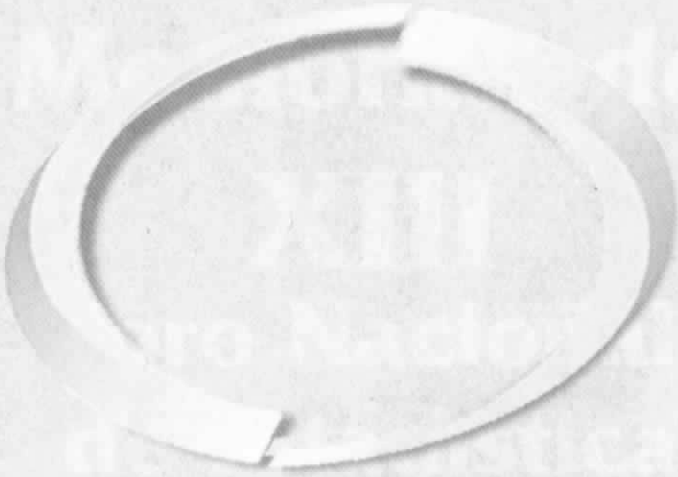
Memorias del XIII Foro Nacional de Estadística

5 AL 9 DE OCTUBRE 1998
ITESM, CAMPUS MONTERREY
MONTERREY, N.L.

INEGI

AME

Memorias del XIII Foro Nacional de Estadística



5 AL 9 DE OCTUBRE 1998
ITESM, CAMPUS MONTERREY
MONTERREY, N.L.

INEGI

AME

Memorias del XIII Foro Nacional de Estadística

Impreso en México
ISBN 970-13-2899-X

Presentación

Durante la semana del 5 al 9 de octubre de 1998 tuvo lugar, en la ciudad de Monterrey, Nuevo León, el XIII Foro Nacional de Estadística. En esta ocasión la organización local corrió a cargo del Instituto de Estudios Superiores de Monterrey (ITESM). Como lo muestra el presente volumen, el programa del evento abordó temas muy diversos sobre el desarrollo y las aplicaciones de los métodos estadísticos.

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

Agradecemos a Sigfrido Iglesias y Alberto Molina por su apoyo en la transcripción de los trabajos. En particular, deseamos destacar la valiosa labor de Alberto en la manipulación de las figuras. También expresamos nuestro reconocimiento al ITESM, Campus Monterrey, cuyo esfuerzo y entusiasmo garantizaron el éxito del Foro. Finalmente, la publicación de esta Memoria no hubiera sido posible sin el generoso apoyo del Instituto Nacional de Estadística Geografía e Informática.

El Comité Editorial

Contenido

Medición de la Calidad del Servicio Considerando Expectativas de los Clientes y su Impacto sobre la Conducta

Pilar E. Arroyo López Aleksander Wójcik Rojek

Juan Gaytán Iniestra

ITESM, Campus Toluca

1 Introducción

El interés por medir y asegurar calidad para los productos de una empresa ha sido un tema que se ha cubierto en diversas áreas como Administración, Estadística e Ingeniería Industrial, desarrollándose enfoques de administración total para la calidad, herramientas estadísticas para su control y procedimientos para la mejora y optimización de procesos y productos. Todas estas herramientas y modelos desarrollados se enfocan hacia procesos de manufactura y aspectos tangibles que reflejen lo adecuado de un producto para el usuario. Actualmente, la contribución del sector manufactura a la economía ha disminuido a favor del sector de servicios, siendo también relevante dentro de este sector, el considerar el aspecto de calidad del servicio. Dado que los servicios son intangibles, resulta difícil proporcionar una medida verificable y objetiva de la superioridad de un servicio, requiriéndose mas que un estándar normativo, un juicio de excelencia y superioridad por parte de los clientes. Ha sido en el área de Mercadotecnia donde se han desarrollado los elementos para medir calidad del servicio, entendiéndose que se pretende medir una actitud de los clientes, lo que implica no una medida del servicio entregado sino una percepción (Gummesson, 1993). Adicional a lo intangible de los servicios, se tiene que no hay una separación entre el resultado del servicio y el proceso completo de entrega de éste, llevando a que calidad del servicio se considere un concepto multidimensional. Las dimensiones que componen el concepto de calidad del servicio, consideran aspectos varios del proceso de servicio como es el caso de la capacidad de responder a las demandas del cliente (respuesta), el trato del proveedor del servicio (empatía) o la habilidad percibida por parte del cliente de que se es capaz de entregar un buen servicio (confiabilidad). (Parasuraman, et al. 1988) identificaron hasta cinco dimensiones para el concepto y propusieron su operacionalización a través del instrumento llamado SERVQUAL. Este instrumento está basado en el modelo de

