binder (1983) refiriendo los modelos lineales generalizados.

Para la realización de procesos inferenciales hay dos enfoques en el caso de inferencia descriptiva. El llamado enfoque basado en el diseño, que considera que si se observara toda la población finita no restaría ningún elemento de incertidumbre, y por tanto para la conducción de inferencias solo se podrá involucrar la incertidumbre a través del proceso aleatorio usado en la selección de la muestra; es decir, la distribución de probabilidad que se sigue de la consideración particular de p(s/z). El otro enfoque es el denominado enfoque basado en modelos, ó también conocido como enfoque de predicción; aquí se asume que si Y es la variable de interés, entonces esta se puede describir a través de un modelo, el cual le da una naturaleza estocástica; es decir, que a pesar de conocer todos los valores en la población finita persistiría la incertidumbre, pero, como ya señalamos, dado que las poblaciones finitas no son estáticas y que al investigador le interesa la tendencia ó patrón general de comportamiento en la población esto resulta mucho más razonable. El carácter descriptivo de la inferencia en este contexto se preserva en el sentido de que el interés del investigador se centra sobre parámetros explícitos como funciones de las observaciones de la población finita. En algunos casos muy simples los dos enfoques para inferencia descriptiva proveen los mismos resultados.

Asumamos que dada una población finita U en cada una de las unidades se pueden medir un conjunto de variables $\{X_1,\ldots,X_p,\ Y_1,\ldots,Y_q\}$, que produce una serie de datos $d = \{(\mathbf{x}_i,\mathbf{y}_i)\}$; $i=1,2,\ldots,N$. Asumamos que las variables X son consideradas como independientes y las Y como dependientes en algún sentido directamente relacionado con la investigación en cuyo contexto la información de U interesa.

Un modelo superpoblacional ν plantea la relación

$$g_Y(y) = f(x;\|) + e \qquad (1)$$

donde f y g son funciones conocidas y e es una variable aleatoria no observable.

Este modelo debe, de acuerdo a los postulados de la investigación y la naturaleza de las variables medidas en las unidades de la población finita, tener sentido. Debe ser un modelo razonable y que en principio describa con cierta bondad los datos en la población finita; es decir, debe mostrar alguna bondad cuando

$$g_Y(y_i) = f(x_i; q^*) + e_i; \; i = 1,2,...,N \qquad (2)$$

donde q^* es el parámetro del modelo en la población finita.

El modelo superpoblacional describe el comportamiento asociado de funciones de las variables X y funciones de las variables Y en una población infinita o hipotética. Esta población es como todas las poblaciones infinitas que se generan en el análisis estadístico: el modelo de referencia para construir la teoría que permita definir procedimientos de estimación e inferencia en general sobre q y así mismo determinar las propiedades de estos. Dada esa teoría es posible dar fundamentos ó recomendaciones metodológicas (métodos estadísticos) para los problemas en cuestión.

La idea central en el uso de un modelo superpoblacional es en el sentido de que hay un proceso estocástico general que genera bajo esa ley, el modelo superpoblacional, los valores que toman los datos en la población finita. En este sentido la población finita es una realización del proceso estocástico que bajo el modelo (1) produce los datos en el modelo (2). Es decir que bajo este enfoque nosotros rechazamos el postulado central en la teoría clásica del muestreo para poblaciones finitas que se refiere al hecho de que de ser posible conocer todos los elementos en la población se tendría toda la información y no habría incertidumbre. La práctica ha mostrado que a los investigadores no interesa la población finita como tal, sino más bien los rasgos característicos de las leyes estocásticas inmersas en esa población; es decir las tendencias generales (consumos promedio, gasto total, relaciones ingreso gasto, relaciones de gustos y hábitos, etc.); además la población finita es una entidad que siempre tiene movimiento en el tiempo y en el espacio (su naturaleza es dinámica) y por tanto la información que provee en un instante no es la que proveerá en otro. Por tal motivo el suponer que al conocer todos los valores de la población finita ya no habría

incertidumbre es una argumentación infundada y desde el punto general de las ideas centrales o fundamentos filosóficos de la estadística es discordante.

Nótese que los parámetros en el modelo (1) y (2) no son los mismos, dado que nosotros suponemos que (2) está descrito sobre una realización del proceso estocástico gobernado por (1). Dada la naturaleza de este problema podemos postular que $\hat{q}$ es una estimación de máxima verosimilitud de q. En tal sentido podemos pensar que el muestreo sobre poblaciones finitas es un muestreo en dos etapas; en la primera a través de un mecanismo aleatorio no controlado se está generando la población finita, y en el segundo a través de un mecanismo aleatorio controlado se está generando la muestra que el investigador analiza. El aspecto interesante en este sentido es que la componente aleatoria que el investigador determina se puede incluir en el trabajo que se haga con base en el modelo. Adelante precisaremos y abundaremos sobre esto.

Las dos direcciones a que se hace referencia en el primer párrafo de esta sección se determinan por la naturaleza de la inferencia que se hace a partir de la última muestra, la cual puede ser descriptiva o analítica. De hecho todos los casos de interés en el enfoque clásico de muestreo, dado que no existe el modelo superpoblacional, son considerados bajo el enfoque superpoblacional como de inferencia descriptiva. De aquí se podría seguir que el análisis de datos a través de modelos lineales generales o generalizados, se debería considerar como casos de inferencia analítica. En general es así, pero hay ejemplos que siguen un enfoque descriptivo a pesar de que usan este tipo de modelos [Kish y Frankel (1974)].

IV.- ESTRATEGIAS PARA EL ANALISIS DE ENCUESTAS

Una vez que se tienen los datos, generalmente, se piensa que el resto es un problema del Estadístico. Esto debería ser así siempre y cuando la planeación de la realización de la investigación se haga considerando a esta de manera global; es decir, en todas las fases como se enuncia en la introducción de este trabajo. La experiencia que el autor tiene indica que rara vez se planifica el desarrollo completo de las tareas de investigación, conceptualizando sus interrelaciones y evolución retroalimentandose en el todo y en las partes. Es por esto que muchas veces el Estadístico se solicita solo cuando los datos han aparecido; y principalmente cuando las exigencias de la investigación van más alla de la estimación de promedios y porcentajes. Ante esta panorámica el Estadístico debe realizar una actividad de rediseño en lo que se refiere al análisis de la encuesta. En lo que sigue daremos algunas ideas sobre cómo abordar esta problemática.

Consideraremos que la tarea del levantamiento de la encuesta se realizó de manera satisfactoria, y que los datos que se tienen son suficientes y describen representativamente el fenómeno bajo estudio. Hay que señalar que bajo el enfoque superpoblacional, no es tan esencial el esquema probabilístico de selección de las unidades en la muestra, salvo por la representatividad y la no introducción de sesgos por factor humano en la selección. En adelante este punto de vista soporta nuestras ideas.

Una estrategia para el análisis de datos se debe orientar en estrícta concordancia con los objetivos de la invetigación y a partir de las hipótesis a corroborar. El Estadístico debe lograr una claridad suficiente en este nivel para comprender y plantear claramente los objetivos particulares de los análisis estadísticos. Muchas veces los objetivos de los investigadores pueden ser enriquecidos con sugerencias del Estadístico emanadas de los análisis exploratorios y de la verificación de la calidad de la información, pero jamás se debe olvidar que hay una directriz fundamental a seguir: los objetivos generales y la hipótesis de trabajo. Es por esto que la planeación del análisis de los datos se debe realizar con el concurso del équipo de investigación, y hacerse con sumo cuidado y meticulosidad.

La clasificación del tipo de variables, fundamentación del estudio de asociaciones entre variables y la postulación de modelos a referir jamás puede realizarse sin una discusión amplia del Estadístico o Estadísticos con los investigadores. El adecuado nivel de comunicación del Estadístico, del uso de técnicas exploratorias y gráficas y la comprensión clara de los objetivos de la investigación, son elementos fundamentales para la definición de las estrategias de modelaje estadístico y verificación de hipótesis.

V.- EL USO DE MODELOS PARA ANALIZAR ENCUESTAS

Gracias a la proliferación de las facilidades computacionales, dados los datos de una encuesta, se pueden realizar, siempre que tengan adecuada justificación teórica y práctica, los más complejos análisis estadísticos. Así,análisis de regresión, de tablas de contingencia, de modelos loglineales, de correspondencia, de componentes principales, de correlación canónica, y otros, se pueden usar para analizar encuestas.

En la forma estándar, los métodos estadísticos consideran una muestra aleatoria simple, lo cual en el caso de las encuestas es bastante raro. Los procedimientos de análisis y selección de modelos que proveen los paquetes computacionales no consideran muestras complejas y además los fundamentos teóricos para el análisis de datos de este tipo de muestras

es todavía tema de investigación. En el enfoque clásico del muestreo, el artículo de Kish y Frankel (1974) es el punto de partida, y realmente en esa dirección pocos avances se han logrado (ver Smith (1984)). En el enfoque superpoblacional múltiples aportes se han hecho. Pfefferman y Smith (1985) tratan abundantemente este tema en lo que respecta a los modelos de regresión. En el caso del análisis de datos categóricos Binder y otros (1985) plantean una revisión de los aportes y tendencias recientes. Sin embargo debemos señalar que en la práctica, con bastante frecuencia, los análisis de datos de encuestas que se hacen utilizando métodos de los mencionados, incurren en la no consideración de la muestra compleja, dado, principalmente, en razón de que los paquetes estadísticos más conocidos no permiten considerar esto.

Los métodos multivariados exploratorios, como lo son los componentes principales y la correlación canónica se pueden usar generalmente en el análisis de encuestas, pero siempre es recomendable considerar la estructura grupal de la muestra en el momento de ejecutar los análisis. Aspectos más profundos del modelaje a través del modelo lineal general y los modelos generalizados se deben abordar solo bajo la recomendación de expertos en el tema, ya que los desarrollos metodológicos no son aún del conocimiento generalizado. Smith (1984) hace una revisión en este sentido.

Recientemente varios paquetes computacionales que incorporan el uso de técnicas para análizar encuestas se han elaborado. El más conocido y más completo de estos sistemas es el SUPERCARP producido por Fuller y otros investigadores en la Universidad norteamericana de Iowa. La revisión del manual del usuario de este sistema permite determinar los conocimientos necesarios para la realización del modelaje estadístico bajo la consideración de datos de encuestas. El conocimiento de la teoría del modelo lineal y los modelos generalizados es fundamental.

Por todo lo expuesto se infiere que es sumamente necesario recurrir a especialistas de esta área desde la fase misma de la planeación de la investigación y hasta la fase de la planeación de los análisis estadísticos. Esto hace mucho sentido en razón de la amplia gama de problemas planteada en la primera sección de este escrito, y que está presente en cada encuesta.

VI.- CONCLUSIONES

El análisis de datos de encuestas es una especialidad en la que el Estadístico en vinculación bastante estrecha con investigadores sociales puede acceder a niveles de investigación, teórica y aplicada, para resolver múltiples problemas relacionados con la implementación de los métodos estadísticos sobre los datos que ha producido una encuesta.

Las encuestas son costosas y los investigadores y el Estadístico deben optimizar al máximo la cantidad de información y de conclusiones que de esta se obtengan. Han pasado ya los tiempos en los que de una encuesta sólo se obtenían frecuencias, totales y porcentajes por variable. El uso de métodos estadísticos multivariados y de modelaje estocástico permite maximizar la obtención de información y dar mayor fuerza a las conclusiones, pero su adecuada implementación requiere no sólo de paquetes computacionales sino de conocimientos teóricos y metodológicos profundos.

El enfoque superpoblacional permite considerar a la inferencia y modelaje con propósitos analíticos en el contexto de muestras complejas, como problemas particulares del modelaje estadístico. Sin embargo, es necesario que los Estadísticos dedicados a esta área pongan mayor atención sobre el uso adecuado de los paquetes computacionales, ya que la inferencia en estos casos requiere procesos especiales (ver Pfefferman y Smith (1985)). Sería recomendable que esta problemática fuera discutida en los cursos tradicionales de muestreo y de métodos estadísticos avanzados

En el terreno teórico esta área presenta una serie de problemas que aún se consideran objeto de investigación, por ejemplo en lo relacionado al análisis de datos discretos, la consideración de los llamados modelos generalizados es bastante reciente asumiendo muestras en grupos (ver Binder (1983)).

AGRADECIMIENTOS

Agradezco de manera especial los comentarios y sugerencias que sobre una versión preliminar del presente me turnara el Dr. Nicolás Hernández Guillén. Fueron asimismo valiosas las observaciones y consideraciones de un árbitro anónimo. En suma estas aportaciones lograron clarificar y mejorar sustancialmente este artículo, cuyas deficiencias son, desde luego, responsabilidad únicamente mía.

BIBLIOGRAFIA

Chatfield C (1985) The initial examination of data; Jour. of Roy. Stat. Soc. series A vol. 148 p. p. 214-253.

Cochran C. W. (1979) Técnicas de Muestreo. CECSA.

Binder D. A. Gratton M., Hidiroglou M. ., Kumar S. and Rao J. N. K. (1984) Analysis of categorical data from surveys with complex design; Survey Methodology vol. 10 p. p. 141-156.

Godambe V. P.(1955) A unified Theory of sampling from finite

populations; Jour. Roy Stat. Soc. B; vol. 17, p. p. 269-278.

----------(1960) An optimum property of regular maximum likelihood estimation; The Ann. Math. Stat. vol. 31, p. p. 1208-11.

Hidiroglou M. A. and Rao J. N. K. (1987) Chi-squared test with categorical data from complex surveys (Two parts); Jour. Off. Stat. vol. 3 p. p. 117-140.

Jessen M. (1979) Samplig Surveys Methods; Wiley.

Kish L. and Frankel M. R. (1974) Inference from complex samples; Jour. Roy Stat. Soc. series vol. 36 p. p. 1-37.

Pfefferman D. and Smith. T. M. F. (1985) Regresion models for grouped populations; Int. Stat. Rview vol. 3 p. p. 37-59.

Rao J.N.K. and Scott A. J. (1981) The analysis of categorical data from complex surveys; JASA, vol. 76 p. p. 221-230.

----------(1984) On Chi-squared tests for multiway contingence tables with cell proportions from survey data; Ann. of Stat. vol. 12, p. p. 46-60.

Rodrigues J., Bolfarine H. and Rogatko A. (1985) A general theory of prediction in finite populations; Int. Stat. Review vol.53 p. p. 239-254.

Smith T. M. F. (1984) Present position and potencial developments:some personal views about sampling surveys; Jour. Roy. Stat. Soc., series A vol. 147 p. p. 208-227.

PROPOSICION PARA LA ESTIMACION EN MUESTREO DE POBLACIONES FINITAS.

Ramírez Valverde Gustavo . Castillo Morales Alberto
Colegio de Postgraduados
Centro de Estadística y Calculo
Chapingo, Méx. c.p. 56230

Introducción

La estimación en muestreo de poblaciones finitas, se ha apoyado desde sus inicios en los procedimientos de inferencia que se usan en poblaciones infinitas con distribuciones hipotéticas. Al no tener el mismo fundamento teórico esas distribuciones que las que ocurren en poblaciones finitas, surgen dudas sobre la aplicación de algunos procedimientos típicos de poblaciones hipotéticas, para la inferencia sobre poblaciones finitas.En la actualidad ya se han dado algunos criterios o algunos supuestos que tratan de detectar los factores de optimización en las estimaciones en muestreo de poblaciones finitas (Godambe 1966 y 1969 ,Godambe y Jhosi 1965, Jhosi 1965 y 1966; Hartley y Rao 1968, Roy y Chavarati 1960, Casel et. al. 1977 y Särndal 1976), sin embargo, estos criterios no son del todo satisfactorios, permaneciendo vigente el problema de estimación en poblaciones finitas.

Como una alternativa se ha propuesto por un lado el enfoque de escalas ponderadas (Hartley y Rao, 1968 y Royal, 1968), que permite realizar inferencia basada en la función de verosimilitud. Por otro lado se ha dado el enfoque de superpoblaciones, que contempla a la población finita como

una realización de una variable N-variada que a su vez constituye una población infinita (Cassel et. al., 1977).

Notación y Definiciones

El punto de partida del muestreo es la definición de una población finita, como el conjunto U de N individuos, donde N es un número finito; al conjunto U se le llama población finita y al número N se le conoce como el tamaño de la población.

Así, la población finita U estará constituida por N unidades u_i, con i = 1,2,3,...,N, donde cada unidad u_i está perfectamente identificada, pudiendo representar la población como: $U = \{ u_i \mid i = 1,2,3,...,N \}$.

Si x es la variable de interes, a cada unidad u_i de la población estará asociado un valor real x_i, de tal forma que x_i es el valor que toma la variable X en la unidad u_i.

Al identificar a los individuos de una población y observar que tienen un valor x_i asociado a cada uno de ellos, se puede construir una función h que relacione a la población U con un vector x de valores $(x_1,x_2,...,x_N)$ para cada variable que alli se estudie. Al vector x se le conoce como parámetro de la población finita U; es decir, $h(U)=x=(x_1,x_2,...,x_N)$, donde x es el parámetro de la población U.

El vector x es un punto del espacio N-dimensional euclideano. Se denotará por Ω al espacio donde se encuentran los valores posibles de x y se le llamará espacio

paramétrico. Es frecuente que se considere Ω = RN, donde RN es el espacio N-dimensional euclideano; sin embargo, Ω puede estar formado por un subespacio de RN. A cualquier función real sobre el parámetro $x=(x_1,x_2,\ldots,x_N)$ se le llamará función paramétrica.

El uso más generalizado del muestreo de poblaciones finitas es para inferir con respecto a algunas funciones paramétricas y no sobre la población en sí.

Se llamará muestra a un subconjunto s no vacio de la población finita U; al número n de elementos de la muestra se le llama tamaño de la misma.

Al conjunto de todas las muestras posibles en una población se le denota por Γ. Con éstos elementos se puede construir el concepto de diseño de muestreo como una función p en Γ, el conjunto de todos los subconjuntos s de U, tales que: $p(s)\geq 0$, y $\Sigma p(s)= 1$.

Enfoque de Escalas Ponderadas

La no existencia de un único estimador lineal e insesgado y el hecho de que la función de verosimilitud es constante para la muestra en todo Ω consistente con la muestra s, limita la inferencia basada en ella. Esto motivó que Hartley y Rao (1968) propusieran el esquema de escalas ponderadas (Royal, en 1968 llega a resultados similares independientemente). Suponen que los elementos x_i con $i=1,2,\ldots,N$, son medidos en una escala discreta y que el conjunto de valores en la escala discreta que pueden tomar

los x_i es menor que N.

De esta forma, la población está caracterizada por t valores x_i', con i=1,2,...,t, donde t es el número de valores discretos distintos que pueden tomar los x_i del parámetro x. El número de veces que ocurre cada x_i en la población se denotara por N_i con i=1,2,...,t, de tal forma que $\sum_{i=1}^{t} N_i = N$.

Por tanto, se tiene que la información sobre el parámetro $x=(x_1,x_2,...,x_N)$ se reduce a dos vectores: $y=(N_1,N_2,...,N_t)$ y $b=(x_1',x_2',...,x_t')$, donde x_i' es el valor que toma la i-ésima escala discreta y N_i es el número de valores x_i' presentes en el parámetro x.

El esquema de muestreo de escalas ponderadas, considera una estructura para la población muestreada, esto es, supone conocidos t valores $x_1',x_2',...,x_t'$, que son las escalas discretas, y supone desconocido el número N_i de los elementos de la población en cada una de las escalas discretas, es decir, el vector $y=(N_1,N_2,...,N_t)$.

Lo importante de este enfoque es que se producen resultados sobre la estimación insesgada, además de que se tiene una función de verosimilitud que no es uniforme, por lo que la inferencia puede estar basada en ella. Para ejemplificar la forma de la inferencia en este enfoque, se utilizará el muestreo simple aleatorio.

Bajo este enfoque, para el muestreo simple aleatorio se tiene tiene que si n_i con i=1,2,...,t es el número de veces que aparece en la muestra s el valor de la i-ésima escala

discreta, donde $\sum_{i=1}^{t} n_i = n$; entonces, la probabilidad de obtener los datos muestrales, sin considerar la etiqueta para escalas iguales, $d = \{n_1, n_2, \ldots, n_t\}$ forma una distribución hipergeométrica para t clases.

La función de verosimilitud para los datos muestrales sin etiqueta x_d es:

$$L(y/d) = \begin{cases} \dfrac{\prod_{i=1}^{t} \left(\binom{N_i}{n_i}\right)}{\left(\binom{N}{n}\right)} & \text{para todo } y \in M \\ 0 & \text{de otra forma} \end{cases}$$

donde, $M = \{y = (N_1, N_2, \ldots, N_t) \mid N_i$ es entero positivo; $N_i \geq n_i$ y $\sum_{i=1}^{t} N_i = N\}$

Claramente se puede observar que esta función de verosimilitud no es constante para los vectores $y = (N_1, N_2, \ldots, N_t)$. Se puede buscar a los valores $\hat{N}_i$ con $i = 1, 2, \ldots, t$ que maximizan $L(y/d)$, los cuales no siempre son únicos, estos son los estimadores de máxima verosimilitud de $y = (N_1, N_2, \ldots, N_t)$.

Como puede verse, dada la estructura originalmente propuesta, la media poblacional puede reescribirse como:

$$\bar{X} = \sum_{i=1}^{t} N_i \, x_i' / N$$

donde, x_i' es el valor de la i-ésima escala discreta; de aquí que se pueda construir el estimador de máxima verosimilitud para la media poblacional al sustituir los estimadores $\hat{N}_i$ en lugar de los parámetros N_i, quedando de la siguiente forma:

$$\hat{\bar{X}} = \sum_{i=1}^{t} \hat{N}_i \; x_i/N,$$

Cuando N/n es un entero la función de verosimilitud L(y/d) es maximizada por $\hat{N}_i = Nn_i/n$, por lo que, la media x de máxima verosimilitud coincide con la media muestral $\bar{x} = \sum_A x_i/n$. Cuando N/n no es un entero los $\hat{N}_i$ que maximizan la función de verosimilitud $L_{x_d}(y)$ se encuentran alrededor de los números Nn_i/n, Hartley y Rao (1969) dan un algoritmo para encontrar $\hat{N}_i$, cuando N/n no es un entero.

Enfoque de Superpoblaciones

En este enfoque la poblacion finita es considerada como una realización de una variable X N-variada, por lo que cada elemento X_i con i=1,2,...,N es una variable aleatoria. Es claro que el parámetro poblacional (de la población finita) $x=(x_1,x_2,\ldots,x_N)$ es una ocurrencia de la variable aleatoria $X=(X_1,X_2,\ldots,X_N)$.

A la población hipotética generada por las posibles realizaciones de la variable aleatoria se le llama superpoblación.

Como la muestra de tamaño n nos da n valores exactos del vector x, el problema se reduce a predecir los N-n valores no conocidos.

En este enfoque se supone que la distribución de la superpoblación $X=(X_1,X_2,\ldots,X_N)$ es conocida o que esta depende de pocos parámetros desconocidos. Entonces, la estimación de funciones parámetricas como el total y la media de la población finita, se hará considerando las predicciones

de los valores $x_i \notin s$. Así para la media, su estimación será:

$$\hat{\bar{X}} = [\sum_{A} x_i + \sum_{B} \hat{x}_i]/N = \sum_{A} x_i/N + \sum_{B} \hat{x}_i/N$$

donde, $A=\{i|u_i \in s\}$, $B=\{i|u_i \notin s\}$ y $\hat{x}_i$ es el predictor de x_i basado en la distribución de la superpoblación.

Por ejemplo, si se supone que la superpoblación $X=(X_1,X_2,\ldots,X_N)$ está constituida de N variables aleatorias independientes con distribución normal con media μ y varianza σ^2, con μ y σ^2 desconocidos, se puede pensar en predecir los valores $x_i \in \{x_i|u_i \notin s\}$ por medio de la media poblacional de X, por lo que el problema se reduce a estimar la media poblacional de una variable X con distribución normal con media μ y varianza σ^2.

Seleccionando una muestra cualquiera de la población finita $x=(x_1,x_2,\ldots,x_N)$, sin importar la aleatorización, tendremos una muestra aleatoria de tamaño n de una distribución normal con media μ y varianza σ . De la muestra se obtiene $\bar{x}$ que es el mejor valor para predecir futuros valores de x, asi, el mejor estimador de la media de la población finita será:

$$\hat{\bar{X}} = [\sum_{A} x + \sum_{B} \hat{x}_i]/N = [n\bar{x}+(N-n)\bar{x}]/N = \bar{x} = \sum_{A} x_i/n$$

donde, $A=\{i|u_i \in s\}$ y $B=\{i|u_i \notin s\}$.

Resulta interesante observar que bajo este tipo de enfoque, de acuerdo al modelo supuesto, la aleatorización puede ser irrelevante, incluso, en algunos casos resulta mejor el muestreo dirigido. Por ejemplo, si se supone (correctamente) un modelo de regresión lineal, se tendrá que

la mejor estimación se realizará muestreando en los puntos de ambos extremos de la recta.

El problema de este enfoque es que la inferencia va a estar en función de que tan adecuado y cercano a la realidad esté el modelo que se suponga para la distribución de $X=(X_1,X_2,\ldots,X_N)$. Por ejemplo, en el caso anterior el supuesto de independencia es muy difícil que se de en la realidad. Los fenómenos reales son complejos, por lo que suponer un modelo más o menos adecuado (suponiendo un amplio conocimiento del fenómeno) para la superpoblación resulta complejo. Esto no significa que en algunas situaciones no podamos suponer un modelo realista que no sea muy complejo.

Enfoque Castillo-Ramírez

Si se tiene una población finita $U=(u_1,u_2,\ldots,u_N)$ y la variable de interés x es real, se tiene cierta estructura que se puede representar por una función escalonada F, de tal forma que si $x_{(1)},x_{(2)},\ldots,x_{(N)}$ son los valores de x ordenados, se tiene: $F(x)=(1/N)\sum_{i=1}^{N} I_{[x\geq x_{(i)}]}$

$$F(x) = \frac{\text{Número de valores } x_{(i)} \text{ menores o iguales a } x}{N}$$

Esta función F es una función desconocida que se asemeja en su construcción y ecuación a una función de distribución empírica y su forma define la población en forma exacta. Nótese que como no se conocen los valores $x_1,x_2,\ldots,x_N$, esta función F es desconocida. Además, F cumple con las condiciones teóricas de una función de distribución.

El objetivo en el esquema propuesto es utilizar la función de distribución muestral F_n para estimar a la función de distribución de la población. La función de distribución muestral F_n se construye utilizando los estadísticos de orden de la muestra, asi, si la muestra es $s=\{x_1,x_2,\ldots,x_n\}$, y los estadísticos de orden son $x_{(1)},x_{(2)},\ldots,x_{(n)}$, entonces:

$$F_n(x)=(1/n)\sum_{i=1}^{n} I_{[x\geq x_{(i)}]}$$

$$=\frac{\text{Número de valores } x_{(i)} \text{ menores o iguales a } x}{n}$$

Las propiedades de la función de distribución muestral F_n como estimador de la función de distribución poblacional deben verse con cuidado, ya que los términos usuales pueden llevar a errores de interpretación. Puede pensarse en el criterio de consistencia en error cuadrático medio (vease Mood, Graybil y Bose, 1983), definido por.

$$p\left[\sup |F_n-F| \xrightarrow[n\to N]{} o\right] = 1.$$

A menos de que se suponga un esquema de aleatorización no resulta claro sobre que probabilidad se calcula lo anterior, aunque si es claro que: $|F_n-F| \longrightarrow 0$ cuando $n\longrightarrow N$.

En este enfoque se observa nuevamente la relevancia de obtener muestras típicas (Casian y castillo en prensa), ya que si la muestra es típica F_n se parecerá a F. Conciderando que no se conoce a la población, se puede observar que el uso de la aleatorización, si bien no garantiza la ocurrencia de una muestra típica, si aumenta la probabilidad de que esta ocurra.

La media poblacional se puede escribir en función de la

función de distribución F de la población en la siguiente forma: $E(x) = \bar{X} = \int_0^{\infty} [1-F(x)]dx - \int_{-\infty}^{0} F(x)dx$.

Por tanto un estimador de la media poblacional utilizando el estimador dado por la función de distribución muestral F_n en la ecuación anterior, queda de la siguiente forma:

$$\hat{\bar{X}} = \int_0^{\infty} [1-F_n(x)]dx - \int_{-\infty}^{0} F_n(x)dx \ .$$

El valor del estimador $\hat{\bar{x}}$ coincide con el valor de la media muestral. La importancia de este nuevo enfoque en el muestreo de poblaciones finitas, es que bajo este esquema de muestreo de poblaciones finitas se tiene un mayor y mejor conocimiento sobre la población, permitiendo los criterios usuales en muestreo finito clásico, y la busqueda de nuevos criterios que deberán definirse en el futuro.

Conclusiones

De acuerdo a lo anteriormente discutido se pueden obtener las siguientes conclusiones:

1.- Bajo el enfoque de escalas ponderadas se puede construir una función de verosimilitud que no es uniforme, sin embargo, presupone conocidos los valores distintos presentes en el parámetro x, lo cual puede ser poco realista.

2.- El enfoque de superpoblaciones propone que el parámetro poblacional x es una realización de una variable X N-variada que a su vez conforma una superpoblación infinita, de la cual se supone conocida su distribución o que ésta depende de pocos parámetros desconocidos. De esta forma la inferencia

esta basada en poblaciones infinitas, sin embargo, la inferencia va a estar en función de que tan adecuado y cercano a la realidad sea el modelo que se suponga a la distribución de la superpoblación.

3.- En muestreo de poblaciones finitas bajo el enfoque de superpoblaciones, la aleatorización puede ser irrelevante, incluso, en algunos casos resulta mejor el muestreo dirigido.

4.- En el enfoque propuesto, se tiene que la inferencia obtenida con una muestra s es la misma independientemente de la forma en que se obtuvo la muestra, lo que es congruente con las condiciones que presenta la función de verosimilitud de las poblaciones finitas. De aqui que lo más importante en muestreo de poblaciones finitas sea obtener una muestra lo más típica posible.

Literatura Citada

Cassian M.A. y A.M.Castillo (en prensa) El muestreo en los estudios de la agricultura; reflexiones acerca de sus fundamentos, Agrociencia, México.

Cassel C., C.Sarndal y J.Wretman (1977) Foundations of inference in Survey Sampling, John Wiley & Sons, New York, EUA.

Godambe V.P. (1955), A unifred Theory of sampling from finite populations, J.R.Statist.Soc.B,17,268-278.

Godambe V.P.(1966), A new approach to sampling from finite populations, I. J.R. Statist.Soc B 28:310-328.

Godambe V.P.(1969), Some aspects of the theorical

developments in Survey-Sampling, en N.L.Johnson y H. Smith eds., New. develoments in survey sampling, New York, Wiley, 27-59.

Godambe V.P. y V.M.Jhosi (1965), Admissibility and Bayes estimation in sampling finite populations I, Ann. Math. Statist, 36:1707-1722.

Hartley, H.O. y J.N.K.Rao (1968), A new estimation theory for sample surveys, Biometrika, 55:547-557.

Jhosi V.M. (1965), Admissibility and Bayes estimation in sampling finite populations II, Ann. Math. Statist., 36:1723-1729.

Jhosi V.M. (1966), Admissibility and Bayes estimation in sampling finite populations III, Ann.Math.Statist., 37:1658-1670.

Mood A.M. F.A.Graybill D.C.Boes (1983), Introduction to the theory of statistics, Mc.Graw-Hill, tercera edición, U.S.A..

Roy J. y J.M.Chakravarti (1960) Estimating the mean of a finite population, Ann.Math.Statist.,31:392-398.

Royall R. (1968) An old approach to finite population sampling theory, J. Amer Statist.Asoc., 63: 1269-1279.

Särndal C.E. (1976), On uniformly minimum variance estimation in finite populations, Ann.Statist., 4:993-997.p

VIC-SITEM, UN SISTEMA PARA CALCULAR TAMAÑOS DE MUESTRA Y ESTIMADORES EN ESTUDIOS POR MUESTREO.

Victor Serrano Altamirano; Gilberto Rendón Sánchez;
Graciela Bueno de Arjona; Vicente González Romero.
Colegio de Postgraduados, Centro de Estadística y Cálculo,
Chapingo, Méx.

RESUMEN

Con el propósito de que el usuario del muestreo estadístico disponga de un procedimiento que le ayude a resolver los problemas del cálculo y la selección de una muestra, así como la estimación de parámetros de una población, se desarrolló un sistema interactivo, programado en Turbo Pascal versión 5.0 para microcomputadoras IBM-PC y compatibles. El sistema está dividido en unidades y programas ejecutables, los cuales son enlazados por un programa principal. La salida de información puede dirigirse a pantalla, disco o impresora. Por su organización, puede usarse con facilidad y con pocos conocimientos sobre computación y teoría del muestreo.

INTRODUCCION

Uno de los aspectos más importantes del muestreo es la determinación del tamaño de muestra que permita obtener estimaciones de los parámetros poblacionales, con la mayor precisión y confiabilidad posibles y al menor costo.

Para conseguir ésto, existen diversos procedimientos dependiendo de la información y recursos con que se cuente y el esquema de muestreo que se utilice. Dichos procedimientos se encuentran en diferentes fuentes, por lo que no son fácilmente accesibles para el usuario común, y no son fáciles de utilizar por el investigador que no está familiarizado con la teoría del muestreo.

Una vez conocido el tamaño de muestra, la selección y listado de las unidades de muestreo puede constituir un problema, ya que no siempre se tiene clara la metodología que se debe seguir, particularmente en los diseños no tan conocidos como son el muestreo por conglomerados en una o más etapas.

El cálculo de los estimadores una vez capturada la información, es otro problema al que se tiene que enfrentar el usuario del muestreo; puede suceder que no conozca la mecánica para obtener el estimador, o que por la cantidad de datos le resulte laboriosa la estimación de los parámetros.

Ante la problemática anterior, se pensó en desarrollar un sistema de cómputo que permitiera al usuario, la aplicación de los esquemas de muestreo de mayor utilización de una manera sencilla y eficiente, que evite los errores propios de la falta de comprensión de los mecanismos de cálculo involucrados en las distintas etapas del muestreo y que sea de fácil uso.

VIC-SITEM fue desarrollado con los objetivos anteriores, escrito en TUR-

BO PASCAL versión 5.0 para microcomputadoras IBM-PC y compatibles.

PRESENTACION DEL SISTEMA

VIC-SITEM es un sistema completamente interactivo que opera a base de menús que van guiando al usuario por la variedad de alternativas con que cuenta.

Su operación es tan sencilla que no requiere de un manual de uso, por lo que únicamente se describirán sus facilidades.

El sistema maneja los esquemas de muestreo:

- Simple aleatorio
- Estratificado aleatorio
- Sistemático
- Conglomerados en una etapa
- Conglomerados en dos etapas

Para cada uno de estos esquemas se puede:

1) Construir un marco de muestreo.
2) Agregar o modificar la información del marco de muestreo, o los datos obtenidos de la muestra.
3) Calcular para la estimación de una media, un total, una proporción, o en la estimación de razón, el tamaño de la muestra definitiva, ba-

jo suposiciones de normalidad o no-normalidad en la distribución del estimador, o bien, utilizando procedimientos óptimos que requieren funciones de costo.

4) Seleccionar y listar las unidades de una muestra preliminar, o de una muestra definitiva.

5) Obtener las estimaciones de los parámetros: media, total, proporción o razón, usando el método directo o los indirectos de razón o regresión.

REQUERIMENTOS DEL SISTEMA

Para hacer uso del sistema VIC-SITEM se requiere:

1) Una microcomputadora IBM-PC o compatible, con al menos 256 Kb de memoria RAM.
2) Un monitor a color o monocromático.
3) Una impresora, sólo en el caso de que se quieran en papel algunos o todos los resultados de la sesión de trabajo.
4) Dos discos flexibles que tengan el sistema VIC-SITEM o disco duro que lo aloje.
5) Uno o más discos flexibles de trabajo o disco duro, dependiendo de la cantidad de información a manejar y los archivos a grabar.
6) Sistema operativo MS-DOS o PC-DOS versión 2.11 en adelante.

OPERACION DEL SISTEMA

El sistema VIC-SITEM puede trabajarse con discos flexibles, pero resulta más cómodo si se instala en disco duro, para lo cual se debe crear un subdirectorio que aloja al sistema con el comando:

1) md vic

   Entrar al subdirectorio con el comando:

2) cd vic

   Insertar el disco que contiene el sistema en el drive A y copiar los programas ejecutables con el comando:

3) COPY a:*.exe c:

   Hecho lo anterior se puede iniciar la sesión de trabajo tecleando:

4) vic

   Una vez instalado el sistema sólo se requiere los pasos 2 y 4 para ejecutar el sistema.

Si no se instala en disco duro, hay que insertar el disco en la unidad A, instalar la unidad de trabajo como la unidad A, tecleando A: y entrar al sistema escribiendo la palabra VIC.

El disco deberá estar desprotegido pues el sistema en ocasiones intentará escribir archivos en el disco.

Una vez iniciada la sesión con el sistema VIC-SITEM una serie de menús guiarán al usuario por las distintas alternativas con que cuenta.

En el primer menú o menú principal se presentan las opciones:

Conocer conceptos básicos

Muestreo simple aleatorio

Muestreo sistemático

Muestreo estratificado aleatorio

Muestreo por conglomerados en una o dos etapas.

Con la opción de conceptos básicos se despliega al usuario aquellos conceptos básicos que debe conocer para utilizar adecuadamente los esquemas de muestreo y el sistema VIC-SITEM, como son los conceptos de población, precisión, confiabilidad, etc.

Dependiendo de la opción elegida y que representa el esquema de muestreo que se va a utilizar, el sistema presentará otro menú, en el que como opciones se tendrá:

1) Conocer conceptos básicos.
2) Marco de muestreo.
3) Selección de una muestra.
4) Manejo de información.
5) Calcular tamaño de muestra.
6) Cálculo de estimadores.
7) Conocer precisión o confiabilidad.

Con la opción de conceptos básicos se presenta una descripción del es-

quema de muestreo utilizado y sé despliegan aquellos conceptos cuya comprensión es importante cuando se utiliza ese esquema de muestreo.

Con la opción de marco de muestreo se puede captar, revisar, corregir e imprimir el marco de muestreo de una población. El marco de muestreo consta únicamente de tres campos alfanuméricos.

Con la opción selección de una muestra, se puede seleccionar una muestra preliminar, o una definitiva. Cuando la muestra es preliminar, el sitema proporciona las unidades a muestrear que constituyen toda la muestra. Si la muestra es definitiva, sólo lista aquellas unidades que adicionadas a las de la muestra preliminar constituirán la definitiva.

Si se ha capturado el marco de muestreo, lista la información general de las unidades seleccionadas; si no es así, entonces únicamente lista los números de las unidades en el marco que se incluirán en la muestra. Cuando se selecciona la muestra definitiva, le solicitará el tamaño de la misma, el cual se puede obtener previamente con la opción 5 o con otro método.

La opción (5) puede tener otros sub-menús de opciones, como es el caso en el esquema de muestreo estratificado, o en el de conglomerados.

Con la opción manejo de información se graba, corrige, despliega o aumenta la información que se usará en el cálculo de un tamaño de muestra, o en el cálculo de los estimadores.

Esta opción se debe ejecutar cuando ya se tiene la información de una muestra preliminar y se desea conocer el tamaño de muestra definitivo, o bien, cuando se va a hacer la estimación preliminar o definitiva de parámetros, para lo cual se pueden usar los métodos directo, indirecto o de proporción.

Dependiendo del método que se use en la estimación y en el cálculo del tamaño de muestra se solicitarán los datos de 1 o de 2 variables.

Con la opción calcular tamaño de muestra, se calcula un tamaño de muestra para la estimación de los parámetros: media, total, razón o proporción; suponiendo normalidad en el estimador o sin esta suposición; con precisión y confiabilidad dados o con precisión relativa. Se selecciona la muestra y se listan las unidades correspondientes.

Para este proceso se utilizan los datos obtenidos en el muestreo preliminar, los cuales deberán estar en un archivo creado con la opción (4).

En los muestreos estratificado y por conglomerados, aparecerán otras opciones para el cálculo del tamaño de muestra.

Con la opción cálculo de estimadores, se obtiene el estimador puntual del parámetro solicitado (media, total, razón, proporción) e intervalos de confianza al 90, 95 y 99%.

Los datos para el cálculo deberán estar en un archivo creado con la opción (4).