"discrepancia" el cual considera que la evaluación de la calidad del servicio es la diferencia entre las expectativas y percepciones de los clientes para los diferentes componentes del servicio. En la literatura de satisfacción del cliente, este paradigma establece que entre más elevadas las expectativas en relación al desempeño real, mayor el grado de disconfirmación y menor la satisfacción del cliente en la transacción específica. A diferencia del concepto satisfacción del cliente, el cual es específico a un encuentro de servicio, el de calidad del servicio es un concepto acumulativo y dinámico que se ve afectado directamente por la no satisfacción en un servicio pero también indirectamente por las expectativas del cliente. (Boulding et al. 1993) propusieron un modelo dinámico para calidad del servicio que considera la forma en que los clientes cambian sus percepciones sobre calidad y su conducta según su experiencia en encuentros de servicio sucesivos y las expectativas que tienen al inicio de cada encuentro de servicio. El modelo distingue entre expectativas de dos tipos: normativas (should), las cuales se refieren a lo que debería ocurrir según la información del ambiente competitivo y predictivas (will), que refieren a lo que puede ocurrir según servicios previos y comentarios de otros clientes. El modelo fue validado empleando datos transversales generados en el laboratorio y en el campo. Dado que no se realizaron mediciones de expectativas a diferentes tiempos, varios supuestos se impusieron al modelo para simplificarlo. Las expectativas de los clientes también se han empleado como una base relevante para la segmentación tanto del mercado del consumidor (Webster, 1989) como del mercado industrial (Pitt et al., 1996). La identificación de diferentes grupos de clientes que asignan distintos niveles de importancia a las varias dimensiones del servicio, permite definir estrategias para atender a los clientes según los beneficios que demandan. El objetivo de este trabajo es verificar la validez del modelo dinámico propuesto por Boulding et al. (1993) empleando datos longitudinales respecto a la percepción de la calidad del servicio académico de una institución de educación superior. Así como segmentar a la población de ingreso a la licenciatura respecto a sus expectativas para el servicio educativo. Finalmente, se desea establecer la relación que hay entre la percepción de calidad del servicio con mediciones sobre las intenciones de conducta de los clientes.

2 Metodología

La primera etapa para este trabajo fue desarrollar un instrumento de medición de la calidad del servicio académico para instituciones de educación superior basado en SERVQUAL y proporcionar evidencia de su validez y confiabilidad. Esta etapa es necesaria ya que los reportes del uso de SERVQUAL en diversos ambientes, indican en varios casos la necesidad de adecuarlo al área específica de servicio (Bresinger y Lambert, 1990; Babakus y Boller, 1991; Cronin y Taylor, 1992). Para validar el instrumento, se aplicó una encuesta a una muestra aleatoria de 100 estudiantes en cursos de verano. La segunda etapa fue realizar una segunda encuesta que proporcionara datos sobre expectativas del servicio académico

que ofrecería la institución para dos períodos de tiempo, así como datos sobre percepciones e intenciones de conducta después de un encuentro de servicio. Para ello se aplicaron dos cuestionarios (uno para expectativas predictivas o will y otro para normativas o should) a 120 estudiantes de primer ingreso de licenciatura del ITESM campus Toluca elegidos al azar empleando como marco de muestreo la lista del departamento de escolar. Después de un semestre de estancia en el campus, se aplicaron los mismos cuestionarios de expectativas, más uno adicional sobre percepciones del servicio el cual también incluyó preguntas sobre intenciones de conducta. Estos datos se emplearon para comprobar la validez del modelo dinámico de calidad del servicio propuesto por (Boulding et al. 1993). Además la percepción global de calidad se relacionó con las intenciones de conducta que declararon los estudiantes. En una tercera etapa, los datos sobre expectativas normativas, los cuales son representativos de las demandas de servicio de los estudiantes, se utilizaron para segmentar a la población de primer ingreso.

3 Resultados

3.1 Diseño del instrumento de medición

Para el diseño del instrumento se revisaron las preguntas de SERVQUAL en su traducción al español (Zeithaml et al., 1996), modificándose las preguntas para adecuarlas al servicio académico. Esta lista base de preguntas se enriqueció al incluir preguntas adicionales que cubren aspectos importantes reconocidos en los instrumentos de evaluación de profesores por parte de los estudiantes (Brightman, et al. 1993). El instrumento final de 32 reactivos, incluyó seis dimensiones, las cinco del SERVQUAL (respuesta, confiabilidad, empatía, tangibles y aseguramiento) y una dimensión adicional denominada "involucramiento del cliente durante el servicio". El instrumento se evaluó empleando datos de encuesta, encontrándose evidencia de su confiabilidad interna (alfa de Cronbach global igual a 0.86) y de su validez a través de evaluar su validez facial, convergente y predictiva. Al realizarse un análisis factorial (Hair et al., 1992) sobre los datos de la encuesta, la extracción de seis factores explicó el 63 % de la varianza total, siendo los valores característicos de todos los factores extraídos mayores que 1. Después de una rotación VARIMAX se obtuvo un agrupamiento interpretable de los reactivos en cada factor, identificándose las siguientes seis dimensiones para el concepto calidad del servicio académico:

a) Aseguramiento, que refiere a percibir que el instructor es capaz de dar un buen servicio. Resultó el primer factor extraído y explica el mayor porcentaje de la varianza (16.1 %),

b) interacción profesor-alumno, relacionado a dar una atención personalizada al alumno,

c) tangibles, involucra aspectos físicos del servicio,

d) evaluación, considera aspectos de diseño de tareas y examenes,

e) apoyo en la resolución de problemas, que refiere a preocupación por resolver apropiadamente problemas personales y académicos de los alumnos, y

f) servicios académicos, relacionado directamente con aspectos administrativos.

Las seis dimensiones se correlacionaron con un reactivo global sobre superioridad del servicio académico, para ello se calculó por cada dimensión un promedio ponderado de los reactivos que la forman, utilizando como ponderaciones, las cargas (loadings) que el análisis factorial asignó a cada reactivo. Todos los scores por dimensión dieron correlaciones altamente significantes con este reactivo (menor coeficiente de correlación igual a 0.63), lo cual es evidencia de validez convergente. Se realizó también un análisis de regresión para identificar la importancia de cada dimensión como predictor de la opinión global. Los resultados al aplicar Stepwise indicaron que el modelo de regresión múltiple más adecuado es aquel en cuatro dimensiones (Aseguramiento, Interacción profesor-alumno, Apoyo en resolución de problemas y Servicios administrativos). El coeficiente de determinación ajustado para este modelo fue de 78.21 %. Analizando los coeficientes de regresión parcial y dado que todas las dimensiones están en la misma escala, se determinó como la dimensión más relevante la de aseguramiento (coeficiente de regresión = 0.48), seguida de Interacción profesor-alumno (0.42).