Con la opción conocer precisión o confiabilidad, se puede conocer la precisión que se logra, para una confiabilidad y un tamaño de muestra fijos; y viceversa, conocer la confiabilidad, dada una precisión deseada y un tamaño de muestra.

Algunos esquemas de muestreo tienen una opción más que es la de análisis de varianza. Asimismo, cada opción tiene otros menús de acuerdo al esquema de muestreo y la opción elegida, lo cual lleva al usuario por la gama de alternativas que se pueden presentar.

BIBLIOGRAFIA

Alvarez, C. V. M. (1988). Tamaño de muestra: procedimientos usuales para su determinación. Tesis de M. C., Centro de Estadística y Cálculo, Colegio de Postgraduados. Chapingo, Méx.

Borland International (1988). Turbo pascal reference guide versión 5.0. Borland International. Scotts Valley, California, U. S. A.

Cochran, W. G. (1977). Sampling techniques. Third edition. John Wiley and Sons., Inc. New York, U. S. A.

Dale, N. y Orshalick. (1986). Pascal. Trad. del inglés por José M. Troya. Mc Graw-Hill. México, D. F.

Jamsa, K. y Nameroff, S. (1986). Turbo pascal programmer's library. Osborne Mc Graw-Hill. Berkeley, California, U. S. A.

Rendón, S. G. y González, R. V. (1989). Tamaño de muestra: una alternativa para su determinación con extensión a estudios con propósitos multiples. Serie de Comunicaciones en Estadística y Computo. Colegio de Postgraduados (En prensa).

Serrano, A. V. (1989). VIC-SITEM: un sistema para calcular tamaños de muestra y estimadores en estudios por muestreo. Tesis de M. C. Centro de Estadística y Cálculo. Colegio de Postgraduados. Chapingo, Méx.

AGRUPAMIENTO ESTADISTICO DE HELECHOS FOSILES.

Gustavo J. Valencia.
Departamento de Matemáticas
Facultad de Ciencias, UNAM.
Ciudad Universitaria, México D.F., México.
Tel.: (915)-548-51-65.

RESUMEN.
Se realiza una agrupación de helechos fósiles, mediante análisis Cluster, utilizando distancias entre centroides y distancias Euclideanas. Los datos corresponden a ejemplares fósiles que pueden clasificarse como fértiles, estériles y otros con características tanto fértiles, como estériles (EF) de helechos del orden Marattiales. Los datos provienen de la Formación Matzitzi de la región de Tehuacán, Puebla. Además de comentar sobre la agrupación de helechos, se discuten aspectos generales de la aplicación de los procediemientos estadísticos multivariados.

INTRODUCCION.

Este trabajo presenta parte del análisis estadístico de los datos proporcionados por el Dr. Reinhard Weber (Instituto de Geología) y la Bióloga Susana Magañon (Facultad de Ciencias) de la UNAM.

Los datos corresponden a ejemplares fósiles que pueden clasificarse como fértiles, estériles y otros con características tanto fértiles, como estériles (se denotan mediante EF) de helechos del orden Marattiales. Los datos provienen de la Formación Matzitzi de la región de Tehuacán, Puebla.

Con base en estudios paleobotánicos realizados por Silva (1970) se ha asignado una edad Pensilvánica a la Formación Matzitzi, confirmándose esta edad en 1978 (Carrillo Martínez y colaboradores). Estudios geológicos recientes, aún sin publicar, han mostrado la necesidad de revisar y precisar la edad de esta formación. Para ésto los datos paleobotánicos son lo más indicado.

ANTECEDENTES.

En general, el orden Marattiales comprende 6 géneros y cerca de 100 especies vivientes. Los helechos Marattiales surgen en el Carbonífero (hace 345 millones de años) y la enorme cantidad de especimenes fósiles encontrados da una idea de la diversidad de especies pertenecientes a este orden.

Uno de los géneros más conocidos es *Psaronius*, con un rango geológico que va desde el Carbonífero Superior (Pensilvánico, 310 millones de años) hasta el Pérmico (230 millones de años). El nombre que recibe el tipo de follaje estéril presente en las frondas de *Psaronius* es *Pecopteris*. Este género tiene un rango geológico que cubre

el periodo del Pensilvánico Inferior (posiblemente Mississípico, 325 a 310 millones de años) hasta el Pérmico (230 millones de años).

Cuando las pínulas en material en compresiones están en estado fértil son asignadas al género *Asterotheca*. Las pínulas de este genero son parecidas morfológicamente a las de *Pecopteris*, aunque no se pueden observar detalles de la venación, pues los organos reproductores, llamados sinangios, se sitúan directamente sobre la lamina de la hoja, cubriendo la venación.

Considerando lo discutido anteriormente y el hecho que resulta muy raro encontrar ejemplares fósiles en los que aparecen en una misma pina, pínulas estériles y fértiles, resulta que no ha sido posible para los paleobotánicos hacer asociaciones certeras de las especies de *Pecopteris*, con sus correspondientes formas fértiles (*Asterotheca*).

OBJETIVOS.

En este estudio se persiguen varios objetivos, principalmente: Determinar las especies pertenecientes a los géneros *Pecopteris* y *Asterotheca*, asignar a cada especie de *Pecopteris* su forma correspondiente en estado fértil y determinar la edad precisa de la Formación Matzitzi.

VARIABLES.

La información disponible fue obtenida por el Dr. Weber y la Biól. Magañon, utilizando los ejemplares de *Pecopteris* y *Asterotheca* que se encuentran en las colecciones paleobotánicas del Instituto de Geología y del Museo de Paleontología de la Facultad de Ciencias de la UNAM, cuyo sitio de colección era conocido precisamente, además del material que obtenido en recolecciones directas.

Las variables consideradas para este estudio son aquellas utilizadas tradicionalmente para describir y delimitar especies de helechos. Los fósiles elegidos para el estudio son aquellos que mostraban las características de interés en forma clara.

Para los distintos tipos de ejemplares las variables consideradas son: para los especímenes estériles se consideraron las variables 1 a 9. Para los fértiles: 1 a 6 y 10 y 11. Para los ejemplares con características fértiles y estériles (EF): 1 a 11.

Las variables consideradas son:

1.- **Longitud de las pínulas.**
2.- Relación **longitud/anchura de las pínulas.**
3.- **Angulo de inserción de la pínula al raquis.**
4.- **Separación** (distancia) entre pínulas contiguas.
5.- **Anchura raquis.**
6.- Relación **anchura raquis/longitud.**
7.- **Número de venas laterales** que salen de la vena media.
8.- Relación **número de venas laterales/longitud.**
9.- Relación **venas al borde/venas laterales.** Donde venas al borde corresponde al número de venas que llegan al borde de la pínula.

10.- **Número de organos reproductores.**
11.- Relación **número de organos reproductores/longitud.**

Las variables se estandarizaron para evitar efectos por las diferentes escalas (ver Valencia, 1988).

AVANCE DEL ESTUDIO.

El estudio apenas se ha iniciado, hasta ahora se ha realizado un análisis de Cúmulos o Conglomerados (Cluster analysis) para agrupar observaciones, buscando identificar especies y tratar de establecer la edad de la formación geológica.

METODO.

Se utilizó el análisis de cúmulos o cluster para buscar indicaciones de las posibles especies presentes en las muestras de fósiles provenientes de la Formación Matzitzi. El propósito general de este análisis es encontrar el "agrupamiento natural" de objetos o individuos similares. El análisis cluster permite la búsqueda de patrones en los datos y con base en estos tratar de determinar las posibles especies presentes.

Con el análisis cluster, se pretende transformar el conjunto de individuos fósiles en una colección de subconjuntos o grupos exhaustivos y excluyentes, tales que los elementos dentro de cada subconjunto sean similares, en tanto que los elementos en subconjuntos diferentes sean lo mas disímiles posible.

Los datos para realizar análisis cluster se requieren de la siguiente manera: vectores de datos p-dimensionales[1]

$$X_1, X_2, \ldots, X_n$$

que se obtienen a partir de mediciones u observaciones de p características en n individuos u objetos. Las características de las variables pueden ser cuantitativas (discretas o continuas) o cualitativas (ordinales o nominales). Los datos pueden representarse mediante una matriz $X = [(x_{ij})]$, donde

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

La mayoría de las veces se tiene que el objetivo de cluster es agrupar los vectores de datos $X_i (i = 1, 2, \ldots, n)$, correspondientes a objetos o individuos, en g conjuntos (grupos, tipos, clases, etc.). Con este análisis las características de los grupos y aún

[1] Los elementos del vector X_j son $x_{j1}, x_{j2}, \ldots, x_{jp}$ para $j = 1, 2, \ldots, n$.

—en muchas ocasiones— el número de grupos, deben determinarse a partir de los datos mismos.

Para poder realizar un análisis cluster es necesario determinar cuando dos individuos u objetos están próximos o son similares. La selección de una medida de distancia o de disimilitud es necesaria. Además, debe definirse la medida de distancia entre grupos.

Cada diferente manera en que se defina la distancia entre grupos, determina un método o procedimiento diferente para realizar el análisis cluster. Ejemplos de maneras de definir la distancia entre grupos, así como de diversas medidas de distancia entre individuos pueden hallarse en Valencia (1988).

Diferentes elecciones de distancias entre elementos y diferentes maneras de definir la distancia entre grupos determinan por lo común diferentes agrupaciones.

Después de discutir las características de los diferentes tipos de análisis de cúmulos, se determinó que el apropiado para el caso que nos ocupa era un método jerárquico, aglomerativo, utilizando el método de centroides con la distancia Euclideana[2] como distancia entre observaciones. Esto es, la manera de definir la distancia entre grupos de observaciones sería la distancia entre los centroides (medias) de cada grupo.

Resulta importante observar que en este caso se espera que el análisis de cúmulos produzca una descripción de los datos y permita generar hipótesis sobre las posibles especies presentes en la formación de interés.

El procedimiento jerárquico elegido no permite (sin modificaciones) la elaboración de diagramas de árbol (también conocidos como dendrogramas). Esto se debe a la naturaleza misma del método, ya que al utilizar los centroides de cada grupo para calcular las distancias entre grupos, resulta que los centroides cambian al unir grupos y por tanto las distancias cambian después de unir los grupos, pudiendo resultar que ciertos grupos queden unidos en la etapa K a una distancia menor que la distancia en la que se unieron grupos en una etapa anterior.

En este caso, pensando en que una descripción gráfica del proceso de unión de los grupos era muy importante, se procedió de la manera siguiente: se construyó el dendrograma de la manera usual (ver Valencia, 1988) siempre que la distancia de unión en la etapa K fuera mayor o igual a la distancia de unión en la etapa $K-1$. Si la distancia en la etapa K resultaba menor que la distancia en la etapa $K-1$, entonces se reasignaba la distancia en la etapa K mediante una interpolación lineal entre las distancias correspondientes a las etapas $K-1$ y $K+1$.

Para describir las características medias de cada uno de los grupos se consideró apropiado utilizar gráficas estrella. Esto es, representaciones gráficas multidimensionales en las que cada estrella representa un grupo. La estrella consiste en una serie de rayos (tantos como variables) partiendo de un punto central, cada rayo representa una variable. La longitud del rayo representa el valor medio (muestral) de la variable en

---

[2] Con base en las correlaciones entre las variables de interés (ver tablas 4, 5, y 6) no resulta necesario considerar dentro del análisis a las covarianzas entre las variables. Descartándose de esta forma el uso de la distancia de Mahalanobis.

el grupo[3].

PROGRAMAS DE COMPUTO.

El análisis se realizó en el Laboratorio de Cómputo de la Facultad de Ciencias de la UNAM, utilizando máquinas PC compatibles.

Para realizar este análisis se consideraron las características del software disponible (SAS, versión 6.1, SPSS/PC+, STATGRAPHICS, versiones 1.0 y 2.1, SYSTAT, versión , SOLO-BMDP, versión 2.0, NCSS, versión 4.2, MAPSTAT y SST) resultando que sólo SYSTAT presentaba el procedimiento elegido. Sin embargo, debido al número de observaciones 610, no fue posible utilizar SYSTAT por el tiempo necesario para la ejecución. Por ésto, se procedió en el Laboratorio de Estadística de la Facultad de Ciencias de la UNAM a la elaboración de un programa que realice análisis cluster, con distancia euclideana[4] y con el método de centroides.

El programa esta compilado en QUICKBASIC (de Microsoft), versión 4.0. El programa resultante es rápido y hay una versión que puede hacer uso de coprocesador numérico si se cuenta con éste. Siendo el número de datos posibles de analizar una componente importante a considerar en este problema (uno de los archivos tiene 610 ejemplares[5] y 9 variables) se hizo el programa de tal forma que pudiera ser interrumpido (por ejemplo una falla en la energía eléctrica) y pudiera continuarse el análisis en otra ocasión, aprovechando fácilmente los cálculos ya hechos.

Los resultados pueden enviarse a la impresora o a un archivo en disco (en este caso, si se cuenta con algún disco duro, el programa funciona con más rapidez).

RESULTADOS.

A. EJEMPLARES ESTERILES.

Los ejemplares estériles estudiados fueron 610. En la figura 1, se tiene el diagrama de árbol (también conocido como dendrograma) correspondiente a las últimas 16 uniones de los ejemplares estériles. Con base en este diagrama se consideró que había siete grupos. Estos grupos están formados de la siguiente manera, contiene:

1) 543 ejemplares, el 89%.
2) 7 ejemplares, el 1.1%.
3) 16 ejemplares, el 2.6%.
4) 3 ejemplares, el 0.5%.
5) 9 ejemplares, el 1.5%.
6) 30 ejemplares, el 4.9%.
7) 2 ejemplares, el 0.3%.

---

[3] Ver Chambers, et al. 1983.

[4] Se cuenta ya con una versión que utiliza la distancia de Mahalanobis.

[5] Este número no es grande, sin embargo muchos de los programas comerciales considerados NO lo pueden manejar o lo manejan de tal modo que los tiempos de ejecución son muy grandes y resultan inaceptables prácticamente.

En la tabla 1, se presentan las medias correspondientes a estos grupos de ejemplares, en la figura 4 se encuentran las respectivas gráficas estrella y con los porcentajes anteriores puede mostrarse la importancia relativa del grupo 1 para la representación de los helechos fósiles.

Con base en la información anterior se tiene que la mayoría (el 89%) pertenece al grupo 1. El resto de las observaciones puede describirse en términos de sus medias mediante las gráficas estrella.

Analizando las observaciones pertenecientes al grupo 1, se ha establecido la hipótesis de que está formado posiblemente por 5 especies. Actualmente se esta trabajando mediante el análisis de la parte alta del árbol (esta parte no se ilustra en la figura 1).

B. EJEMPLARES FERTILES.

Los ejemplares fértiles estudiados fueron 335. En la figura 2, se tiene el diagrama de árbol correspondiente a las últimas 16 uniones de los ejemplares fértiles. Con base en este diagrama se consideró que había ocho grupos. Estos grupos están formados de la siguiente manera, contiene:

1) 306 ejemplares, el 91.3%.
2) 1 ejemplar, el 0.3%.
3) 13 ejemplares, el 3.9%.
4) 7 ejemplares, el 2.1%.
5) 2 ejemplares, el 0.6%.
6) 3 ejemplares, el 0.9%.
7) 2 ejemplares, el 0.6%.
8) 1 ejemplar, 0.3%.

En la tabla 2, se presentan las medias correspondientes a estos grupos de ejemplares, en la figura 5 se encuentran las respectivas gráficas estrella y con los porcentajes se muestra la importancia relativa del grupo 1 para la representación de los helechos fósiles.

La información anterior indica que la mayoría de las observaciones están en el grupo 1. Con base en las gráficas estrella pueden describirse el resto de las observaciones. Actualmente se está trabajando sobre el análisis morfológico de las observaciones del grupo 1. Posiblemente se presente una situación similar a la descrita respecto a los ejemplares estériles.

C. EJEMPLARES EF.

Son 38 los ejemplares con características tanto estériles como fértiles (EF) que fueron estudiados. En la figura 3, se tiene el diagrama de árbol correspondiente a las últimas 16 uniones de los ejemplares EF. Con base en este diagrama se consideró que había seis grupos. Estos grupos están formados de la siguiente manera, contiene:

1) 13 ejemplares, el 35.1%.
2) 1 ejemplar, el 2.7%.

3) 1 ejemplar, el 2.7%.
4) 2 ejemplares, el 5.4%.
5) 19 ejemplares, el 51.4%.
6) 1 ejemplar, el 2.7%.

En la tabla 3, se presentan las medias correspondientes a estos grupos de ejemplares, en la figura 6 se encuentran las respectivas gráficas estrella y los porcentajes muestran la importancia relativa de los grupos 1 y 5 para la representación de los helechos fósiles. Esta información indica que hay presentes dos grupos (el 1 y el 5). Hasta ahora se han identificado tentativamente dos especies de helechos. Con base en las gráficas estrella puede verse que la especie asociada al grupo 1 presenta valores medios más pequeños que los de la especie asociada al grupo 5. Los datos indican que las demás observaciones corresponden a individuos deformes pertenecientes a alguno de los dos grupos 1 ó 5.

COMENTARIOS.

Estando esta investigación en curso, no voy a hacer comentarios específicos y sólo quiero señalar algunos puntos de carácter general sobre las aplicaciones de la Estadística Multivariada.

Es por todos conocida la importancia que tienen los métodos y procedimientos de la Estadística Multivariada para la descripción y el análisis de conjuntos de observaciones en varias dimensiones. Es por ésto, que quiero señalar algunos puntos:

a) El condicionar el diseño o análisis de un experimento (o un conjunto de datos) al software disponible.

A menudo resulta que el estadístico y el interesado deciden, después de analizar y discutir el problema, el procedimiento estadístico a seguir y resulta no existe software disponible para llevarlo a cabo. El proceder del estadístico en estos casos debería de ser la elaboración del software adecuado para resolver el problema.

Sin embargo, en algunas ocasiones se llega al absurdo de cambiar el procedimiento elegido por un procedimiento (estadístico) que sí esté disponible computacionalmente, aunque hubiera sido descartado anteriormente en el proceso de discusión.

Otras veces, el estadístico (conocedor del software de que dispone o puede llegar a conseguir fácilmente) manipula la discusión para que resulte electo el procedimiento para el que cuenta con software.

Esto nos lleva al siguiente punto.

b) La carencia de software estadístico apropiado para realizar aplicaciones de la Estadística Multivariada.

El software disponible para realizar aplicaciones de Estadística Multivariada ha tenido un desarrollo relativamente reciente. Esto explica las carencias importantes aún en esta área y ha tenido como resultado que el software disponible no responda en muchos casos, siquiera a las necesidades básicas del investigador.

La consecuencia directa de estos comentarios es la urgente necesidad de contar con software estadístico (Multivariado en particular) apropiado y confiable.

BIBLIOGRAFIA.

- Carrillo Martínez, M. y colaboradores (1978): *Reconstrucción Paleoecológica y Paleogeográfica de la Formación Matzitzi, Tehuacán, Puebla.* Reporte de Biología de Campo de la Facultad de Ciencias de la UNAM, México.

- Chambers, J.M., Cleveland, W.S., Kliner, B. y Tukey, P.A. (1983): *Graphical Methods for data Analysis.* Boston: Duxbury Press.

- Silva Pineda, A. (1970): *Plantas Pensilvánicas de la Región de Tehuacán, Puebla.* Paleontología Mexicana, Instituto de Geología, Num. 29, UNAM, México.

- Valencia, G. (1988): *Análisis Cluster.* Publicaciones del Laboratorio de Estadística, Facultad de Ciencias, Universidad Nacional Autónoma de México, nº 1.

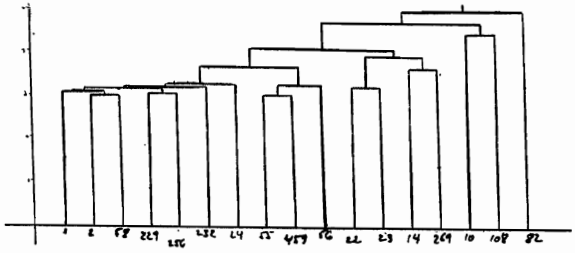


FIGURA 1. DIAGRAMA DE ARBOL CORRESPONDIENTE A EJEMPLARES ESTERILES.

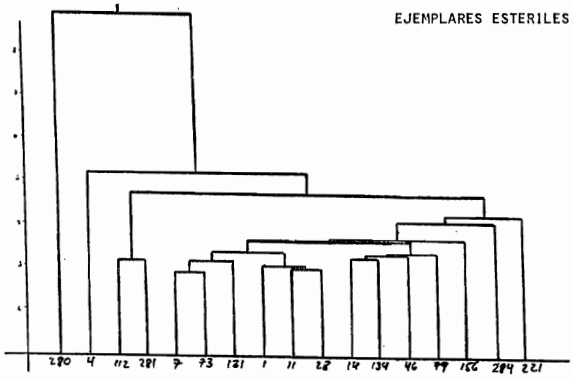


FIGURA 2. DIAGRAMA DE ARBOL CORRESPONDIENTE A EJEMPLARES FERTILES.

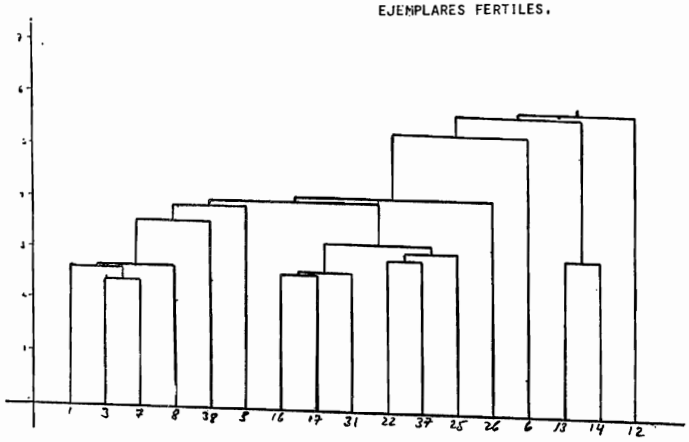


FIGURA 3. DIAGRAMA DE ARBOL CORRESPONDIENTE A EJEMPLARES EF

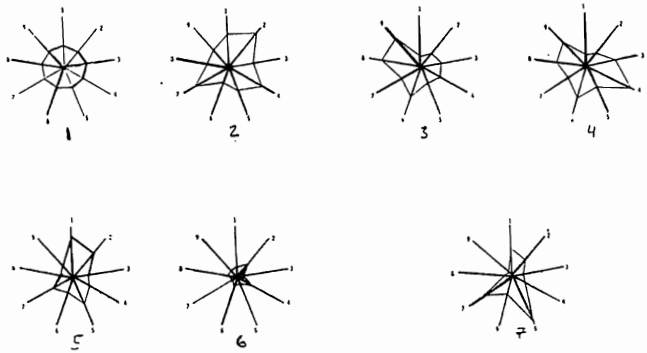


FIGURA 4. GRAFICAS ESTRELLA CORRESPONDIENTES A GRUPOS DE EJEMPLARES ESTERILES.

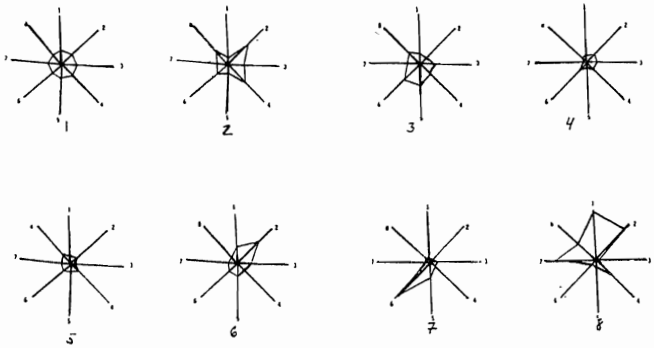


FIGURA 5. GRAFICAS ESTRELLA CORRESPONDIENTES A GRUPOS DE EJEMPLARES FERTILES.

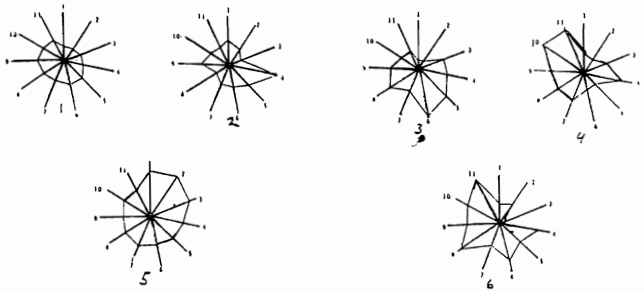


FIGURA 6. GRAFICAS ESTRELLA CORRESPONDIENTES A GRUPOS DE EJEMPLARES EF.

TABLA 1.-MEDIAS POR GRUPOS
EJEMPLARES ESTERILES.

GRUPOS.

1	2	3	4	5	6	7
0.040	1.509	-1.374	-1.226	2.597	-1.213	1.882
-0.043	3.117	-0.380	-0.678	1.423	-0.218	1.254
0.140	0.478	0.049	1.041	-0.833	-2.562	0.346
-0.021	2.078	0.208	3.477	-0.349	-0.487	0.569
0.069	0.382	-0.360	0.176	1.041	-1.749	3.563
0.040	-0.614	1.484	1.839	-0.607	-1.352	1.050
0.065	2.165	0.401	-0.245	0.295	-2.080	2.252
0.017	0.153	2.633	1.217	-1.142	-1.348	0.006
0.037	0.212	2.166	1.374	-0.814	-1.583	-0.532

TABLA 2.-MEDIAS POR GRUPOS
EJEMPLARES FERTILES.

GRUPOS.

1	2	3	4	5	6	7	8
0.076	-1.644	-0.602	-1.614	-0.946	0.828	-2.666	2.226
0.012	3.608	-0.826	-0.529	-0.747	3.755	-2.185	1.759
0.010	0.642	0.668	-0.468	-2.842	-0.082	-0.672	-1.757
0.032	2.751	0.086	-0.994	-0.838	-0.994	-0.994	-0.058
-0.028	-1.136	1.899	-1.902	-1.136	0.007	0.794	-1.136
-0.110	0.116	2.104	-0.834	-0.496	-0.483	7.947	-1.375
0.088	-0.699	0.027	-2.689	-0.924	-0.699	-2.721	1.548
0.047	1.021	0.690	-2.341	-0.271	-1.105	-1.941	-0.243

TABLA 3.-MEDIAS POR GRUPOS
EJEMPLARES EF.

GRUPOS.					
1	2	3	4	5	6
-0.330	1.099	-1.664	-0.436	0.362	-1.511
-0.545	0.172	-1.411	-0.570	0.538	-1.292
-0.016	-0.713	0.798	0.445	0.088	-2.527
0.070	4.327	1.119	1.921	-0.485	-0.485
0.736	1.364	2.323	-0.795	-0.532	-1.035
0.783	0.471	3.680	-0.579	-0.638	-0.405
0.535	0.414	0.414	1.200	-0.451	-1.159
0.692	-0.441	1.951	1.184	-0.645	0.025
0.918	1.330	0.905	1.277	-0.796	-0.796
0.430	0.392	-0.104	3.121	-0.576	-0.601
0.532	-0.204	0.836	2.936	-0.677	0.118

EL USO DEL CONCEPTO DE CONFUSION EN DISEÑOS CON RELACIONES DE ANIDAMIENTO

Vaquera Huerta, Humberto
Zarate de Lara,Guillermo P.
Burguete Hernandez, Francisco
Centro de Estadística y Cálculo, Colegio de Postgraduados , Chapingo, Méx.

RESUMEN

En este trabajo se presentan los diagramas de estructuras que se usan para representar diseños balanceados y completos y su relación con los modelos estadísticos. Se incluye además un algoritmo que permite relacionar algebráicamente, mediante el concepto de confusión, las relaciones existentes entre los términos de un diseño con relaciones de anidamiento y las interacciones entre factores de un diseño cruzado.

DIAGRAMAS DE ESTRUCTURAS

Los diagramas de estructura de Throckmorton (1961), Taylor y Hilton (1981), son una representación diagramática útil para tipificar los factores de clasificación de un conjunto de datos y por consiguiente para representar los

diseños experimentales.

Por medio de un diagrama de estructura se pueden conocer las relaciones de anidamiento y cruzamiento de los factores, lo cual es ventajoso cuando se tienen diseños de experimentos multifactoriales muy complejos. Lee (1966), Taylor y Hilton (1981) indican que se puede clarificar y simplificar el proceso de identificación del modelo lineal del diseño experimental que se tenga , usando diagramas de estructuras.

Throckmorton (1961) señala que siempre se tiene la posibilidad de simplificar y representar cualquier diseño experimental sin importar la complejidad de este, dibujando pequeñas figuras conocidas como diagramas de estructuras ó diagramas de Hasse, con la restricción de que los diseños caigan en la gran clase de los completos y balanceados. Tales figuras son el resultado de una correspondencia matemática entre estructuras de enrejado parcialmente ordenadas y estructuras de diseños experimentales.

Para representar un diseño experimental en forma diagrámatica, se le asigna una letra mayúscula a cada factor. El anidamiento entre factores se representa por la presencia de líneas conectadas hacia arriba entre las letras que representan a los factores; la ausencia de líneas indica cruzamientos. Cualquier factor conectado por líneas hacia arriba con otros factores se considera que esta anidado con

los factores de arriba. Cuando los factores no estan conectados por líneas hacia arriba se dice que estan cruzados.

La figura 1 representa un diseño experimental con 3 factores.

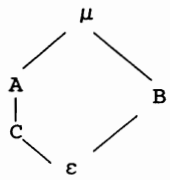


Figura 1

Donde, μ; es la media A, B, C son 3 factores y ε representa el error experimental.

En la figura anterior el factor A esta cruzado con B; C esta anidado en A; ε esta anidado en C, en B, y en A; El factor C esta cruzado con B y anidado en μ y en A, y anida a ε; El termino μ anida a todos los factores.

La representación de un diseño mediante diagramas de estructura lleva no solo implícitas las relaciones entre factores, sino que también indica el modelo lineal correspondiente a la estructura. Salvo la identificación de si ocurren o no los términos de interacciones entre factores cruzados, si se supone que ocurren todas las interacciones entre factores cruzados que son posibles, si queda determinado el modelo con la estructura.

En el trabajo de Vaquera (1989) se presenta el algoritmo de Lee (1975), el cual es un algoritmo general para encontrar el modelo asociado a un diagrama de estructuras .

El modelo asociado a la figura 1 resulta ser :

$$Y_{ijkl}=\mu + A_i + B_j + AB_{ij} + C_{(i)k} + BC_{(i)jk} + \varepsilon_{(ijk)l} \qquad (1)$$

En la expresión (1) se tiene el modelo correspondiente a la Figura 1.

En la Figura 2 se representa un diseño más complejo con 6 factores.

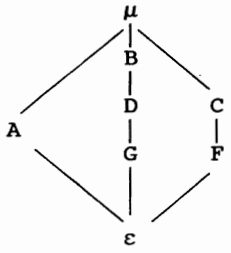


Figura (2)

Modelo correspondiente:

$$\begin{aligned} Y_{ijklmnp} = \mu &+ A_i + B_j + C_k + D_{l(j)} + AB_{ij} + AD_{il(j)} + AG_{in(jl)} + \\ &AC_{ik} + F_{n(lj)} + AF_{im(k)} + BC_{jk} + BF_{jm(k)} + DC_{lk(j)} + \\ &DF_{lm(jk)} + G_{m(k)} + GC_{kn(jl)} + GF_{mn(jkl)} + ABC_{ijk} + \\ &ADC_{ilk(j)} + ADF_{ilm(jk)} + ABF_{ijm(k)} + AGC_{ink(jl)} + \\ &AGF_{imn(jlk)} + \varepsilon_{p(ijklmn)} \end{aligned} \qquad (2)$$

UN ALGORITMO PARA RELACIONAR ALGEBRAICAMENTE LAS INTERACCIONES QUE EXISTEN EN UN DISEÑO TOTALMENTE CRUZADO CON LOS TERMINOS DE OTRO QUE PRESENTA RELACIONES DE ANIDAMIENTO.

Una vez que se conoce en forma explícita el modelo de un diseño que presenta factores anidados, la primer pregunta que se puede formular es: ¿que representa cada término en el modelo sobre todo los términos con subíndices de anidamiento ?.

En el Modelo 1 los términos $C_{k(j)}$ y $AC_{ik(j)}$ relacionan algebráicamente en terminos de grados de libertad y de sumas de cuadrados con las interacciones de un modelo cruzado de la siguiente forma

$$C_{k(j)} = C_k + BC_{kj}$$

$$AC_{ik(j)} = AC_{ik} + ABC_{ijk}$$

Las expresiones anteriores conducen a decir que en el Modelo 1 se tienen realacionados los efectos de C con la interacción BC y los de AC con la interacción triple ABC. A partir de estas relaciones algebráicas se puede mostrar que las sumas de cuadrados resultan ser:

$$SC(C_{k(j)}) = SC(C_k) + SC(BC_{kj})$$
$$SC(AC_{ik(j)}) = SC(AC_{ik}) + SC(ABC_{ijk})$$

Por lo que su identificación resulta el extremo util para llevara cabo el análisis estadístico de modelos con relaciones de anidamiento. Las relaciones algebraicas también se cumplen en los grados de libertad.

En el Modelo 2 se tienen las siguientes relaciones algebraicas:

$$D_{(j)i} = D_j + BD_{ij}$$

$$AD_{il(j)} = AD_{il} + ABD_{ijl}$$

$$G_{n(lj)} = G_n + DG_{ln} + BD_{jn} + BDG_{jln}$$

$$AG_{in(jl)} = AG_{in} + ABG_{ijn} + ADG_{iln} + ABDG_{ijln}$$

$$F_{m(k)} = F_m + CF_{mk}$$

$$AF_{im(k)} = AF_{im} + ACF_{imk}$$

$$BF_{jm(k)} = BF_{jm} + BCF_{jkm}$$

$$DC_{kl(j)} = DC_{kl} + BCD_{jkl}$$

$$DF_{lm(jk)} = DF_{lm} + BDF_{jlm} + CDF_{klm} + BCDF_{jklm}$$

$$CG_{kn(jl)} = CG_{kn} + BCG_{jkn} + CDG_{kln} + BCDG_{jkln}$$

$$FG_{mn(jkl)} = FG_{mn} + BFG_{jmn} + CFG_{kmn} + DFG_{lmn} + BCFG_{jkmn} + CDFG_{klmn} + BDFG_{jlmn} + BCDFG_{jklmn}$$

$$ACD_{ikl(j)} = ACD_{ikl} + ABCD_{ijkl}$$

$$ADF_{ilm(jk)} = ADF_{ilm} + ABDF_{ijlm} + ACDF_{iklm} + ABCDF_{ijklm}$$

$$ABF_{ijm(k)} = ABF_{ijm} + ABCF_{ijkm}$$

$$ACG_{ikn(jl)} = ACG_{ikn} + ABCG_{ijkn} + ACDG_{ilkn} + ABCDG_{ijkln}$$

$$AFG_{imn(jkl)} = AFG_{imn} + ABFG_{ijmn} + ACFG_{ikmn} + ADFG_{ilmn} +$$

$$ABCFG_{ijkmn} + ABDFG_{ijlmn} + ACDFG_{iklmn} + ABCDFG_{ijklmn}$$

De los ejemplos anteriores, se puede ver que en los diseños que presentan factores anidados, conforme aumenta el número de factores, se tiene un incremento en el número de efectos involucrados en las relaciones algebraicas en el modelo. En el modelo 2 se tienen 56 efectos en esas relaciones.

A continuación se presenta un algoritmo para obtener las relaciones presentes en un diseño anidado.

Una vez que se tengan los términos del modelo estadístico, es posible saber qué expresión se identifica con cada término anidado, adoptando los siguientes pasos:

Paso 1.

Denote la asociación entre los efectos de cada factor y los subíndices:

Ejemplo:Para los factores del Modelo 2 se establece la

siguiente correspondencia:

A$\longleftrightarrow$ i C $\longleftrightarrow$ k F $\longleftrightarrow$m

B $\longleftrightarrow$ j D $\longleftrightarrow$ l G $\longleftrightarrow$ n ε $\longleftrightarrow$ p

Paso 2.

Para cada término anidado se deben hacer todas las posibles combinaciones de los subíndices que estan dentro del paréntesis y ademas el número 1.

Ejemplo: Para el término $AFG_{imn(jkl)}$ Las combinaciones son:

jkl, jk, jl, kl, j, k, l, 1

Paso 3.

A cada combinación de subíndices obtenida en el paso anterior agregar los subíndices que estan fuera del paréntesis del término bajo consideración.

Ejemplo: En el término $AFG_{imn(jkl)}$ se tiene fuera del paréntesis imn por lo que obtenemos:

imn, ijmn, ikmn, ilmn, ijkmn, ijlmn, iklmn, ijklmn

Paso 4.

Finalmente escriba las interacciones o términos asociados a los subíndices escribiéndolas en forma aditiva e

igualarlas al término anidado bajo consideración, y finalmente se encuentra la relación de factores que representa.

Ejemplo:El término $AFG_{imn(jkl)}$ representa la relación de factores.

$$AFG_{imn(jkl)} = AFG_{imn} + ABFG_{ijmn} + ACFG_{ikmn} + ADFG_{ilmn} + ABCFG_{ijkmn} + ABDFG_{ijlmn} + ACDFG_{iklmn} + ABCDFG_{ijklmn}$$

Una vez que se tiene conocido el modelo estadístico asociado a un diagrama de estructuras, y además se conocen las relaciones de factores; lo siguiente es probar los efectos de los diferentes factores del modelo. Para el caso de los diseños anidados (completos y balanceados) es posible realizar el análisis de la varianza (ANVA), usando reglas para encontrar, sumas de cuadrados y esperanzas de cuadrados medios . En los textos de Winer (1971), Montgomery (1985), Méndez (1977), Searle (1971), Hicks (1982), Anderson (1974),se dan tales reglas .

BIBLIOGRAFIA

Anderson V. L., y McLean R. A. (1974) "Design of experiments, a realistic approach". Marcel Dekker. New York

Hicks C. R. (1982) "Fundamental concepts in the design of experiments". C B S College Publishing. New York.

Lee W. (1975) "Experimental design and analysis". W. H. Freeman and Company. San Francisco.

Lee W. (1966) "Experimental design symbolizationand model derivation". Psychometrika 31: 397-412.

Méndez Ramírez, Ignacio. (1977) "Modelos mixtos y aleatorios en el diseño y análisis de experimentos". serie azul: Monografías. IIMAS, UNAM.Vol.4 No.31

Montgomery D. C.(1985)."Design and analysis of experiments" John Wiley & Sons. New York.

Searle S. R. (1971) "Linear models". John Wiley & Sons. New York.

Taylor W. H., y Hilton H.G., (1981) "A structure diagram symbolization for analysis of variance". The American Statistician 35:85-93

Throckmorton T. N. (1961) "Structures of classification data". Unpublished Ph.D. Dissertation. Iowa State University, Dept. of Statistics. Ames, Iowa.

Vaquera H., H. (1989) "Diseños anidados, un enfoque metodologico" . Tesis de Maestría en Ciencias . Centro de Estadistica y Calculo, Colegio de

Postgraduados. Chapingo, Méx.

Winer B. J.(1971) "Statistical principles in experimental design". McGraw-Hill Book Company. New York.

SELECCION DE NIVELES DE OPERACION EN EL PROCESO DE EXTRUSION DE UN CEREAL.

Vargas Chanes, Delfino
Centro de Inevstigación en Matemáticas
Plaza la Valenciana s/n Apdo.Post. 402
C.P. 36000. Guanajuato, Gto.

Georgina Calderón Domínguez
Escuela Nacional de Ciencias Biológicas
Instituto Politécnico Nacional
Prolonganción Carpio y Plan de Ayala
México, D.F.

RESUMEN

Se presenta la descripción de la Metodología de Superficie de Respuesta, y se utiliza ésta para determinar los niveles óptimos de cuatro factores cuantitativos, para dos respuestas dependientes de forma univariada. Los factores controlados son: temperatura en la zona de compresión, temperatura en la zona de medición, velocidad de rotación del tornillo y humedad de alimentación; las variables de respuesta son contenido de lisina y dureza. Se utilizó un diseño central compuesto rotable 2^4, con repetición al centro. Se ajusta un modelo cuadrático y se observa buen ajuste para lisina y dureza. El punto óptimo seleccionado es 150 ^{o}C en la zona de compresión, 90 ^{o}C en la zona de medición, 70 RPM y 25 % de humedad.