3.2 Ajuste del modelo de Boulding

El modelo utilizado como base en el estudio considera expectativas de dos tipos, las cuales se modifican con cada encuentro de servicio según se describe a continuación.

Las expectativas normativas o "should", se proponen no decrecientes y se incrementan sólo si el servicio entregado las supera, según la siguiente función de actualización

$$SE_{ijt} = SE_{ijt-t} + \beta_{jt}(k \cdot DS_{ijt}) + \varepsilon_{1jt} \tag{1}$$

donde DS_{ijt} se refiere al verdadero nivel de calidad del servicio entregado en el tiempo t respecto a la dimensión j segun el cliente i. SE_{ijt} denota las expectativas normativas o should y ε es el componente aleatorio.

Las expectativas predictivas o "will", cambian en el tiempo dependiendo del servicio entregado, pero cada vez están menos influenciadas por él, y se actualizan de acuerdo a un proceso de suavizamiento representado por la siguiente ecuación, donde α está entre 0 y 1.

$$WE_{ijt} = \alpha_{jt} WE_{ijt-t} + (1 - \alpha_{jt}) DS_{ijt} + \varepsilon_{2jt} \tag{2}$$

Integrando estas dos expectativas, la percepción del servicio se representa de acuerdo a la siguiente ecuación

$$PS_{ijt} = \alpha_{jt} WE_{ijt-t} + (1 - \alpha_{jt}) DS_{ijt} + \gamma_{jt} SE_{ijt-t} + \varepsilon_{3jt} \tag{3}$$

Dado que el verdadero nivel de servicio entregado (DS) es no observable, éste se despeja de la ecuación 2 y se substituye en la 3, después de simplificar se obtiene el siguiente modelo

$$PS_{ijt} = (1 - \alpha_{jt})WE_{ijt} + \alpha_{jt}^2 WE_{ijt-t} + \gamma_{jt} SE_{ijt-t} + \varepsilon_{4jt} \tag{4}$$

Para validar los supuestos del modelo, se compararon los datos correspondientes a expectativas "should" con aquellos de expectativas "will", empleando una t-Student para datos apareados. Se declararon diferencias significativas en todas las dimensiones, lo cual implica una apropiada distinción de ambos tipos de expectativas, así como posibilidades de mejora en todos los aspectos por parte de la institución. Para comprobar la dinámica de las expectativas, se aplicaron los dos cuestionarios a estudiantes de 5. a 8. semestre. Se declararon diferencias mayores entre expectativas "will" y "should" para los estudiantes de 5. a 8 semestre respecto a los de primer ingreso. Las expectativas "should" fueron comparables en ambos grupos, en tanto las "will" fueron significativamente menores para tres dimensiones (prueba Z): aseguramiento, tangibles y resolución de problemas. Estos resultados apoyan las funciones de actualización propuestas para las expectativas. Para verificar la estructura del modelo en la ecuación 4, se realizó un análisis de regresión, considerando tres variables independientes y luego se probaron hipótesis respecto a que los coeficientes se ajustaran a las restricciones implícitas en la ecuación 4. Se ajustó además un modelo no restringido, el cual se formuló como sigue: asumiendo que todos los clientes recibieron un mismo nivel de servicio (DS) durante el semestre, las percepciones de calidad para un estudiante dependerían solamente de las expectativas que se formó ante de entrar al semestre y recibir el servicio académico en el tiempo t. A estas expectativas no se les impone ningún proceso específico de actualización como el sugerido en las ecuaciones 1 y 2, de donde el modelo no restringido tendría la forma de la ecuación 4, en donde a los coeficientes del modelo no se les impone ninguna restricción

$$PS_{ijt} = \beta_1 SE_{ijt-t} + \beta_2 WE_{ijt-t} + \varepsilon_{5jt} \tag{5}$$

Para todas las dimensiones, el coeficiente de determinación ajustado para el modelo de Boulding, fue mayor que para el modelo no restringido. El menor coeficiente de determinación para el modelo no restringido corresponde a la dimensión 6 (Servicios Administrativos), el cual se incrementa a 0.818 al ajustar el modelo restringido de Boulding. Para la dimensión 1 (Aseguramiento) se observaron los mayores coeficientes de determinación: 0.890 para el modelo no restringido y 0.929 para el modelo dinámico de calidad propuesto en la ecuación 5. La hipótesis de que el coeficiente de la primera variable (WE_{ijt-t}) es igual a (1 - raíz cuadrada del coeficiente de la segunda variable) no fue rechazada para ninguna de las seis dimensiones (t-student desde 0.11 a 1.36), lo que apoya la estructura del modelo en la ecuación (4). A diferencia del estudio de (Boulding et al. 1993), el coeficiente para las expectativas "should" resultó ser positivo para todas las dimensiones, lo cual está de acuerdo con la estructura de las ecuaciones 1 y 2, ya que como se hace notar

en el trabajo de (Boulding et al. 1993), las expectativas normativas o should son iguales a las expectativas normativas más las demandas específicas que impone un individuo.

3.3 Impacto de las dimensiones de calidad del servicio a intenciones de conducta

Se calculó un índice global de percepción de calidad del servicio (OS) ponderando el score total de cada dimensión, las ponderaciones usadas fueron los coeficientes de regresión parciales estimados en la última etapa de validación del instrumento (ver sección 2.1). Este índice de percepción global de calidad se relacionó con las intenciones de conducta (BI) declaradas en la encuesta, proponiéndose el siguiente modelo de relación, donde ε es el error aleatorio $BI_k = \varphi OS_{kt} + \varepsilon_k$.