INTRODUCCION

La metodología de la superficie de respuesta (MSR), en los últimos años ha recobrado interés. El auge se debe en gran parte a que los procesos de optimización de la industria han sido abordados utilizando herramientas estadísticas. Por ejemplo en la industria alimentaria, intervienen alrededor de 100 ó más variables en el proceso de elaboración de algún producto. En este caso una estrategia estadística adecuada, redunda en un ahorro de recursos y tiempo considerables, suficientes para justificar el uso de herramientas estadísticas.

La incursión de la MSR en la industria es relativamente reciente, pero los primeros trabajos publicados se reportan en la década de los cincuenta por Box y Wilson (1951), los trabajos posteriores de Box culminan con la publicación de un libro de Estadística para Experimentadores, Box, Hunter & Hunter (1978). El lector interesado en el tema puede consultar también Hill & Hunter (1966) en el que se resume la MSR, incluye una revisión bibliográfica del tema. El artículo de Steinberg & Hunter (1984) contiene una revisión bibliográfica más actualizada, con comentarios. Entre los artículos recientes más relevantes se encuentran Derringer & Suich (1980), Khuri & Conlon (1981) y Myers & Carter (1973), estos autores abordan el tema desde el enfoque multivariado, es decir proponen estrategias y metodologías para la respuesta óptima simultánea; aun cuando este problema es abierto, las aportaciones realizadas son importantes.

METODOLOGIA DE LA SUPERFICIE DE RESPUESTA

En esta sección se expone brevemente un resumen de la MSR, el lector interesado en aspectos teóricos más rebuscados es preferible remitirlo a Myers (1971) y Box, Hunter & Hunter (1978), también resulta recomendable el libro de Montgomery (1976).

Postulación del Modelo Matemático.

El modelo univariado general es de la forma

$$Y = \mathbf{X}\,\beta + \varepsilon \qquad (1)$$

donde

$Y^T = [Y_1,\ldots,Y_n]$ define la respuesta para n repeticiones.

$\beta^T = [\beta_1,\ldots,\beta_k]$ define los k parámetros.

$\varepsilon^T = [\varepsilon_1,\ldots,\varepsilon_n]$ define el vector de errores correspondiente.

y $\mathbf{X}_{nxk}$ es una matriz diseño con $r(\mathbf{X}) = k$

Afortunadamente en la mayoría de los casos un modelo de segundo

orden polinomial ajusta adecuadamente a los datos, cuando se está dentro de la región cercana al óptimo.

El modelo polinomial de segundo orden es:

$$Y = \beta_0 + \sum_i \beta_i x_i + \sum_i \beta_{ii} x_i^2 + \sum_{i>j} \sum_j \beta_{ij} x_{ij}. \qquad (2)$$

El fundamento de este modelo es la aproximación polinomial de Y, basada en la expansión en series de Taylor de Y alrededor del punto $x_1=x_2=....=x_k= 0$.

Los supuestos de este modelo son:

1. Existe una estructura de Y que es muy complicada o desconocida. Las variables de estudio son cuantitativas y contínuas.
2. La función Y se puede aproximar en la región de interés por un modelo polinomial de orden inferior.
3. Las variables independientes $x_1,...,x_k$ son controladas en el proceso de experimentación y medidas con un error casi nulo.

El Análisis de Datos.

A partir del modelo general propuesto, si la matriz $\mathbf{X}^T\mathbf{X}$ es no singular, los estimadores $\hat{\beta}$ de mínimos cuadrados se obtienen de la ecuación

$$\hat{\beta} = [\mathbf{X}^T\mathbf{X}]^{-1} \mathbf{X}^T\mathbf{Y} \qquad (3)$$

entonces podremos construir una superficie con los valores predichos

$$\hat{y} = \mathbf{x} \hat{\beta}, \qquad (4)$$

como función de las variables.

La siguiente etapa consiste en usar técnicas del análisis de varianza para evaluar el ajuste del modelo. En situaciones comunes el modelo ajustado es de primer o segundo orden, rara vez es de tercer orden. Para la determinación de las condiciones óptimas se debe tener en cuenta si 1) la región de estudio es una vecindad del ótimo y 2) la región de estudio está fuera de la vecindad dèl óptimo; nuestro caso corresponde a la situación (1).

Bajo el supuesto (1) se procede a analizar los datos con la

metodología de Box & Wilson (1951). Otro aspecto importante es utilizar el análisis canónico para conocer el comportamiento de las variables controladas, e interpretar adecuadamente el modelo. Cuando resulta particularmente difícil encontrar el óptimo, después de un análisis canónico, Myers (1971) recomienda un análisis de cordillera. El método de análisis de cordillera consiste en estimar una función de la respuesta que permite estudiar el comportamiento de las variables a lo largo de la superficie, alrededor de un radio.

Diseño Experimental.

Otro aspecto es la selección del diseño experimental, en la actualidad la mayoría de los experimentadores utilizan los factoriales fraccionados como diseños centrales, a estos les aumentan puntos exteriores que cumplan condiciones de rotabilidad además de las repeticiones centrales. Estos diseños se llaman centrales compuestos rotables y fueron propuestos por Box & Wilson (1951).

La selección del diseño consiste en determinar el modelo a estimar. Por ejemplo, si se estima un modelo de segundo orden, se utiliza un diseño de segundo orden.

Análisis Canónico.

La expresión (2) se puede reescribir en forma matricial como

$$y = \beta_o + \mathbf{x}'\beta + \mathbf{x}'\mathbf{B}\mathbf{x} \qquad (5)$$

donde:

$$\mathbf{x}=\begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_k \end{bmatrix}, \quad \beta=\begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}, \quad \mathbf{B}=\begin{bmatrix} \beta_{11} & \beta_{12}/2 & \dots & \beta_{1k}/2 \\ & \beta_{22} & \dots & \beta_{2k}/2 \\ & & \dots & \dots \\ \text{sym} & & \dots & \beta_{k-1,k}/2 \\ & & \dots & \beta_{k,k} \end{bmatrix}$$

Aquí se tiene que $\mathbf{x}'\beta$ és la porción de (5) que contiene a los términos

de primer orden y $\mathbf{x'Bx}$ contiene las contribuciones cuadráticas. El análisis canónico realiza una translación de la función de respuesta del origen (x_1=0, ... x_k=0) al punto estacionario $\mathbf{x}_o$; entonces la función de respuesta se puede reexpresar en términos de nuevas variables, W_1, W_2, ... , W_k. El origen se translada al centro del sistema de respuestas y se forman los ejes W_1, W_2, ... , W_k.

La forma de la función en términos de estas variables se llama forma canónica:

$$\hat{y} = \hat{y}_o + \lambda_1 W_1^2 + \lambda_2 W_2^2 + \ldots + \lambda_k W_k^2 \qquad (6)$$

donde $\hat{y}_o$ es la respuesta estimada en el punto estacionario, $\mathbf{x}_o = -\mathbf{B}^{-1}\mathbf{b}/2$ y λ_i son los valores carcterísticos de la matriz **B**. Usando la forma canónica se puede ver de manera más inmediata la contribución de los factores en la respuesta $\hat{y}$, los valores más grandes de λ_i denotan un mayor aportación del factor W_i.

UN ESTUDIO DE CASO

Calderón (1986) y Calderón *et al* (1987), estudiaron el proceso de cocimiento por extrusión de un cereal para desayuno. Al comienzo del experiemnto se tuvieron varios factores de estudio, al final se conservaron cuatro. En el presente trabajo los factores constantes son: temperatura en la zona de alimentación (100°C), velocidad de alimentación (71 g/min), diámetro del orificio del dado de salida (2 mm), relación de compresión del Tornillo (1:1); los factores controlados son: temperatura en la zona de compresión (A) en °C, temperatura en la zona de medición (B) en °C y humedad de alimentación (D) en %. En seguida se muestran los valores codificados de los niveles de cada factor:

Variables x_i	Valores Codificados ξ_i				
	-2	-1	0	1	2
A= T_2 °C	100	120	140	160	180
B= T_3 °C	80	85	90	95	100
C= RPM	50	62.5	75	87.5	100
D= Humedad	21.1	22.8	24.5	26.2	27.9

donde

$$\xi_i = \frac{x_i - \bar{x}_i}{x_{\Delta i}},$$

la variable x_i es el nivel de i-ésimo factor, $\bar{x}_i$ es la media del i-ésimo factor y $x_{\Delta i}$ es el incremento en el nivel de estudio del i-ésimo factor.

Se utilizó un diseño 2^4 central compuesto con tres repeticiones al centro y seis puntos axiales (α=2). Se optimizaron las respuestas Lisina (Y_1) en g aa/100 g de proteína y Dureza (Y_2) en Kg. En el cuadro 1 se tiene la matriz diseño y las variables de respuesta estudiadas. Se ajustó un modelo cuadrático para lisina y otro para dureza, de manera univariada.

Para la LISINA en el cuadro 2a se observa una R^2=0.92, se considera muy satisfactoria. En el cuadro 2b el análisis de varianza para el modelo de regresión ajustado, reporta significancia para los efectos lineal y cuadrático. En el cuadro 2d se muestran los parámetros estimados del modelo completo, los factores que más influyen son A (T_2 °C), B (T_3 °C) y C (RPM) para las componentes lineales y los términos cuadráticos A^2 , B^2, C^2 y D^2. En los cuadros 2e se puede apreciar que el punto estacionario es un mínimo y se encuentra dentro de la región de experimentación.

Para DUREZA las estadísticas básicas del cuadro 2a son aceptables, se tiene una R^2=0.99. La hipótesis de falta de ajuste se rechaza y los términos lineal, cuadrático y productos cruzados, son significativos (ver cuadros 2b y 2c). Se puede apreciar en el cuadro 2d que los términos lineales más relevantes son A (T_2 °C) y D (% de humedad), los términos cuadráticos y de productos cruzados significativos son A^2 , C^2

y AC. El punto estacionario es un punto silla y se encuentra fuera de la región de estudio para el factor D; debido a este hecho se realizó el análisis de cordillera descendiente en un radio de 1 unidad (ver cuadro 2e y 3).

Las ecuaciones empíricas son las siguientes, sólo se muestran las variables más importantes ($\hat{\alpha} < 0.3$). Ver cuadro 2d para los modelos completos:

$$\hat{LIS}_{emp} = 3.3 - 0.24A - 0.11B + 0.05C + 0.08A^2 + 0.17B^2 + 0.07BC + 0.27C^2 + 0.23D^2 \quad (5)$$

$$\hat{DUR}_{emp} = 2 - 0.07A + 0.02D - 0.16A^2 - 0.02AB - 0.18AC - 0.02BC + 0.34C^2. \quad (6)$$

Las ecuaciones canónicas son las siguientes:

$$\hat{LIS}_{can} = 3.065 + 0.277W_1^2 + 0.230W_2^2 + 0.155W_3^2 + 0.073W_4^2 \quad (7)$$

$$\hat{DUR}_{can} = 1.976 + 0.357W_1^2 + 0.008W_2^2 + 0.003W_3^2 + 0.173W_4^2 \quad (8)$$

Las gráficas tridimensionales se muestran en las figuras 1 y 2, para lisina y dureza, respectivamente. Las gráficas de contornos se muestran en las figuras 3 y 4 para lisina y dureza respectivamente. Las condiciones óptimas para generar la figura 1 están seleccionadas por el método canónico, el punto estacionario Ose puede leer en la interseccion de las rectas en la gráfica de contornos de la figura 3. Mientras que las condiciones óptimas para generar la figura 2 están determinadas por el análisis de cordillera descendente y el punto encontrado se puede leer en la grafica de contornos, figura 4, en la interseccion de las rectas.

CONCLUSIONES

La lisina se ve afectada por la temperatura en la zona de compresión y en la zona de medición, así como por la velociadad del extrusor. Mientras que la dureza, solo se ve afectada por la temperatura en la zona de compresión y el porciento de humedad. Este efecto se puede observar claramente en la magnitud de los coeficientes de las ecuaciones

canónicas 7 y 8, mientras más grandes sean los coeficientes, es mayor la contribución del factor. Para ambas variables el modelo ajustado se considera satisfactorio.

El caso típico de la MSR univariada es que los óptimos se encuentran en regiones distintas para cada variable. En el cuadro 4 se aprecia que las coordenadas para LISINA y DUREZA son distintas para los óptimos univariados. Se asume una solución única de compromiso, sobreponiendo las gráficas de contornos, figuras 3 y 4, y se propone como óptimo:

	U.Exp.	U.Cod.
A =	150°C	0.5
B =	90°C	0
C =	70 RPM	-0.4
D =	25 %	0

obteniendo una respuesta para Lisina de 3.2 g aa/100 g de proteína y para dureza 2.01 kg.

REFERENCIAS

Box, G.E.P.; Hunter W. & Hunter J. (1978). "Statistics for Experimenters". John Wiley & Sons, U.S.A.

Box, G.E.P. & Wilson, K.B. (1951). "On the Experimental Attainment of Optium Conditions". J. Roy. Statist. Soc., Ser. B, 13, 1-45.

Calderón G. (1986). "Estudio de las Condiciones del Proceso de Extrusión en la Elaboración de un Cereal para Desayuno". Tesis de Maestría en Ciencias (Alimentos) Escuela Nacional de Ciencias Biológicas - IPN. México.

Calderón G.; Zárate S. y Vargas D. (1987). " Extrusión de una Mezcla Maiz-Salvado: I. Selección de los Parámetros de Operación Mediante Superficie de Respuesta". Memorias Tecnología de Alimentos. Vol. 22, No. 5, México.

Derringer G. & Suich R. (1980). "Simultaneous Optimization of Several Response Variable". Journal of Quality Technology 12: 214-219.

Hill W.J. and Hunter W.G. (1966). A Review of Response Surface Methodology: A Literature Survey. Technometrics, Vol. 8, No. 4, 571-590.

Khuri A.J. and Conlon M. (1981). "Simultaneous Optimization Multiple Responses Represented by Polynomial Regression Functions".

Technometrics 23: 363-375.

Montgomery D.C. (1976). "Design and Analysis of Experiments". (2^a ed.) John Wiley & Sons. U.S.A.

Myers R.H. (1971). "Response Surface Methodology". Allyn and Bacon, ed. Boston, U.S.A.

Myers R.H. & Carter W.H. Jr. (1973). "Response Surface Techniques for Dual Response Systems". Technometrics 15: 301-317.

Steinberg D.M. and Hunter W.G. (1984). Experimental Design: Review and Comment. Technometrics Vol 26, No. 2, 71-97.

	VARIABLES ORIGINALES				VARIABLES CODIFICADAS				REPUESTAS	
OBS	T2	T3	N	H	A	B	C	D	LIS	DUR
1	120	85	62.5	22.8	-1	-1	-1	-1	4.4	2.07
2	120	85	62.5	26.2	-1	-1	-1	1	4.4	2.10
3	120	85	87.5	22.8	-1	-1	1	-1	4.4	2.42
4	120	85	87.5	26.2	-1	-1	1	1	4.3	2.51
5	120	95	62.5	22.8	-1	1	-1	-1	4.0	2.06
6	120	95	62.5	26.2	-1	1	-1	1	4.1	2.19
7	120	95	87.5	22.8	-1	1	1	-1	4.2	2.41
8	120	95	87.5	26.2	-1	1	1	1	4.3	2.52
9	160	85	62.5	22.8	1	-1	-1	-1	4.1	2.58
10	160	85	62.5	26.2	1	-1	-1	1	4.0	2.31
11	160	85	87.5	22.8	1	-1	1	-1	4.1	1.93
12	160	85	87.5	26.2	1	-1	1	1	4.0	2.10
13	160	95	62.5	22.8	1	1	-1	-1	3.6	2.29
14	160	95	62.5	26.2	1	1	-1	1	3.5	2.31
15	160	95	87.5	22.8	1	1	1	-1	3.8	1.91
16	160	95	87.5	26.2	1	1	1	1	3.9	1.89
17	140	90	75.0	21.1	0	0	0	-2	3.9	2.00
18	140	90	75.0	27.9	0	0	0	2	4.4	2.01
19	140	90	50.0	24.5	0	0	-2	0	4.2	3.35
20	140	90	100.0	24.5	0	0	2	0	4.4	3.35
21	140	80	75.0	24.5	0	-2	0	0	4.0	1.99
22	140	100	75.0	24.5	0	2	0	0	3.8	2.01
23	100	90	75.0	24.5	-2	0	0	0	4.2	1.48
24	180	90	75.0	24.5	2	0	0	0	2.9	1.22
25	140	90	75.0	24.5	0	0	0	0	3.1	2.0
26	140	90	75.0	24.5	0	0	0	0	3.4	2.01
27	140	90	75.0	24.5	0	0	0	0	3.4	1.99

Cuadro 1. Datos originales del diseño central compuesto rotable.

	Variable: Lisina	Dureza
Respuesta Media	3.955	2.174
Raiz cuadrada ECM	0.170	0.046
R-Cuadrada	0.918	0.994
Coef. of Variación	4.322	2.134

Cuadro 2a. Estadísticas básicas.

	LISINA			DUREZA	
Regresión	G.L.	S.C.	Pr>F	S.C.	Pr>F
Lineal	4	1.761	0.000	0.146	0.0001
Cuadrático	4	2.095	0.000	4.344	0.0000
Prod. cruzados	6	0.118	0.671	0.505	0.0000
Regresión Tot.	14	3.975	0.000	4.996	0.0000

Cuadro 2b. Análisis de la varianza.

Residual	g.l	Sumas de Cuadrados para LISINA	DUREZA
Falta de Ajuste	10	0.290	0.0256
Error puro	2	0.060	0.0002
Error total	12	0.350	0.0258

Cuadro 2c. Falta de ajuse.

		LISINA		DUREZA	
Parámetro	g.l.	Parámetro Estimado	Pr>F	Parámetro Estimado	Pr>F
Cte.	1	3.300	0.000	2.000	0.0000
A	1	-0.237	0.000	-0.074	0.0000
B	1	-0.112	0.007	-0.004	0.6679
C	1	0.054	0.146	0.003	0.7311
D	1	0.037	0.303	0.024	0.0254
A*A	1	0.078	0.056	-0.157	0.0000
B*A	1	-0.031	0.478	-0.018	0.1321
B*B	1	0.165	0.000	0.005	0.6278
C*A	1	0.018	0.668	-0.175	0.0000
C*B	1	0.068	0.137	-0.020	0.1105
C*C	1	0.265	0.000	0.342	0.0000
D*A	1	-0.018	0.668	-0.010	0.4057
D*B	1	0.031	0.478	-0.005	0.6742
D*C	1	0.006	0.886	0.008	0.4654
D*D	1	0.228	0.000	0.006	0.5457

Cuadro 2d. Nivel de significancia de los parámetros del modelo.

LIS: Optimo Individual				
$\hat{LIS}_{OPT} = 3.1$				
Variables de Estudio	A	B	C	D
Unidades Experimentales	173	92.8	72.1	24.4
Unidades Codificadas	1.65	0.55	-.23	-.05
DUR: Optimo Individual				
$\hat{DUR}_{OPT} = 1.75$				
Variables de Estudio	A	B	C	D
Unidades Experimentales	159.8	90.3	76.9	24.4
Unidades Codificadas	0.99	0.06	0.15	-.04

Cuadro 2e. Respuesta óptima para las variables del estudio de caso.

Radio	Respuesta Estimada	Factores A	B	C	D
0.0	2.000	0	0	0	0
0.1	1.990	0.097	0.005	0.007	-0.021
0.2	1.978	0.196	0.011	0.022	-0.030
0.3	1.962	0.294	0.017	0.038	-0.035
0.4	1.942	0.393	0.023	0.054	-0.038
0.5	1.919	0.492	0.029	0.071	-0.039
0.6	1.893	0.591	0.035	0.087	-0.039
0.7	1.863	0.689	0.041	0.104	-0.038
0.8	1.830	0.788	0.047	0.121	-0.038
0.9	1.794	0.886	0.053	0.137	-0.036
1.0	1 753	0.985	0.059	0.154	-0.035

Cuadro 3. Cordillera estimada para respuesta mínima de dureza.

PREDICCION DE AVENIDAS CON PERIODO DE RETORNO CONOCIDO

José A. Villaseñor Alva

Centro de Estadística y Cálculo, Colegio de Postgraduados, Chapingo, México, CP 56230, México

Philippe Bois

Ecole Nationale Supérieur d'Hydraulique et Mécanique de Grenoble. Domaine Universitaire, B.P. 95. 38402 Saint Martin d'Heres, Francia.