El análisis de regresión indicó que la percepción global de calidad se relaciona significativamente ($P = 0.000$) con las intenciones de recomendar el campus para realizar estudios a otras personas y de realizar estudios posteriores en el mismo campus. Los coeficientes de regresión fueron de 0.1375 para recomendaciones y 0.1841 para estudios posteriores. No se declaró una relación significante entre percepción global de calidad y la intención de no cambiar de campus ($P = 0.512$). Considerando que respecto prestigio de campus y atractivo de la ciudad, el campus Monterrey es generalmente la primera elección de un estudiante, se explica la no-significancia para esta intención de conducta.

3.4 Segmentación

Inicialmente, se forman segmentos a priori en función de la preparatoria de origen del estudiante de primer ingreso, considerando que esta variable estaba directamente relacionada con expectativas. Al realizar un ANOVA en un sentido, no se encontraron diferencias significantes en la medida de tendencia central para las expectativas de estos tres estratos considerados a priori.

Con los datos de expectativas al inicio del semestre, se realizó un análisis de conglomerados empleando un algoritmo jerárquico (método de Ward y distancia de Pearson) agrupándose los 68 datos disponibles en tres conglomerados. Para verificar la segmentación en estos tres grupos, se realizó un análisis de conglomerados no jerárquico (K-means en MINITAB) empleando como semillas observaciones en el centro de los conglomerados generados con el algoritmo jerárquico. Al analizar los centroides de cada conglomerado, se logró la identificación de tres segmentos:

a) Alta Exigencia, que asignó altas expectativas a cada dimensión de calidad del servicio.

b) Exigencia en Servicio Básico, con expectativas intermedias excepto en la dimensión de evaluación la cual no es fundamental en su evaluación de calidad del servicio.

c) Preocupados por Evaluación, con expectativas altas en una evaluación objetiva y justa.

4 Conclusiones y recomendaciones

El instrumento base para realizar este estudio tiene la validez y confiabilidad adecuadas para ser utilizado rutinariamente en el proceso de mejoramiento continuo de la calidad, identificando los puntos débiles del servicio al analizar cada dimensión. Habiéndose identificado como las dimensiones más importantes que definen la calidad global del servicio académico las siguientes: Aseguramiento, Interacción profesor-alumno, Apoyo en resolución de problemas y Servicios Administrativos La percepción de calidad depende significativamente de las expectativas normativas ("should") que tienen los estudiantes y fueron considerablemente elevadas, no observándose el efecto negativo esperado sobre la percepción de calidad. También hay dependencia de las expectativas predictivas ("will"), las cuales al estar por debajo de las expectativas "should" en las dimensiones Aseguramiento, Tangibles y Resolución de problemas indican un diferencial relevante para mejorar la calidad del servicio. En particular las expectativas normativas fueron relevantes en la definición de segmentos a posteriori, útiles para definir la estrategia de enseñanza según las demandas. La segmentación a priori por preparatoria de origen no resultó significante. Los datos longitudinales obtenidos se ajustan al modelo dinámico propuesto por (Boulding et al. 1993) indicando que la percepción de calidad es una función de expectativas normativas y predictivas, las cuales se modifican con cada servicio recibido. Aún cuando este verdadero nivel de servicio no pueda ser medido objetivamente, las relaciones propuestas con las expectativas permiten considerar su influencia en la percepción global de calidad. La calidad global del servicio influye directamente sobre las intenciones de conducta de los estudiantes, en particular sobre las recomendaciones que harán del ITESM y la selección posterior de la institución para estudios posteriores. No encontrándose una relación estadísticamente significante de la calidad global con la lealtad (no cambiarse de campus) a la institución, posiblemente debido a la fuerte imagen de prestigio de otros campus.

Referencias

Babakus, E. and Boller G. W., (1991). Empirical Assesment of SERVQUAL Scale. *Journal of Business Research*, 24, 253-268.

Boulding, W., Kalra, A., Stealin, R. and Zeithmal, V.A. (1993). A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions. *Journal of Marketing Research*, 30, 7-27.

Brensinger, R. and Lambert, D. M. (1990). Can the SERVQUAL Scale Be Generalized to Business-to-Business Services?. *American Marketing Association, Chicago.*

Brightman, H.J., Elliott, M.L. and Bhada, Y. (1993). Increasing the Effectiveness of Student Evaluation of Instructor Data through a Factor Score Comparative Report. *Decision Sciences*, 24, 192-199.

Cronin, J.J.Jr. y Taylor S.A. (1992). Measuring Service Quality: A Re-examination and Extension. *Journal of Marketing*, 56, 55-68.

Gummesson, E. (1993). *Quality Management in Service Organizations.* International Service Quality Association, Stockholm.

Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C. (1992). *Multivariate Data Analysis*, Macmillan, Publishing Company, New York.

Parasuraman, A., Zeithmal, V.A. and Berry, L.L. (1988). SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing*, 64, 12-40.

Pitt, L., Morris, M.H., and Oosthuizen, P. (1996). Expectations of Service Quality as an Industrial Market Segmentation Variable. *The Services Industries Journal*, 16, 1-9.

Webster, C. (1989). Can Consumers be Segmented on the Basis of Their Service Quality Expectations? *Journal of Services Marketing*, 3, 35-53.

Zeithmal , V. A., Parasuraman , A . and Berry, L.L. (1996). *Calidad Total en la Gestión de Servicios.* Díaz de Santos, México

Un Modelo Paramétrico de Regresión para Datos de Tiempos de Vida

Mario Cantú Sifuentes — *CIMA UA de C*
José A. Villaseñor Alva — *Colegio de Postgraduados*
Barry C. Arnold — *University of California, Riverside*

1 Introducción

En muchos contextos, particularmente en Ingeniería es de interés encontrar modelos que describan el comportamiento estadístico de datos de tiempos de vida en presencia de censura y variables explicatorias (o covariables).

Se han usado varios modelos para explicar el efecto que tiene una variable explicatoria sobre el tiempo de vida de una clase de artículos manufacturados. En general, se ha usado un modelo para cada caso particular. En seguida se dan dos ejemplos:

En la explicación del *efecto de longitud*, es decir del efecto que tiene la longitud sobre el tiempo de vida de artículos manufacturados, se usa frecuentemente el principio del eslabón más débil (Galambos, 1978). Tal principio establece:

1. Si la longitud del artículo se divide conceptualmente en partes de cualquier tamaño de interés, se supone que los tiempos a falla de las partes son estocásticamente independientes.

2. El artículo falla cuando cualquiera de las partes falla.

El modelo del eslabón más débil puede expresarse matemáticamente como

$$S(t;x) = (S_0(t))^x,$$

donde $S(t;x)$ es la probabilidad de que un artículo de longitud x sobreviva más allá del tiempo t. $S_0(t)$ es llamada la función de sobrevivencia de referencia, la cual se supone que sea la función de sobrevivencia de un artículo de longitud $x = 1$. Comunmente se usa como referencia a la función de sobrevivencia correspondiente a la densidad Weibull. Si suponemos que $S_0(t)$ denota la función de sobrevivencia de una densidad Weibull con

parámetro de forma $\beta > 0$ y parámetro de escala $\alpha > 0$, entonces el tiempo de vida de un artículo de longitud x tendrá una función de sobrevivencia dada por

$$S_T(t) = \exp\left\{-x\left(\frac{t}{\alpha}\right)^{\beta}\right\} = \exp\left\{-\left(\frac{t}{\alpha_x}\right)^{\beta}\right\},$$

un modelo Weibull con el mismo parámetro de forma, pero con parámetro de escala $\alpha_x = \alpha x^{-1/\beta}$. Bajo el modelo del eslabón más débil, el parámetro de escala refleja el efecto que tiene la longitud sobre el tiempo de vida.

Es claro que la independencia se cumple solo aproximadamente. Además, hay materiales, por ejemplo fibras de carbón al aire (Wolstenholme, 1995), para los cuales el principio del eslabón más débil no se cumple. Una estrategia posible para tales situaciones, consiste en suponer que la función de sobrevivencia corresponde a una distribución más general, la cual contenga como caso particular a la densidad Weibull, y además, permitir que el parámetro de escala sea una función más general de la longitud.

El modelo de Arrhenius (Nelson, 1990) es ampliamente usado para describir el tiempo de vida de una clase amplia de productos como función de su temperatura. En tal clase se incluyen productos tales como aislantes eléctricos, acumuladores, lubricantes y grasas, plásticos y filamentos de lamparas incandescentes. Las suposiciones del modelo son:

1. Dependiendo del producto, para cualquier temperatura, la distribución del tiempo a falla es Exponencial, Weibull o Lognormal.

2. Si se supone que la densidad es Lognormal, la desviación estándar del logaritmo del tiempo de vida se asume constante, mientras que la media se supone ser función de la temperatura.

   Si el modelo es Weibull, se supone que el parámetro de forma es constante y que el parámetro de escala es función de la temperatura.

3. Sin importar la densidad de que se trate, se supone que el logaritmo del parámetro variable es una función lineal del recíproco de la temperatura absoluta.

En este caso, se asume que el efecto de la temperatura sobre el tiempo de vida es reflejado por un parámetro particular y de una manera específica. Sin embargo, la selección de la distribución del tiempo de vida se hace empíricamente.

Como en el caso del modelo del eslabón más débil, el modelo de Arrhenius puede ser incorporado en un marco más general. Las técnicas estándares de inferencia nos pemitirán discriminar entre los modelos en competencia y, quizá, concluir que ninguno de los modelos Weibull, Exponencial o Lognormal son apropiados.

El modelo Gamma generalizado (GG) propuesto por (Stacy, 1962) es útil en tal sentido. La densidad GG es

$$f_T(t) = \frac{\beta}{\alpha\Gamma(k)} \left(\frac{t}{\alpha}\right)^{\beta k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\} I_{(0,\infty)}(t), \tag{1}$$

donde $k > 0$ y $\beta > 0$ son parámetros de forma y $\alpha > 0$ es un parámetro de escala. Las densidades Exponencial ($\beta = k = 1$), Weibull ($k = 1$) y Gamma ($\beta = 1$) son casos particulares de la Gamma generalizada. Más aún, la densidad Lognormal es un caso límite cuando $k \to \infty$ (Prentice, 1974). La familia de densidades (1) proporciona un "supermodelo" conveniente el cual incluye varios modelos en competencia como casos particulares.

2 El modelo propuesto

Para el estudio del comportamiento probabilístico del tiempo de vida T como función de una covariable x proponemos el siguiente modelo general en términos de la función de sobrevivencia de T:

$$S_T(t; x) = P\left(T > t/\alpha(x)\right), \tag{2}$$

donde $\alpha(x)$ es una función continua de x. Entonces, para cada $t > 0$ fijo, $S_T(t; x)$ es una función de x. Siguiendo la discusión de los ejemplos en la Sección 1, proponemos una extensión de la GG a un modelo de regresión en términos de la densidad de T de la forma

$$f_T(t) = \frac{\beta}{\alpha(x)\Gamma(k)} \left(\frac{t}{\alpha(x)}\right)^{\beta k-1} \exp\left\{-\left(\frac{t}{\alpha(x)}\right)^{\beta}\right\} I_{(0,\infty)}(t). \tag{3}$$

En (3) se supone que solo α, el parámetro de escala, refleja el efecto de la covariable x sobre el tiempo de vida del artículo. Esto es, en el modelo general en (2) T tiene la distribución GG como en (1) con parámetro de escala $\alpha = 1$.