RESUMEN

En el diseño de estructuras hidraúlicas a lo largo de un río (puentes, presas, ...) es de primordial importancia el conocimiento del comportamiento probabilístico de las avenidas que se presentan. Una avenida es un flujo relativamente fuerte que sobrepasa un umbral crítico. Para que una construcción hidráulica sea eficiente durante un período largo de tiempo, es necesario conocer de la manera más precisa posible, los flujos de las avenidas para diversos períodos de retorno, principalmente los grandes.

En este trabajo se propone un método estadístico con base en el uso del número de avenidas k_i que sobrepasan un nivel fijo x_0, que constituyen las excedencias, y en el flujo máximo anual x_i del año i, para un número de años n. Este método ha sido implementado en el lenguaje Pascal (Turbo Pascal versión 4.0) para una computadora personal. El programa proporciona

al usuario el flujo de una avenida correspondiente a un período de retorno conocido.

1. EL MARCO TEORICO

Para este estudio el problema puede ser enmarcado teóricamente como un muestreo en dos etapas. Para una población dada, el tamaño k de la muestra es un valor tomado por una variable aleatoria K con función de densidad p_K sobre el conjunto de enteros positivos. Para K=k se observa la variable aleatoria X (la excedencia máxima). La función de distribución condicionada de X dado que K=k es F_o^k, en donde F_o es una función de distribución contínua definida sobre el intervalo $(x_o, +\infty)$, que constituye la distribución de las excedencias que a su vez se consideran independientes. En general se dipone de datos (x_1, k_1), (x_2, k_2), ..., (x_n, k_n), los cuales pueden ser considerados como una realización de las variables aleatorias bidimensionales (X_1, K_1), (X_2, K_2), ..., (X_n, K_n), independientes con la misma distribución.

Las excedencias máximas $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes con distribución común

$$F_M(x) = P(X \leq x) = \sum_{k=1}^{\infty} P(X \leq x | K = k) P(K = k)$$

$$= \sum_{k=1}^{\infty} F_O^k(x) p_K(k)$$

$$= \varphi_K(F_O(x))$$

en donde φ_K es la función generatríz de probabilidades de K.

El período de retorno t correspondiente a un valor x se define como

$$t = 1/(1-F_M(x)).$$

El valor de t es el número promedio de años que hay que esperar para que la excedencia máxima anual exceda al valor x.

Nótese que a períodos de retorno (t) grandes les corresponden excedencias (x) grandes. Debido a que estamos interesados en valores grandes de t, nuestras estimaciones estarán basadas en valores grandes de x.

Si la variable aleatoria K tiene una media finita $\mu(K)$, entonces

$$\lim_{x\to\infty} \frac{1-F_M(x)}{1-F_O(x)} = \lim_{x\to\infty} \frac{1-\varphi_K(F_O(x))}{1-F_O(x)} = \varphi'_K(1) = \mu(K) \quad .$$

Es decir, que para valores grandes de x, $1-F_M(x)$ puede ser aproximada por $\mu(K)(1-F_O(x))$.

2. ESTIMACION DE $F_M(x)$

De la sección anterior, se concluye que la estimación de $F_M(x)$ para valores grandes de x, requiere un estimador de

$\mu(K)$ y un estimador de $F_o(x)$ para x grande.

El estimador que se propone para $\mu(K)$ es la media muestral $\hat{\mu}_{K,n} = \sum_{i=1}^{n} K_i/n$.

Un estimador de $F_o(x)$, el cual se adapta convenientemente a nuestras necesidades, ha sido propuesto en Boyles y Samaniego (1986); este estimador es llamado "el estimador no paramétrico de máxima verosimilitud" (ENPMV).

El ENPMV es definido de la siguiente manera: para una realización $(x_1, k_1), (x_2, k_2), \ldots, (x_n, k_n)$, sean $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$ los valores $x_1, x_2, \ldots, x_n$ ordenados en forma creciente (es decir, $x_{1:n} \leq x_{2:n} \leq \ldots \leq x_{n:n}$), y sean $k_{1:n}, k_{2:n}, \ldots, k_{n:n}$ los valores k correspondientes originalmente asociados a los valores x .

Para $i=1,\ldots,n-1$ defínase

$$\hat{p}_i = \left(\sum_{j=1}^{i} k_{j:n} / \sum_{j=1}^{i+1} k_{j:n} \right)^{1/k_{i+1:n}} .$$

El ENPMV de $F_o(x)$ es dado por

$$\hat{F}_n(x) = 0, \quad x \leq x_{1:n}$$

$$= \prod_{j=1}^{n-1} \hat{p}_j, \quad x_{i:n} \leq x < x_{i+1:n}, \quad i=1,2,\ldots,n-1$$

$$= 1, \quad x \geq x_{n:n} .$$

Debido al hecho de que este estimador está basado en los valores máximos de muestras de F_o, es razonable esperar que las estimaciones de $\hat{F}_o$ tendrán una tendencia a ser más precisos para los valores grandes de x que para los pequeños. Este estimador tiene la propiedad de que su desviación estándar tiende a cero cuando x tiende a infinito (ver Boyles y Samaniego, 1986). Por lo tanto la parte superior de la gráfica de $\hat{F}_n$ proporciona una buena estimación de la misma parte de la gráfica de F_o.

3. ALISAMIENTO DE $\hat{F}_n(x)$

Para estimar $F_o(x)$ para valores grandes de x, es decir superiores a $x_{n:n}$, se propone realizar un alisamiento sobre la parte superior correspondiente al 30% superior de la gráfica de $\hat{F}_n(x)$.

Para este propósito, se propone ajustar cuatro modelos contínuos distíntos a la parte inferior correspondiente al 30% de la gráfica de $1-\hat{F}_n(x)$. Los modelos considerados son:

(1) $G_1(x) = a/(1+\exp\{cx+b\})$, decaimiento exponencial .

(2) $G_2(x) = a/(1+bx^c)$, decaimiento Pareto.

(3) $G_3(x) = a/(1+\exp\{bx^c\})$, decaimiento Weibull.

(4) $G_4(x) = a(1-\exp\{-\exp\{-(cx+b)\}\})$, decaimiento Gumbel.

Aún cuando los modelos (1) y (4) son equivalentes cuando x tiende a infinito, estos 4 modelos, en efecto incluyen casi todos los comportamientos de las colas a la derecha de las distribuciones que se presentan en las aplicaciones. El modelo que produce un ajuste con suma de desviaciones al cuadrado mínima es el que es seleccionado para realizar la estimación.

4. LA DISTRIBUCION DEL MAXIMO HISTORICO

De las secciones 1, 2 y 3 podemos concluir que una estimación de $1-F_M(x)$ para valores grandes de x, es proporcionada por $\hat{\mu}_{K,n} G_i(x)$ para alguna i en el conjunto {1, 2, 3, 4}, es decir, $\hat{F}_M(x) = 1-\hat{\mu}_{K,n} G_i(x)$, para alguna i y valores grandes de x.

Si recordamos que F_M es la distribución de la excedencia máxima anual, y que las excedencia máximas observadas X_1, X_2, ..., X_n son variables aleatorias independientes con la misma distribución, entonces el máximo histórico, definido por

$$M_n = \max(X_1, X_2, \ldots, X_n)$$

tiene la distribución $F_M^n(x) = P(M_n \leq x)$.

Por lo tanto un estimador de la distribución de M_n es

proporcionado por $\hat{F}_M^n(x)$.

Debido al hecho de que F_M está definida en el intervalo (x_o, ∞), de acuerdo con la teoría de valores extremos para muestras de tamaño grande n de $\hat{F}_M$, existen sucesiones de constantes $\{\alpha_n > 0\}$ y $\{\beta_n\}$ (que dependen de $1-\hat{F}_M(x)$) tales que

$$\lim_{n\to\infty} \hat{F}_M^n (\alpha_n x + \beta_n) = R(x)$$

para toda x punto de continuidad de la función R(x), en donde R(x) es una función de distribución de uno de los siguientes tipos:

(i) ley de Fréchet: para $\gamma > 0$, $\Phi_\gamma(x) = 0$, $x < 0$
$= \exp\{-x^{-\gamma}\}$, $x \geq 0$.

(ii) ley de Gumbel: $\Lambda(x) = \exp\{-\exp(-x)\}$, $-\infty < x < +\infty$.

Así, para valores grandes de n, sabemos que $P(M_n \leq x)$ es casi igual a $R((x - \beta_n)/\alpha_n)$.

Con base en los resultados de Villaseñor (1981), la ley de distribución R(x) y las sucesiones de constantes $\{\alpha_n > 0\}$ y $\{\beta_n\}$ son determinadas para cada uno de los modelos G_i, i = 1,2,3,4 (ver el apéndice), y las expresiones son las siguientes:

i=1: $R(x) = \Lambda(x)$, $\alpha_n = 1/c$, $\beta_n = (-b+\ln(n\hat{\mu}_{K,n}))/c$.

i=2: $R(x) = \Phi_c(x)$, $\alpha_n = (n\hat{\mu}_{K,n}/b)^{1/c}$, $\beta_n = 0$.

$$i=3:\ R(x) = \Lambda(x),\ \alpha_n = c^{-1}b^{-1/c}(\ln(n\hat{a}_{K,n}))^{-1+1/c}$$

$$\beta_n = (b^{-1}\ln(n\hat{a}_{K,n}))^{1/c} .$$

$$i=4:\ R(x) = \Lambda(x),\ \alpha_n = 1/c,\ \beta_n = (-b+\ln(n\hat{a}_{K,n}))/c .$$

Para un período de retorno t dado (en este caso t está dada en unidades de n años), vemos que el caudal histórico máximo (es decir la x tal que $t = 1/P(M_n > x)$) es estimado por:

$$i=2:\ x(t) = \alpha_n(-\ln(1-1/t))^{-1/c}$$

$$i=1,3,4:\ x(t) = \beta_n - \alpha_n\ln(\ln(1-1/t)) ,$$

en donde las constantes α_n y β_n son las correspondientes a cada unas de las i.

BIBLIOGRAFIA

Boyles, R.A. y Samaniego, F.J. (1986). "Estimating a distribution function based on nomination sampling". Journal of the American Statistical Association, 81, 1039-1045.

Villaseñor, J.A. (1981). "Norming constants for maxima attracted to exp(-exp(-x))". Proceedings of the 43rd Session of the International Statistical Institute, Buenos Aires, Argentina, 147-149.

APENDICE

El cálculo de las constantes de normalización $\{\alpha_n > 0\}$ y $\{\beta_n\}$ está fuertemente ligado al comportamiento de la cola derecha de la distribución $\hat{F}_M(x)$, es decir a la función $1-\hat{F}_M(x)$ cuando x tiende a infinito (ver Villaseñor (1981)).

Sea $\hat{\mu} = \hat{\mu}_{K,n}$. A continuación se considera cada modelo separadamente.

o

Modelo (1): Nótese que

$$\lim_{x\to\infty} e^x(1-\hat{F}_M(x/c)) = \hat{\mu}ae^{-b} .$$

Entonces por la Proposición 1 en Villaseñor (1981), se tiene que

$$\lim_{n\to\infty} \hat{F}_M^n(x/c+\ln(n\hat{\mu}ae^{-b})/c) = \Lambda(x) .$$

Es decir, $\alpha_n = 1/c$ y $\beta_n = (-b+\ln(n\hat{\mu}a))/c$.

Modelo (2): Nótese que

$$\lim_{x\to\infty} x^c(1-\hat{F}_M(x)) = \hat{\mu}a/b .$$

Entonces
$$\lim_{n\to\infty} \hat{F}_M^n((n\hat{\mu}a/b)^{1/c}x) = \lim_{n\to\infty}\left(1- \frac{x_n^c(1-\hat{F}_M(x_n))x^{-c}}{n\hat{\mu}a/b}\right)^n$$

$$= \Phi_c(x) ,$$

en donde $x_n = (n\hat{\mu}a/b)^{1/c}x$. Es decir

$\alpha_n = (n\hat{\mu}a/b)^{1/c}$, $\beta_n = 0$.

Modelo (3): Nótese que

$$\lim_{x\to\infty}(1-\hat{F}_M(x/b^{1/c}))\exp\{x^c\} = \hat{\mu}a \ .$$

Entonces por la Proposición 1 de Villaseñor (1981), se tiene que

$$\lim_{n\to\infty}F_M^n((\alpha'_n x + \beta'_n)/b^{1/c}) = \Lambda(x),$$

en donde $\alpha'_n = c^{-1}(\ln(n\hat{\mu}a))^{-1+1/c}$ y $\beta'_n = (\ln(n\hat{\mu}a))^{1/c}$.

Es decir, $\alpha_n = \alpha'_n/b^{1/c}$ y $\beta_n = \beta'_n/b^{1/c}$.

Modelo (4): Nótese que por la regla de L·Hospital,

$$\lim_{x\to\infty}(1-\hat{F}_M(x/c))e^x = \hat{\mu}ae^{-b} \ .$$

Por lo tanto las constantes son las mismas que para el modelo (1).

Software para PC'S en Geoestadística.

Fernando Avila Murillo.

Departamento de Matemáticas.

Universidad de Sonora.

RESUMEN

Una parte esencial de la práctica geoestadística es el uso de programas para computadora, en los cuales se han implementado los diversos algoritmos y métodos sugeridos por la teoría. Debido a la ubicuidad de las computadoras personales, la tendencia en los últimos años ha sido la de diseñar el software geoestadístico en función de las características de estas máquinas. En esta nota se comentan en forma general algunos criterios de evaluación de software y se reseñan dos colecciones de programas ‘paquetes’ populares en el medio geoestadístico.

INTRODUCCION.

Puesto que la Geoestadística se ha desarrollado en función de sus aplicaciones, el uso de programas de cómputo, y de computadoras con equipo periférico, ha sido parte integral de su práctica. El resultado de este énfasis en computación es que el software geoestadístico se presenta en gran variedad de tamaños, sabores y colores.

Se han diseñado programas de "uso general" y programas para resolver problemas específicos de la minería, de la prospección petrolera, de la geoquímica, etc. Algunos programas requieren grandes computadoras y otros pueden ser ejecutados en computadoras de bolsillo. Hay programas que se han desarrollado con fines comerciales y se venden o rentan por miles de dólares, mientras que otros fueron escritos por estudiantes o profesores universitarios con fines académicos y de investigación y son del dominio público.

Durante años, se ha considerado a BLUEPACK [1] como el estándar de software geoestadístico. Se trata de una colección de programas escritos en FORTRAN para una computadora grande (VAX por ejemplo) que corre en forma "batch" con un mínimo de interacción por parte del usuario. Los autores, franceses todos, reconocidos geoestadísticos que han contribuido significativamente al desarrollo de la disciplina, han incorporado gran parte de la teoría geoestadística lineal a su programa, convirtiéndolo en un potente auxiliar de proyectos de investigación teóricos o aplicados. BLUEPACK no es gratis.

En el ámbito académico, la fuente original de muchos programas que se usan actualmente es el conjunto de programas escritos por Knudsen en FORTRAN [2] y que en forma modificada siguen apareciendo en paquetes comerciales y del dominio público. Otra fuente de programas, y de referencia, es el libro de Journel y Huijbregts [3]. El profesor Journel tiene la saludable costumbre de exigir a sus alumnos de Stanford que incluyan los programas usados en sus tésis; también es uno de los iniciadores y promotores de GEO-EAS, que se reseñará más adelante.

Los programas hasta aquí mencionados tienen en común lo siguiente: están escritos en FORTRAN, diseñados para correr en un Centro de Cálculo en forma "batch" con un mínimo de participación del usuario, quien debe esperar los resultados para decidir el siguiente paso de su investigación o para realizar nuevas corridas; el apoyo gráfico de los programas es rudimentario y visible solamente en papel.

Con la disponibilidad actual de las llamadas computadoras personales, el software geoestadístico ha mejorado, principalmente en las áreas de interfase con el usuario y de graficación. A continuación describiré algunas características del "paquete" ideal, aún inexistente.

El software ideal debe ser confiable, eficiente, suficiente, amigable, accesible y gratis:

-Confiable: los algoritmos deben ser estables numericamente y dar resultados correctos cuando los datos y el problema están planteados correctamente.

-Eficiente: no debe desperdiciar tiempo y memoria; debe presentar los resultados de una manera efectiva; no debe pedir al usuario información inecesaria o que pueda ser inferida automaticamente.

-Suficiente: debe producir todos los resultados necesarios para el usuario; minimamente debe calcular variogramas experimentales, modelar variogramas teóricos, hacer kriging ordinario puntual y por bloques y construir mapas.

-Amigable: debe ser fácil de usar, aún por usuarios con un mínimo de conocimientos de programación; debe dar ayuda cuando se presentan errores en la ejecución; debe ser interactivo en todas las etapas, presentando resultados y pidiendo decisiones intermedias.

-Accesible: debe ser fácil de obtener.

-Gratis: !

Como ejemplos de "paquetes" que satisfacen algunos de los criterios mencionados, mencionaré a STATPAC y GEO-EAS los cuales son conocidos en el ámbito académico, del dominio público y facilmente conseguibles.

STATPAC.

STATPAC es un conjunto de programas usado y distribuido por la United States Geological Survey (USGS). La versión que comentamos está fechada el 8 de Marzo de 1988 y consta de 10 discos doble-densidad.

Cuatro discos contienen programas fuente escritos en FORTRAN 77, BASIC, ASSEMBLER 8086 y Turbo PASCAL. Casi todos los programas fueron escritos por W. D. Grundy y A. T. Miesch.

Un disco contiene una biblioteca de módulos objeto de los subprogramas escritos en FORTRAN y ASSEMBLER, y resulta muy útil cuando es necesario recompilar algún programa principal.

Los cinco discos restantes contiene programas ejecutables (*.EXE), entre los que se encuentran los de la serie SS2D****, los cuales sirven para realizar análisis geoestadístico en dos dimensiones. Todos los programas ejecutables requieren el coprocesador numérico 8087 de INTEL (o el 80287 para una IBM PC AT) y al menos 320K bytes de memoria.

Los discos están formateados para el sistema operativo de Microsoft™ versión 2.* o más reciente. Los programas escritos en FORTRAN pueden ser compilados con el compilador de Microsoft™ versión 3.30 o posterior. Los programas en BASIC están escritos en IBM BASIC y podrían ser compilados para obtener ejecutables más rapidos que los programas interpretados.

La información anterior, y mucha más, está disponible en un archivo README incluido en los discos. El archivo README es un archivo de texto que se puede imprimir (35 páginas) y que sirve como manual de STATPAC. El archivo STPINFO.ONE es un archivo de texto que sirve como introducción (tutor) general al uso de algunos de los programas; en cambio, STPINFO.TWO se refiere especificamente a los programas de uso geoestadístico.

En términos generales, los archivos de texto y los comentarios incluidos en los programa fuente son suficientes para orientar a los usuarios que nunca han usado STATPAC. Los ejemplos son claros y la documentación está bien redactada. A la fecha no he tenido necesidad de recurrir a los autores para aclaraciones o explicaciones adicionales. Los programas han corrido satisfactoriamente y no han provocado problemas que requieran el uso del botón de RESET. Internamente, STATPAC es un conjunto de programas consistente, eficiente y bien estructurado.

A continuación mencionaré algunas virtudes y "defectos" específicos:

-El análisis geoestadístico está restringido a problemas en dos dimensiones. Se pueden modelar los tipos de variograma mas comunes, hacer kriging ordinario o universal, puntual o por bloques y construir mapas de contorno. Los modelos se pueden validar usando el método de validación cruzada o ajustar con una técnica iterativa que usa un criterio de mínimos cuadrados.

-STATPAC incluye programas que implementan las tareas y métodos más comunes de la Estadística: ANOVA, ajuste de funciones de dos variables, regresión lineal múltiple, análisis discriminante, análisis de componentes principales y prueba chi-cuadrada de normalidad, entre otros. Los algoritmos usados son los estándares más confiables para cada método.

-Los archivos de datos STATPAC pueden ser manipulados, transformados, transferidos, revisados, etc., a través de una serie de programas de apoyo.

-Los resultados quedan guardados, en archivos de texto que pueden imprimirse directamente, o en archivos con formato especial que requieren de un programa específico para su impresión.

-Las gráficas son, en general, rudimentarias; la mayoría se escriben en archivos de texto que deben imprimirse para ser analizadas.

-Como el paquete está orientado al análisis de datos geoquímicos y petrológicos, los archivos de datos tienen una estructura peculiar. Un registro típico consiste de un identificador, dos enteros que pueden ser coordenadas geográficas, valores de las variables numéricas y códigos calificadores de los valores numéricos.

GEO-EAS.

GEO-EAS es una colección de programas distribuida por la Agencia de Protección al Ambiente de los Estados Unidos (US EPA). Los programa fuente están escritos en FORTRAN y se compilan con el compilador de Microsoft™ V4.01 o posterior. Los programas compilados ocupan alrededor de 3 megabytes de memoria y pueden ser obtenidos en discos de 5 1/4", doble o alta densidad, o en discos de 3 1/2". Los programas corren bajo DOS, necesitan 640 Kb de RAM y no requieren coprocesador aritmético.

Los programas de GEO-EAS pueden ser ejecutados en forma individual o llamados desde un menú, cuando se guardan todos los programas en disco duro. Una de las principales virtudes de GEO-EAS es que, con excepción de algunos archivos para graficación, todos los archivos de entrada de datos y todos los archivos de resultados son archivos de texto (ASCII) con la misma estructura, y por lo tanto, pueden ser usados por todos los programas de la colección.

Todos los programas se controlan interactivamente por medio de pantallas de menú que ofrecen diversas opciones al usuario. Todas las pantallas tienen el mismo formato y su manejo es muy sencillo. Se puede obtener una explicación muy detallada del uso de cada uno de los programas en un manual de más de 200 páginas escrito por Evan Englund y Allen Sparks, principales promotores y diseñadores de los programas.

GEO-EAS surgió como resultado de varios proyectos que la US EPA condujo a través de convenios cooperativos entre su Laboratorio de Sistemas de Monitoreo Ambiental - Las Vegas y las universidades de Stanford, Wyoming y Arizona. La programación fue realizada en su mayor parte por Computer Sciences Corporation de Las Vegas, a partir de programas fuentes escritos en las citadas universidades.

Señalamos ahora algunos puntos específicos de GEO-EAS, versión 1.1:

-La interfase con el usuario a través de menús por pantallas es efectiva, eficiente y fácil de usar. Los programas calculan valores "default" en cada etapa, pero el usuario siempre tiene la opción de modificar los valores propuestos, haciendo que GEO-EAS sea verdaderamente interactivo.

-La estructura de los archivos de datos y de los archivos de resultados es muy sencilla: con excepción de algunos archivos METACODE para gráficas, todos los programas de GEO-EAS usan y generan archivos ASCII, que se pueden leer en pantalla, imprimir directamente en impresora o modificar con cualquier editor de textos.

-La selección de técnicas es un poco idiosincrática: se pueden escoger alternativas al variograma, pero no se puede modelar una "deriva"; tampoco hay un gran surtido de modelos de variograma, así que puede resultar necesario usar varias estructuras anidadas.

-GEO-EAS no tiene muchos programas que realicen tareas estadísticas tradicionales. Por ejemplo, la regresión lineal aparece como es una opción en los programas que hacen diagramas de dispersión, y sus resultados no son muy confiables.

-Se enfatizan las gráficas, las cuales son excelentes; en un monitor de colores con una computadora equipada con tarjeta EGA se pueden examinar gráficas a 4 colores, verdaderamente útiles para el análisis estadístico.

-Casi todos los programas han tenido errores de programación. La reciente versión 1.2 corrige la mayoría de los problemas detectados, pero es posible que haya errores ocultos aún.

CONCLUSIONES.

No existe el software que satisfaga todas las necesidades de todos los usuarios. En Geoestadística se ha popularizado el uso de programas para computadoras personales y los ejemplos reseñados son muestras de la gran calidad que es posible obtener cuando se conjuntan los esfuerzos de profesionales de la Geoestadística y la computación.

REFERENCIAS.

1.- P. Delfiner, J.P. Delhomme, J.P. Chiles, D. Renard et F. Drigoin (1980). BLUEPACK-3D Notice L'Utilisation. Centre de Géostatistique et de Morphologie Mathématique.

2.- Knudsen, H.P. and Kim, Y.C. (1978). A short course on geostatistical ore reserve estimation. Departament of Mining and Geological Engineering, University of Arizona.

3.- Journel, A.G. and Huijbregts, Ch. J. (1978). Mining Geostatistics. Academic Press.

HIBRIDIZACION DE LA GEOESTADISTICA

Fernando Paz P.
CISIUS y Depto. de Minas, Escuela de Ingeniería, Universidad de Sonora, Apdo. Postal B-94, Hermosillo, Sonora

RESUMEN

La hibridización de la Geoestadística es presentada como una consecuencia de utilizar información objetiva y subjetiva. De una descripción de situaciones de decisión y de su liga en una jerarquía de niveles de información, se introducen variables y procesos híbridos. El kriging híbrido sirve como un vehículo normal para tratar con procesos de estimación, tanto para valores discretos como contínuos.

INTRODUCCION

El problema de estimación y simulación de campos aleatorios ha sido abordado por la Geoestadística en base a la función variograma y al método Kriging. La Geoestadística, en la práctica, se ocupa del estudio de las mediciones hechas sobre un fenómeno natural, el cual es interpretado como un proceso estocástico. En base a ciertas hipótesis, estacionariedad por ejemplo, la Geoestadística interpreta los valores observados de un proceso como realizaciones de variables aleatorias regionalizadas. En un sentido estricto, pueden verse las hipótesis constitutivas de la teoría geoestadística como un modelo matemático para caracterizar un posible proceso determinístico, el cual no puede ser definido completamente por restricciones de información. Por consideraciones prácticas en el muestreo de un proceso, la hipótesis geoestadística no puede ser validada con completa certidumbre.

En otro nivel de información, existe un gran número de situaciones prácticas donde los valores de un proceso son una mezcla de información "dura" y "blanda". En este trabajo, el concepto duro se interpreta en términos probabilísticos (con la constante como un caso especial). La información "blanda" es tratada bajo la perspectiva de la Teoría de los Conjuntos Borrosos desarrollada por Zadeh (1965). De esta forma, es introducido el aparato intuitivo y matemático de las variables borrosas y su unión con las variables aleatorias. A partir de diferentes situaciones, la hibridización de la Geoestadística es presentada como una mezcla de información probabilística y borrosa. Por sus divergencias de niveles de información entre las dos teorías, se ha utilizado el término "híbrido" para connotar situaciones o procesos combinados. En lo siguiente, nos limitaremos a la exposición de ideas que puedan ser usadas con fines prácticos, con énfasis en Geotecnia,

y que son las más interesantes para el desarrollo de una Geoestadística híbrida.

TIPOLOGIA DE NIVELES DE INFORMACION

El objetivo de la Geoestadística puede ser puesto como un problema de toma de decisiones, donde se debe optimizar el uso de la información disponible para definir una función de interpolación en un espacio dado. En este contexto, es importante tipificar los diferentes niveles de información a ser encontrados: certidumbre, riesgo, incertidumbrre, borrosidad e ignorancia.

a) Certidumbre. Se refiere a la situación donde se conoce con completa certeza el valor que tomará una variable o proceso, Fig. 1. En términos corrientes, a ésta variable se le denomina constante y es un caso especial de una variable aleatoria (distribución Dirac, $f(z_o) = 1$) y de una variable borrosa ($\mu(z_o) = 1.0$).

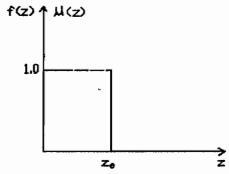


Figura 1. Función densidad o función de membresía de una Constante.

b) Riesgo. Es la situación donde los valores de una variable o proceso no se conocen con certeza, sino sus probabilidades. Estas últimas pueden ser medidas (enfoque frecuentista) o inferidas (enfoque bayesiano). La Fig. 2 muestra ésta variable. Una variable aleatoria está basada en mediciones, ya sea en términos de frecuencias de ocurrencia o inferidas a través de hipótesis como

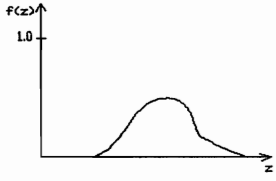


Figura 2. Función densidad de una Variable Aleatoria

ocurre en Geoestadística. La constante es un caso particular de una variable aleatoria que implica que su probabilidad de ocurrencia es 1.0.

f(z) se denomina funcion densidad de probabilidad de la variable aleatoria y su integral es F(z), que es la función de distribución acumulada de la variable aleatoria y tiene como dominio:

$$F(z) \in [0,1] \tag{1}$$

y es monóticamente creciente.

c) Incertidumbre. Denota la situación donde se desconocen las probabilidades asociadas a los valores de una variable, pero el rango de ocurrencia $[z_1, z_2]$ es conocido. La Fig. 3 muestra el concepto de intervalo asociado con la situación de incertidumbre.

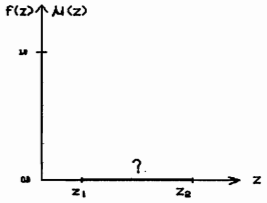


Figura 3. Función densidad o función de membresía de un Intervalo.

El intervalo $[z_1, z_2]$ tiene un nivel de probabilidad pequeño, casi cero ($f[z_1, z_2] \approx 0$), pero un grado de membresía μ de 1.0 (ver Fig. 7b):

$$\mu[z_1, z_2] = 1.0$$
$$f[z_1, z_2] \approx 0.0 \tag{2}$$

La variable aleatoria es una generalización del concepto de intervalo con diferentes niveles de probabilidad $f[z_i, z_{i+1}] \geq 0$, todos con $\mu[z_i, z_{i+1}] = 1.0$. Así, un intervalo puede ser definido en base a diferentes niveles de probabilidad asociados, siempre y cuando exista ese tipo de información o pueda ser inferida en base a los datos muestrales.

d) Borrosidad. El término borrosidad es interpretado en función de la teoría de los conjuntos borrosos (Zadeh, 1965). La Fig. 4 muestra una variable borrosa. Una variable borrosa es un conjunto borroso que es convexo y normal (Nahmias, 1978; Kaufmann y Gupta, 1985). La función $\mu(z)$ se denomina función de membresía y mide

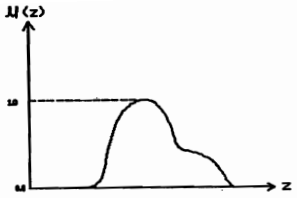


Figura 4. Función de membresía de una Variable Borrosa.

el grado de pertenencia de los valores z a la variable borrosa en estudio. $\mu(z)$ generaliza el concepto de pertenencia o no pertenencia a valores intermedios de verdad. Por ejemplo, en una variable aleatoria todos sus valores tienen $\mu = 1.0$, puesto que se conoce que éstos valores son ciertos (verdaderos) a presentarse, con diferentes probabilidades. La función de membresía está definida en:

$$\mu(z) \in [0,1] \qquad (3)$$

donde $\mu = 0$ implica la no pertenencia total al conjunto y $\mu = 1.0$, la pertenencia total. La constante es un caso especial de la variable borrosa ($\mu = 1.0$). La variable borrosa puede interpretarse como una generalización del concepto de intervalo a diferentes niveles de membresía μ, pero con probabilidades desconocidas. La variable borrosa es determinada en forma subjetiva y no proviene de mediciones. $\mu(z)$ es una medida de menor nivel que $f(z)$.

Las operaciones con variables borrosas están definidas por la convolución Max-Min (Kaufmann y Gupta, 1985):

$\forall$ x, y $\in$ R :

$$\mu_{A(*)B}(z) = \underset{z=x*y}{\text{MAX}} \{\text{MIN} [\mu_A(x), \mu_B(y)]\} \qquad (4)$$

donde * es cualquier operador matemático del par x,y y A y B son variables borrosas asociadas a x e y, respectivamente.

Una forma más compatible con la forma de procesar información de los seres humanos, es la variable lingüística introducida por Zadeh (1975). Una variable lingüística es interpretada como una restricción borrosa que asocia una distribución de posibilidad (Zadeh, 1978) con los valores que puede tomar esta variable en un universo de discurso. Una variable lingüística se define en palabras, relacionadas a valores numéricos de referencia por medio de la distribución de posibilidad. La posibilidad es diferente de la probabilidad y ambos conceptos están relacionados a través del principio de consistencia probabilidad-posibilidad (Zadeh, 1978)

que establece que:

> "El grado de posibilidad de un evento o variable es mayor o igual a su grado de probabilidad"

En términos matemáticos, el principio de consistencia puede ser puesto como:

$$\mu(z) \geq f(z) \qquad (5)$$

Dadas las bases de ambas teorías, la borrosidad y la probabilidad no pueden mezclarse directamente y no existe un plano de comparación entre ellas.

En programas de exploración geotécnica (véase a Guerra, 1989) es típico que las fracturas geológicas en una masa rocosa sean caracterizadas en forma lingüística. Por ejemplo, la Fig. 5, muestra la variable lingüística "rugosidad de una fractura", definida con referencia al índice JRC de Barton (1976). La rugosidad de una superficie o perfil de una fractura mide el grado de irregularidad de sus cotas topográficas. Para una discusión del uso de variables borrosas o lingüísticas en el contexto de la Geotecnia puede consultarse a Paz (1986b, 1987a).

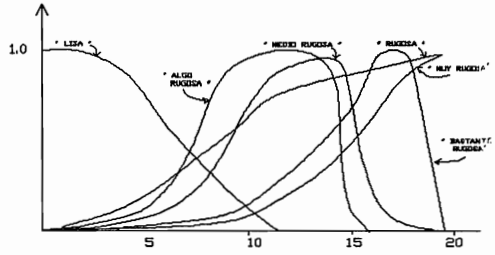


Figura 5. Función de membresía de una Variable Lingüística "Rugosidad de una Fractura".

La borrosidad expresada en términos lingüísticos se refiere a la vaguedad semántica de éstas expresiones. En el caso de variables borrosas, se refiere a la subjetividad asociada con la definición de éstas variables.

e) Ignorancia. El término ignorancia es usado para determinar una situación donde se desconocen las probabilidades y posibilidades de un evento o variable. Sin embargo, por restricciones físicas es posible definir un intervalo límite $[z_I, z_S]$ con membresía $\mu \approx 0$:

$$\mu[z_I, z_S] \approx 0 \qquad (6)$$

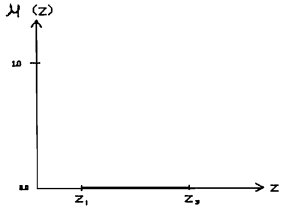


Figura 6. Función de membresía del concepto de Intervalo en la situación de ignorancia.

La Fig. 6 muestra ésta situación. Por ejemplo, en el caso de la evaluación de reservas mineras, es posible definir los límites de los porcentajes de metal de un mineral entre [0, 100], con un soporte no-puntual; aunque ésto no implique que la membresía de este intervalo sea 1.0, sino mas bien que se aproxima a 0.0, ya que los extremos son sumamente raros.

JERARQUIA DE VARIABLES Y SU USO EN GEOESTADISTICA

La tipología de variables desarrollada en la sección anterior puede ser puesta en forma jerárquica, en cuanto a nivel de información, como lo muestra la Fig. 7. La Fig. 7a muestra que el nivel superior está definido por una constante [f(z) = 1.0]. En éste caso, en una estimación dada podemos utilizar el método del Kriging Ordinario. La constante es interpretada como un valor medido. Este caso es el usual en Geoestadística. En la Fig. 7b,

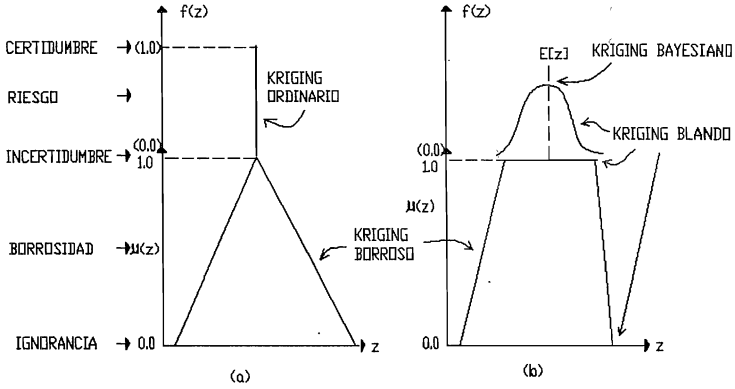


Figura 7. Jerarquía de variables. (a) Con base a una variable borrosa triangular; y, (b) con base a una variable borrosa trapezoidal.

el nivel superior es una distribución de probabilidad. Esta variable aleatoria puede ser inferida (reconstruída por un proceso de estimación) de acuerdo a hipótesis; por ejemplo, la estacionariedad y el caso multigausiano. También, cuando no existen datos, la distribución de probabilidad puede ser estimada en forma subjetiva en base al conocimiento de expertos. La integración de esta información en el Kriging puede ser hecha por medio del Kriging "Blando" (Journel, 1986) usando funciones indicadoras. Cervantes et.al. (1987) presentan un caso estudio en Geotecnia con esta técnica. Otra alternativa es usar un estimado a priori de la media, E(z), y usar el Kriging Bayesiano desarrollado por Omre (1987). Paz (1986a) discute ampliamente las técnicas para evaluar distribuciones de probabilidad subjetivas, asi como sus sesgos de estimación y limitaciones.

En el nivel intermedio de información, Fig. 7b, el intervalo (μ = 1.0) puede ser usado en el contexto del Kriging Blando (Journel, 1986). En forma similar, para el nivel más inferior de información, el intervalo con $\mu \approx 0.0$ es usado por Journel (1986) para restringir al Kriging y que las estimaciones no tomen valores negativos. En la jerarquía mostrada en la Fig. 7, el concepto de riesgo y borrosidad fué hecho independiente para fines ilustrativos. Por ejemplo, para el caso de el intervalo inferior de la ley de un mineral, [0, 100], dependiendo de la escala de observación, estos límites tienen una probabilidad positiva, pero muy pequeña (casi nula), de ocurrencia; por lo tanto el intervalo tiene un μ = 1.0. Por el lado subjetivo, $\mu \approx 0$ puesto que su probabilidad es muy pequeña. Como ya lo habíamos mencionado antes, la relación entre posibilidad y probabilidad está dada por el principio de consistencia. Los dos conceptos se refieren a niveles diferentes de información. Para comprender ésto, cuando definimos una variable borrosa en un punto sin información, y analizamos un intervalo de ésta variable con $\mu < 1.0$, ésto no implica que su probabilidad de ocurrencia sea cero, sino que simplemente ignoramos ésta medida asociada al intervalo.

Las variables borrosas en la base de la Fig.7 pueden ser usadas en una especie de Kriging, aún por desarrollar, que tome en cuenta la borrosidad de las entidades numéricas. El autor está conciente de que Journel(1983) ha utilizado el término Kriging "Borroso" para el caso de distribuciones de probabilidad subjetivas (Kriging Blando). Dada la dificultad de determinar directamente estructuras de correlación subjetivas, más adelante veremos algunas alternativas para tomar en cuenta a las variables borrosas en el Kriging.

El utilizar una variable borrosa con función de membresía triangular o trapezoidal como base de la jerarquía de información, está condicionada a consideraciones de la incertidumbre asociada a un evento o variable. En términos generales, la variable borrosa trapezoidal, con un intervalo de incertidumbre con μ = 1.0, es más consistente con una modelación de procesos.

Desde el punto de subjetividad, es dudoso que los seres humanos

sean capaces de hacer estimaciones de distribuciones de probabilidad consistentes con sus propios conocimientos y creencias. En la opinión del autor, la estimación de variables borrosas es más consistente y ofrece una alternativa para evaluar el tipo de razonamiento aproximado de los expertos. En un plano más estructural, las estimaciones de distribuciones de probabilidad subjetivas son confundidas con estimaciones de variables borrosas. La subjetividad de las estimaciones debe ser modelada por variables borrosas y no por distribuciones de probabilidad. La evidencia experimental en este sentido va en aumento. Paz (1989b) presenta diversos argumentos para apoyar la factibilidad del modelo borroso contra el probabilístico.

HIBRIDIZACION DE VARIABLES

Aunque hay una gran variedad de formas posibles de combinar variables borrosas y aleatorias, en ésta sección solo discutiremos dos de ellas.

a) Variable Híbrida. Definida como la suma de una variable borrosa y una aleatoria (Kaufmann y Gupta, 1985):

$$\mu(z) \; [\; + \;] \; f(z) \qquad (7)$$

La operación con variables híbridas está definida por la convolución híbrida (Kaufmann y Gupta, 1985) que establece que la aritmética de estas variables se da por separado entre variables comunes. Esto es, variables aleatorias con aleatorias y borrosas con borrosas. La Fig. 8 muestra una variable híbrida. Se puede observar en ésta figura que la variable aleatoria posiciona a la variable borrosa sobre el eje z de acuerdo a la función de probabilidad de la variable aleatoria.