Cuando no hay datos censurados en la muestra, el modelo (3) es numéricamente tratable. Sin embargo, cuando la muestra contiene datos censurados, el proceso de estimación por máxima verosimilitud se torna difícil. Esto sucede por que en el caso de censura la función de verosimilitud incluye términos que involucran la función de sobrevivencia para cada uno de los datos censurados. Este hecho puede hacer que la función de verosimilitud se torne numéricamente inestable en el sentido de que los algoritmos usados para estimación por máxima verosimilitud no convergen.

Conforme a nuestra experiencia, una forma de aligerar tales complicaciones, sin perder de vista el efecto que tiene la covariable sobre los parámetros originales, es considerar el

modelo de localización-escala $Y = \log T = \mu + \sigma W$, donde la distribución de W está dada por

$$f_W(w) = \frac{\exp(kw - e^w)}{\Gamma(k)} I_{(-\infty,\infty)}(w).$$

Entonces, la densidad de $Y = \log T$ esta dada por

$$f_Y(y) = \frac{1}{\sigma\Gamma(k)} \exp\left\{k\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\} I_{(-\infty,\infty)}(y), \tag{4}$$

donde $\mu = \log\alpha \quad \varepsilon(-\infty,\infty)$, $\sigma = 1/\beta \quad > 0$ y $k > 0$ son los parámetros de localización, escala y forma, respectivamente. La función de sobrevivencia correspondiente es

$$\begin{aligned} S_Y(y) &= \int_y^\infty \frac{1}{\sigma\Gamma(k)} \exp\left\{k\left(\frac{x-\mu}{\sigma}\right) - \exp\left(\frac{x-\mu}{\sigma}\right)\right\} d \\ &= \frac{1}{\Gamma(k)} \int_{\exp\left\{\frac{y-\mu}{\sigma}\right\}}^\infty z^{k-1}\mathrm{e}^{-z} dz \\ &= \frac{\Gamma\left(k, \exp\left\{\frac{y-\mu}{\sigma}\right\}\right)}{\Gamma(k)}, \end{aligned}$$

donde

$$\Gamma(k,a) = \int_a^\infty z^{k-1}\mathrm{e}^{-z} dz,$$

y $\Gamma(\cdot)$ es la función gamma usual.

Ahora el modelo propuesto tiene la forma equivalente

$$f_Y(y;x) = \frac{1}{\sigma\Gamma(k)} \exp\left\{k\frac{y-\mu(x)}{\sigma} - \exp\left\{\frac{y-\mu(x)}{\sigma}\right\}\right\} I_{(-\infty,\infty)}(y). \tag{5}$$

Cualquier resultado derivado en términos de (5) puede ser transferido facilmente a (3).

3 Metodología de análisis

Para abreviar, se expondrá la metodología de análisis usando el modelo sin reparametrizar para datos no censurados en (3). Sin embargo, lo que se expone es válido para el modelo equivalente en (5) el cual facilita los cálculos cuando existe censura en los datos.

Para estimar los parámetros del modelo es necesario asumir una forma funcional para el parámetro de escala α. Para esto, suponga que se tienen m niveles de la covariable x.

En cada nivel de la covariable hay, digamos n_i tiempos de vida observados $i = 1, 2, \ldots, m$. Suponga que T_{ij} denota el j-ésimo tiempo de vida en el i-ésimo nivel de la covariable. Suponga además que

$$T_{ij} \sim \mathrm{GG}(\alpha_i, k, \beta), \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n_i$$

independientes. Hay entonces $m + 2$ parámetros por estimar. La idea es encontrar los estimadores de máxima verosimilitud (EMV) y buscar una relación simple entre α y x, usando por ejemplo mínimos cuadrados.

Si l_i denota el logaritmo de la función de verosimilitud para el nivel i, $i = 1, 2, \ldots, m$. Entonces, bajo el supuesto de independencia, $l = \sum_{i=1}^{m} l_i$ define el logaritmo de la *verosimilitud global.* Aquí se presenta el problema de que el método de Newton-Raphson (NR) parece no funcionar para resolver las ecuaciones de verosimilitud. Se propone, entonces, encontrar los EMV mediante el perfil de máxima verosimilitud sobre β. Con valores fijos de β, mediante NR, se obtienen EMV parciales para k y α_i, $i = 1, 2, \ldots, m$. Los EMV globales serán los que correspondan al valor de β con el cual se obtenga el valor máximo de l. Las ecuaciones relevantes en el perfil de máxima verosimilitud son:

$$n\psi(k) + \sum_{i=1}^{m} n_i \log \frac{\vec{t}_i}{n_i k} = \beta \sum_{i=1}^{m} n_i \log \tilde{t}_i \tag{6}$$

y

$$\alpha_i^{\beta} = \left(\frac{\vec{t}_i}{n_i k} \right) \quad i = 1, 2, \ldots, m, \tag{7}$$

donde $\vec{t}_i = \sum_{j=1}^{n_i} t_{ij}^{\beta}$, $\tilde{t}_i = \left(\prod_{j=1}^{n_i} t_{ij} \right)^{1/n_i}$ y $n = \sum_{i=1}^{m} n_i$.