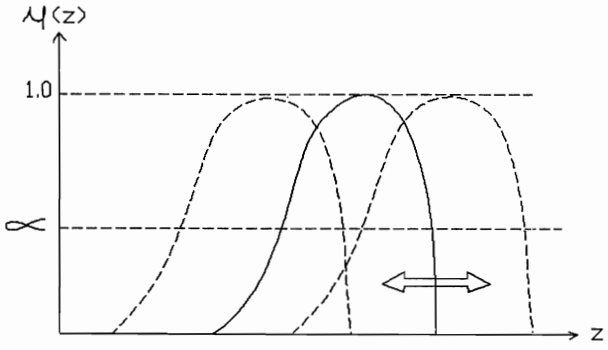


Figura 8. Definición de una Variable Híbrida.

La forma más obvia de relacionar una variable híbrida con las aplicaciones, es para el caso de información contaminada, donde estas variables tienen dos componentes: Una aleatoria y otra borrosa. En la práctica resulta difícil separar éstas componentes por lo que podemos caracterizar datos contaminados como una variable híbrida. Una variable híbrida es producto de operaciones lineales con una mezcla de información objetiva y subjetiva.

b) Variable Aleatoria Borrosa. Este tipo de variable se refiere al caso de un experimento donde los resultados son borrosos (véase a Stein y Talati, 1981, por ejemplo). Así, los valores del experimento no están definidos con certidumbre, sino que son borrosos. La Fig. 9 muestra una variable aleatoria borrosa. Paz (1989a) muestra el uso de éstas variables en Geotecnia.

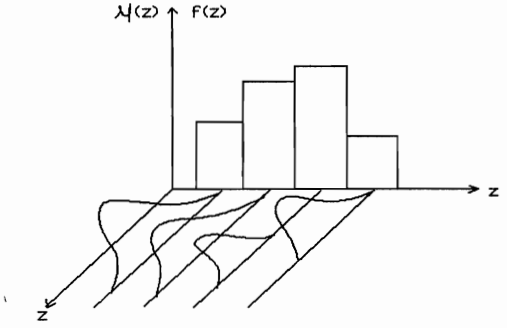


Figura 9. Variable Aleatoria Borrosa.

La relación de estas variables con situaciones prácticas sigue un planteamiento donde los atributos de un fenómeno son evaluados en forma subjetiva y no medidos.

PROCESOS HIBRIDOS

Como un paso natural a la hibridización de variables, los procesos híbridos se refieren a la mezcla de procesos estocásticos y borrosos.

a) Procesos Aleatorios Borrosos. Un caso común de éste tipo está mostrado en la Fig. 10 para un muestreo geotécnico en masas rocosas. Paz y Guerra (1987) han estudiado la correlación espacial del fracturamiento de masas rocosas y han evaluado sus variogramas demostrando asi la existencia de estructuras espaciales del fracturamiento. Tal como se ha discutido anteriormente, en los programas de exploración geotécnica es común caracterizar los atributos de una fractura en términos lingüisticos o borrosos,

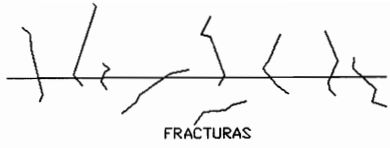


Figura 10. Posicionamiento Espacial de Fracturas Geológicas.

un ejemplo es la rugosidad. De ésta manera, el proceso es aleatorio borroso, puesto que la posición de las fracturas en el espacio es interpretada como una variable aleatoria regionalizada cuyo resultado es una variable lingüística.

b) Procesos Híbridos propiamente dichos. Se refiere al caso de variables híbridas donde su estructura de correlación espacial es producto de la suma de una variable borrosa y una aleatoria. El variograma experimental $\gamma_B(h)$ es la suma de los variogramas de ambos tipos de variables. El variograma de la variable borrosa, $\gamma_B(h)$, se interpreta como el asociado con $\mu = 1.0$. En si, $\gamma_B(h)$ es un variograma producto de variables híbridas donde existe una mezcla de información objetiva y subjetiva. Generalmente, las variables híbridas están asociadas a datos de dudosa reputación.

KRIGING HIBRIDO USANDO VALORES DISCRETOS

El término Kriging híbrido es usado para denotar el uso de una mezcla de variables aleatorias y borrosas en el proceso de estimación. Partiendo del hecho de que en Geoestadística lineal, el valor estimado z* en un punto sin información, está dado por:

$$z^* = \sum_{1}^{n} \lambda_i \, z_i \qquad (8)$$

En este trabajo nos restringiremos a operaciones lineales solamente. La generalización de los conceptos al caso no-lineal, es hecha relativamente simple usando funciones indicadoras.

Si consideramos la situación clásica de un espacio donde tenemos una serie de datos muestrales z_i [$f(z_i) = 1$], Fig. 11, el interés entonces estará centrado en determinar el valor desconocido en un punto cualquiera de ese espacio. Para ésto, a través del Kriging, podemos utilizar la relación (8) y estimar los pesos λ_i a asignar a las variables z_i.

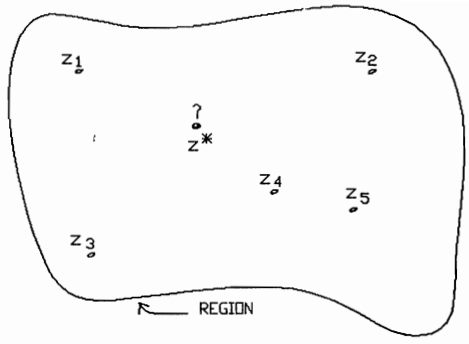


Figura 11. Representación del problema de estimación.

Como primera aproximación del Kriging híbrido, podemos suponer que el variograma de los datos muestrales (variables aleatorias, v.a.) es conocido y que podemos hacer una serie de valuaciones puntuales, en el espacio de estudio, de variables borrosas (v.b.) o lingÜísticas. Bajo la hipótesis de que el variograma de los datos muestrales es igual al de las v.b., tenemos los siguientes casos:

a).- En un punto cualquiera del espacio valuamos una v.b., entonces la z* estimada en un punto de valor desconocido será una variable híbrida (v.h.). Esto es, cada nivel de información permanecerá intacto y no se mezcla información objetiva y subjetiva. A partir de la variable híbrida estimada, podemos evaluar diferentes intervalos de confianza para las variables borrosas y usar ésta información condensada en la toma de decisiones (Kaufmann y Gupta, 1985; Paz y Calles, 1986).

b).- El caso de una v.b. y una v.h. da como resultado una variable híbrida. La jerarquía superior del nivel de información predomina.

c).- La situación de un número n de valuaciones de v.b. o v.h. es similar al caso de las valuaciones únicas. Esto es, el promedio ponderado de variables borrosas es una v.b.

De acuerdo a lo discutido en la sección sobre la jerarquía de variables, es esperado que el tipo de variables borrosas usadas sea trapezoidal o de "moda" múltiple. Así, utilizando el Kriging blando (Journel, 1986), la información subjetiva puede ser usada para condicionar las estimaciones del Kriging a partir del intervalo con $\mu = 1.0$. En ésta situación, las variables borrosas son mezcladas con los datos muestrales solamente cuando el nivel de información es compatible. Sin embargo, en la estimación, ése nivel es puesto por separado.

Un caso muy interesante es el discutido por Paz (1987b) en relación a la situación donde los datos muestrales estan contaminados (posiblemente puedan interpretarse como v.h.), pero no es posible descriminar los valores reales (determinísticos o probabilísticos) de los borrosos, dentro de cada valor muestral. Así, la alternativa es usar algun tipo de valuación externa que cuantifique la confiabilidad de cada una de las muestras. El uso de variables lingüísticas como "confiabilidad alta", "confiabilidad más o menos baja", etc. parece ser un medio apropiado para modelar ésta situación. Calculado el variograma muestral, éste es utilizado para evaluar la confiabilidad asociada a una estimación dada usando la relación (8). Con este procedimiento, la confiabilidad estimada es una medida externa, borrosa, que cuantifica el efecto de la configuración geométrica del muestreo, así como su calidad.

Otra alternativa en el Kriging híbrido es usar el valor con $\mu = 1.0$ para una variable borrosa triangular o el valor medio con $\mu = 1.0$ para una trapezoidal para la estimación del variograma muestral. A partir del variograma calculado, podemos usar los procedimientos ya discutidos. De nuevo, suponiendo el modelo de moda múltiple en la variable borrosa, se puede usar el intervalo con $\mu = 1.0$ y calcular el variograma en base a la aritmética de intervalos. El resultado será un variograma de intervalo, el cual puede ser usado directamente, acondicionado al método kriging para esta situación.

En todos los casos mencionados del Kriging Híbrido, $\gamma(h)$ fue supuesto como una función única de los datos muestrales o de una combinación de estos y de los valores con $\mu = 1.0$ de las variables borrosas. Considérese ahora la situación en la que el variograma muestral es calculado en forma independiente para los datos muestrales y para las variables borrosas, en $\mu = 1.0$. Usando el Kriging como hipótesis, podemos usar los dos variogramas en las estimaciones para producir una v.h. Para el caso borroso, la estimación de las v.b. completas, para $\mu < 1.0$, puede ser hecha suponiendo que el variograma para estos valores es igual al de $\mu = 1.0$. Esto es simple al usar variables semi-simétricas como las triangulares o trapezoidales.

Finalmente, si en vez de usar los conceptos de la Geoestadística para forzar las estimaciones borrosas, consideramos que la función de correlación espacial es borrosa (con cualquier otra métrica diferente de la euclideana) y que el método de estimación también lo es, entonces estaremos haciendo honor a cada nivel de información de acuerdo a su connotación. Sin embargo, la "simplicidad" de las técnicas geoestadísticas se pierde y aun no existe una respuesta a éste problema.
Es importante señalar que las valuaciones de variables borrosas (o de las distribuciones de probabilidad subjetivas) no son independientes de los datos muestrales reales. De hecho, éstas estimaciones pueden ser consideradas como valores interpolados ("krigeados") puesto que usan la información circundante. El volver a Krigear otra vez filtra aun más la información, y el

enfoque se vuelve redundante. Es más simple valuar directamente una variable borrosa en el punto donde se requiere. Para casos de estimaciones masivas de datos, el enfoque discutido es adecuado.

KRIGING HIBRIDO CON VALORES CONTINUOS

Siguiendo a Omre (1987), es dudoso que los expertos puedan hacer un número grande de valuaciones discretas. Resulta más natural el usar funciones borrosas contínuas para éste fin. La Fig. 12 muestra la situación 1-D de una función contínua borrosa de interpolación, z(x), asociada en cada punto a una $\mu[z(x)]$. Esta función es considerada como una estimación proveniente de algún

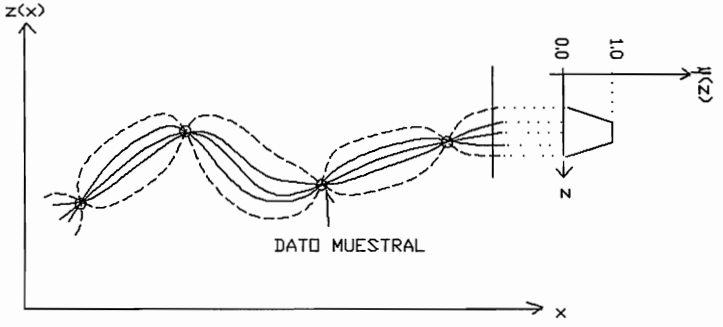


Figura 12. Función Contínua Borrosa de Interpolación.

tipo de "Kriging" mental. El estimador no necesariamente se puede considerar como insesgado y de varianza mínima. En las estimaciones de variables aleatorias subjetivas o borrosas, existen sesgos cognoscitivos y motivacionales que deben ser minimizados para que este estimador mental pueda ser considerado como tal. Paz (1986a, 1989b) discute ampliamente éstos sesgos y su forma de minimizarlos. Nótese que la función contínua borrosa de interpolación usa la información descrita por los datos muestrales z_i, así como algún tipo de información, implícita o explícita, z_{ib} (la b es de borrosa) disponible en el contexto del problema estudiado.

En la Fig. 12 se muestra que la función de interpolación a sido acoplada al concepto de una variable borrosa de moda múltiple. Es dudoso que esto pueda hacerse en la práctica. Una solución simple es utilizar una sola curva de interpolación e interpretarla como el valor medio del intervalo con $\mu = 1.0$. El resto de la variable con $\mu < 1.0$, puede ser asumido como que varía linealmente (lo cual es compatible con las estimaciones mentales) con la distancia de

separación entre puntos muestrales. De ésta forma, la función borrosa queda completamente caracterizada.

Dadas las hipótesis geoestadísticas como estacionariedad, la utilidad de usar una curva contínua borrosa de interpolación es importante porque permite extrapolar la información recuperada a otras regiones con menos información disponible. Esto sugiere que la función de interpolación sea usada en lugares donde la información es abundante y las estimaciones subjetivas pueden ser hechas equivalentes a un proceso tipo Kriging. El problema planteado es del tipo inverso puesto que el resultado del Kriging se conoce pero no las λ_i y el $\gamma(h)$ usados en la estimación. Las alternativas de solución a este problema son:

a).- Asumiendo que el $\gamma(h)$ de los valores muestrales es igual al de la función borrosa, entonces se puede estimar los λ_i asociados a cada dato muestral z_i por un proceso de kriging inverso. Con los λ_i calculados, se procede a evaluar el grado de compatibilidad de los pesos con una variable borrosa de éstos y se itera hasta que la compatibilidad subjetiva sea establecida. La forma del variograma es mantenida y sus parámetros son ajustados. Puesto que existe un número grande de valores interpolados, se puede reconstruir una variable aleatoria borrosa de los pesos en cada punto muestral. El procedimiento de estimación de $\gamma(h)$ es útil cuando éste esta mal caracterizado.

b).- Asumiendo variables borrosas de λ_i y calcular un $\gamma(h)$, para $\mu = 1.0$, compatible con las estimaciones del kriging. La forma y los parámetros de $\gamma(h)$ son dejados abiertos. Esto es útil cuando el $\gamma(h)$ muestral se desconoce.

c).- Hacer borrosos a las λ_i y $\gamma(h)$ e iterar hasta lograr la compatibilidad.

Un problema asociado con los procedimientos discutidos es que los valores estimados pueden interpretarse como:

$$z^* = \sum_{i=1}^{n} \lambda_{ib}\, z_i + \sum_{j=1}^{?} \lambda_{jb}\, z_{jb} \qquad (9)$$

donde z_{jb} es algún tipo de información no declarada en los valores muestrales. Haciendo una interpolación en un espacio restringido, se puede asumir que las z_{jb} son nulas o despreciables.
Otra situación de condicionamiento de información es cuando interpolamos curvas de isovalores, $z > z_c$, Fig. 13. La curva krigeada mentalmente es dependiente de esta acotación. Utilizando el Kriging indicador como proceso de interpolación similar a los casos discutidos, podemos intentar reconstruir las variogramas indicadores.

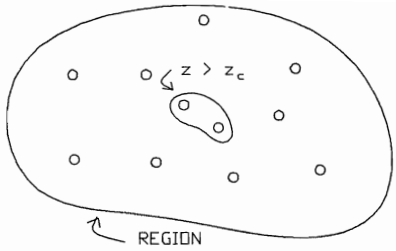


Figura 13. Interpolación de curvas de isovalores.

A final de cuentas, el hacer explícito (por variables borrosas) el estimador mental (exacto, insesgado y de varianza borrosa mínima) es un paso hacia adelante para recuperar toda la información posible de un proceso. Las estimaciones del Kriging y las mentales no difieren significativamente cuando la información es abundante. En este contexto, el término adecuado debe sustituir a óptimo. Un proceso de sub-optimización es más consistente para las situaciones presentadas.

COMENTARIOS FINALES

Como se podrá apreciar, en éste trabajo solo se han mostrado ideas generales para hibridizar la Geoestadística. Esta "punta del iceberg" discutida es con el fin de motivar a los investigadores en el desarrollo de los tópicos emergentes en este entrelazamiento entre la Geoestadística y la Teoría de los Conjuntos Borrosos.

Un punto digno de mención es el referente al concepto de soporte de las estimaciones borrosas. En éste sentido hay una falta de evidencia experimental para decidirse por alguna opción teórica.

Finalmente, debe quedar claro que las pretensiones de ésta hibridización de la Geoestadística es hacer honor a cada uno de los tipos de información usada, sin tener que aumentar artificialmente el nivel de la información subjetiva para mezclarlo con la objetiva.

REFERENCIAS

Barton, N., 1976, The Shear Strength of Rock and Rock Joints, Int. J. Rock Mech. Min. Sci. and Geomech. Abstr., Vol. 13, No.9, pp. 255-279.

Cervantes, J.A., Kim, Y.C. y Farmer, I.W., 1987, Aplicación del Kriging Bayesiano a la Caracterización de Macizos Rocosos en la Mina Subterránea de San Manuel, Arizona, XVII Convención Nacional de la AIMGMMAC, Tomo II, pp. 173-187, Acapulco.

Guerra, M.I., 1989, Programas de Exploración Geotécnica en la Minería, I Simposio Nacional Sobre Mecánica de Rocas Aplicada a la Minería, Universidad de Sonora, Hermosillo.

Journel, A.G., 1983, Fuzzy Kriging, Internal Notes, Departament of Applied Earth Sciences, Stanford University.

Journel, A.G., 1986, Constrained Interpolation and Qualitative Information - The soft Kriging Approach, Mathematical Geology, Vol. , No. , pp.

Kauffmann, A. y Gupta, M.M., 1985, Introduction to Fuzzy Arithmetic, Van Nostrand Co., New York.

Nahmias, S., 1978, Fuzzy Variables, Fuzzy Sets and Systems, No. 1, pp. 97-110.

Omre, H., 1987, Bayesian Kriging - Merging Observations and Qualified Guesses in Kriging, Mathematical Geology, Vol. 19, No. 1, pp. 25-39.

Paz, F., 1986a, Simulación Estocástica de Proyectos de Inversión en la Industria Minera : Un Enfoque Metodológico, Reporte para CAMIMEX, México, D.F.

Paz, F., 1986b, Estabilidad de Taludes: Un Enfoque a través de Variables LingÜísticas, Memorias del 2o. Seminario Nacional sobre Minado a Cielo Abierto, Universidad de Sonora, Hermosillo.

Paz F., 1987a, Rock Mechanics Applications of Fuzzy Sets Theory, Proc. 28th U.S. Rock Mech. Symp., Tucson, Arizona, A.A. Balkema, Rotterdam, pp. 1017-1024.

Paz F., 1987b, Confiabilidad de las Estimaciones Geoestadísticas Usando Diferentes Soportes No-Geométricos, XVII Convención Nacional de la AIMGMMAC, Tomo I, pp. 384-389, Acapulco.

Paz, F., 1989a, Caracterización Aproximada de Masas Rocosas, I Simposio Nacional Sobre Mecánica de Rocas Aplicada a la Minería, Universidad de Sonora, Hermosillo.

Paz, F., 1989b, Modelación de la Incertidumbre en la Programación, Evaluación y Control de Proyectos, Colegio de Ingenieros Civiles de Sonora, A.C., Serie de Divulgación, A Ser Publicado.

Paz, F., y Calles, V.M., 1986, Simulación Híbrida Aplicada a la Evaluación de Proyectos, Memorias del 2o. Seminario Sobre Minado a Cielo Abierto, Universidad de Sonora, Hermosillo.

Paz F., y Guerra, M.I., 1987, Comportamiento Espacial de la Densidad del Fracturamiento, II Reunión Nacional de Mecánica de Rocas, SMMR, México, D.F.

Stein, W.E. y Talati, K., 1981, Convex Fuzzy Random Variables, Fuzzy Sets and Systems, No.6, pp. 271-283.

Zadeh, L.A., 1965, Fuzzy Sets, Information and Control, Vol.8, pp. 338-353.

Zadeh, L.A., 1975, The Concept of a Linguistic Variable and Its Applications to Aproximate Reasoning, Information Science, Vol. 8, No.3, pp. 199-249 (Part.I); Vol. 8, No.4, pp. 301-357 (Part II); Vol. 9, No. 1, pp. 43-80 (Part III)

Zadeh, L.A., 1978, Fuzzy Sets as a Basis for a Theory of Possibility, Fuzzy Sets and Systems, No. 1, pp. 3-28.