El algoritmo es sencillo, para cada valor fijo de β, se resuelve para k, la ecuación (6). El valor estimado de α, para el valor fijo de β se obtiene directamente de la expresión (7). El procedimiento se repite hasta encontrar el valor de β que maximice l. Para obtener estimaciones iniciales de los parámetros se pueden usar los estimadores de momentos propuestos por (Stacy y Mihram, 1965).

Una vez obtenidos los EMV de las α_i, $i = 1, 2, \ldots, m$, se puede construir un diagrama de dispersión de $\log \alpha_i$ v.s. $\log x_i$ para obtener una idea de la relación funcional entre α y x.

Una forma posible para α es la función potencia $\alpha(x) = cx^b$, o en forma equivalente $\log \alpha = \log c + b \log x = \beta_0 + \beta_1 \log x$; donde $\beta_0 = \log c$ y $\beta_1 = b$. En tal caso, habrá 4 parámetros por estimar. Las ecuaciones relevantes para el perfil de máxima verosimilitud son ahora

$$kc^{\beta} = \frac{1}{n} \sum_{i=1}^{m} \frac{\vec{t}_i}{x_i^{b\beta}}, \tag{8}$$

$$\beta\left(\frac{1}{n}\sum_{i=1}^{m}\frac{\vec{t}_i}{x_i^{b\beta}}\right)\sum_{i=1}^{m}n_i\log x_i=\sum_{i=1}^{m}\frac{\vec{t}_i\log x_i^{\beta}}{x_i^{b\beta}}, \tag{9}$$

$$\beta\sum_{i=1}^{m}n_i\log\tilde{t}_i=n\psi(k)+n\log\left(\frac{1}{kn}\sum_{i=1}^{m}\frac{\vec{t}_i}{x_i^{b\beta}}\right)+b\beta\sum_{i=1}^{m}n_i\log x_i, \tag{10}$$

donde $\vec{t}_i$, $\tilde{t}_i$ y n están definidas como antes.

Para este caso, el algoritmo es nuevamente muy simple: Para cada β fija, resuelva la ecuación (9) para b; con tales valores, resuelva la ecuación (10) para k. Finalmente obtenga el estimador parcial de c de (8). Las estimaciones iniciales de c y b pueden ser las obtenidas del ajuste por mínimos cuadrados de la recta $\log\alpha=\log c+b\log x$. Como estimaciones iniciales de k y β se pueden usar las provenientes del antes citado método de momentos de Stacy y Mihram.

Para el caso en que la muestra contenga datos censurados se propone usar el modelo reparametrizado en (5), y aunque las ecuaciones de verosimilitud se ven más complicadas, el método de (NR) no presenta problemas fuertes en su implementación. El algoritmo de análisis es similar; sin embargo, resulta más conveniente realizar los perfiles de máxima verosimilitud sobre el parámetro k en vez de sobre el parámetro β.

4 Aplicaciones

En esta sección se analizan, usando el modelo propuesto, dos conjuntos de datos.

4.1 Datos de tiempos a falla de hilaza

Estos datos fueron producidos por (Picciotto, 1970) al estudiar las características de la fatiga a la tensión de hilaza considerando varios niveles de longitud. Las longitudes fueron 30, 40, 50, 60, 70, 80, 90 y 100 centímetros con tamaños de muestra 99, 99, 100, 99, 100, 100, 100 y 100, respectivamente.

Al ajustar por mínimos cuadrados la recta $\log\alpha=\beta_0+\beta_1\log x$ se obtienen los estimadores $\tilde{\beta}_0=7.99$ y $\tilde{\beta}_1=-1.1$ con un coeficiente de determinación de 0.98. Con lo que es razonable suponer que $\alpha(x)=cx^b$. Además, efectuando una prueba de razón de verosimilitud generalizada para validar tal hipótesis se obtuvo que $-2\log\lambda=6.60$. Este valor es más pequeño que 12.59 el percentil 95 de una χ^2_6. Con lo que no rechazamos la hipótesis nula $\alpha(x)=cx^b$. Este conjunto de datos ha sido estudiado con anterioridad por (Cantú, et al. 1998) donde se encontró evidencia a favor del modelo de regresión Gamma. Esto es $\beta=1$ en (3). Con esto, resulta razonable probar la hipótesis $\beta=1$. Nuevamente, usando la prueba de razón de verosimilitudes generalizada se obtiene $-2\log\lambda=0.24$; este

valor es más pequeño que el valor crítico 3.84, el percentil 95 de una χ^2_1. Por lo tanto no rechazamos tal hipótesis a un nivel de significancia del 5%. Para este tipo de datos, estamos tentados a creer que se cumple el principio del eslabón más débil. Entonces, probamos la hipótesis nula $k = 1$ (el modelo de regresión es un Weibull). Para este caso, obtenemos que $2\log\lambda = 22.84$ el cual es mayor que el valor crítico 3.84 al 5% de significancia. Por lo tanto la hipótesis $k = 1$ es rechazada.

Ajustando el modelo (3) con $\beta = 1$, obtenemos: $\hat{c} = 2202.35$, $\hat{b} = -1.0985$ y $\hat{k} = 3.438$ con $l = -4080.35$.

4.2 Datos de tiempos de vida de aislante de motores

(Nelson y Hahn, 1972) analizaron datos censurados provenientes de una prueba de vida acelerada de aislante de motores. Su análisis se basó en 30 tiempos de vida, 10 para cada nivel de temperatura. Los niveles de temperatura considerados fueron 170, 190 y $220°C$, con 3, 5 y 5 observaciones censuradas, respectivamente. Para el análisis Nelson y Hahn usaron el modelo de Arrhenius Lognormal. Esto es, supusieron que para cualquier temperatura: la distribución del tiempo de vida es Lognormal, la desviación estándar del logaritmo del tiempo de vida es constante y la media de el logaritmo del tiempo de vida es una función lineal del recíproco de la temperatura absoluta (T_a). Es decir se supuso que $x = 1/T_a$. Además, por consideraciones de escala, consideraron conveniente trabajar con $x = 1000/T_a$. Con lo que los valores de la covariable son 2.256,2.159 y 2.028 para 170, 190 y $220°C$, respectivamente.

Aplicando el modelo propuesto en (5), poniendo $\mu(x) = \beta_0 + \beta_1$ con $x = 1000/T_a$ como covariable, se obtuvo $\hat{k} = 0.335$, $\hat{\sigma} = 0.143501$, $\hat{\beta}_0 = -12.5258$ y $\hat{\beta}_1 = 9.42932$ con $l = -22.7364$. Si fijamos $k = 1$, los estimadores correspondientes son: $\hat{\sigma} = 0.361397$, $\hat{\beta}_0 = -11.9539$ y $\hat{\beta}_1 = 9.06729$ con $l = -22.9543$. Note que los valores del logaritmo de la verosimilitud maximizada para $k = 0.335$ y para $k = 1$ no son muy diferentes. Entonces es razonable probar la hipótesis $k = 1$. Nuevamente, usando la prueba de la razón de verosimilitud generalizada, se obtiene $-2\log\lambda = 0.4353$. Dado que 0.4353 es menor que el valor crítico 3.84, el percentil 95 de una χ^2_1, no rechazamos tal hipótesis al 5% de nivel de significancia.

5 Discusión

Es interesante notar que para el caso de los datos de hilaza, el principio del eslabón más débil no se cumple, esto es el modelo de Weibull no es el más apropiado. Además, nuestro análisis de los datos de hilaza produce resultados que están en desacuerdo con los encontrados por (Arnold et al. 1996), donde ellos argumentan que tal conjunto de datos puede ser descrito apropiadamente por un modelo de regresión Weibull con $\alpha(x) = \delta x^{-2/\beta}$.

Tal modelo fué rechazado por nuestro procedimiento de selección de modelos. Aquí se concluyó que el modelo más apropiado es el modelo de regresión Gamma.

El conjunto de datos de aislamiento también fué analizado con anterioridad. En su estudio (Nelson y Hahn, 1972) usaron un modelo de regresión Lognormal. Conforme a nuestro estudio, concluimos que el modelo más apropiado es el Weibull; esto es $k = 1$.

Referencias

Arnold, B. C., Castillo, C. and Sarabia, J. M. (1996), Modelling the Fatigue Life of Longitudinal Elements. *Naval Research Logistics*, **43**, 885–895.

Cantú, S. M., Arnold, B. C., and Villaseñor, J. A. (1998) A Distributional Study of Lifetime Data From Longitudinal Specimens. *Cuadernos de Investigación de la U. A. de C.*

Galambos, J. (1978), *The Asymptotic Theory of Extreme Order Statistics.* Wiley–Interscience, New York.

Nelson, W. B. and Hahn, G. L. (1972). Linear Estimation of Regression Relationship form Censored Data, Part I: Simple Methods and their Applications (with discussion). *Technometrics*, **14**, 247–76.

Nelson, W. (1990), *Accelerated Testing: Statistical Models, Test Plans and Data Analyses.* New York: John Wiley.

Picciotto, R. (1970), Tensile Fatigue Characteristics of Sized Polyester/Viscose Yarn and their Effect on Weaving Performance. Thesis presented to North Carolina State University, at Raleigh, N.C.

Prentice, R. L. (1974), A Log–Gamma Model and its Maximum Likelihood Estimation. *Biometrika*, **61**, 539–544.

Stacy, E. W. (1962), A Generalization of the Gamma Distribution. *Annals of Mathematical Statistics*, **33, 3**, 1187–1192.

Stacy, E. W. and Mihram, G. A. (1965) Parameter Estimation for a Generalized Gamma Distribution. *Technometrics*, **7**, 349–58.

Wostenholme, L. C., (1995) A Non Parametric Test of the Weakest–Link Principle. *Technometrics*, **37**, 169–175.

Optimización de Multirrespuesta

Román De la Vara Salazar y Jorge Domínguez Domínguez

CIMAT, Guanajuato

1 Introducción

En la literatura de diseño de experimentos y de optimización de procesos se hace mucho énfasis en ejemplos y problemas que consideran el estudio de *una sola* propiedad o característica de calidad del producto o proceso, cuando lo típico es que la calidad del producto dependa del valor que asumen más de una de sus propiedades, ver por ejemplo (Myers & Montgomery, 1995). Por ejemplo, un alimento tiene varias propiedades como: digestibilidad, textura, pH, sabor, aspecto, etc., y todas tienen su importancia para que el alimento sea bien aceptado por los consumidores. Muchas veces el optimizar una de las características del producto hace que las otras propiedades se vean afectadas y que el resultado sea un producto con peor calidad global que antes. De aquí la importancia de contar con técnicas que sirvan para optimizar de manera simultánea todas las respuestas de interés, o al menos localizar el punto de operación del proceso en el que todas las variables tienen el "mejor desempeño posible". Así pues, típicamente lo que se obtiene con un método es una solución de compromiso, puesto que algunas características se afectan un poco en aras de lograr mejores resultados en otra de las propiedades.

Se han propuesto varios métodos para optimizar de manera simultánea más de dos respuestas, entre los que se encuentran: Derringer y Suich (1980), Khuri y Conlon (1981), Ames, et al (1997) y el método gráfico. De éstos, el más antiguo es el método gráfico y el más utilizado es el de la función de deseabilidad de Derringer y Suich. Estos métodos, así como referencias adicionales, se describen en De la Vara y Domínguez (1998). Todos los métodos suponen que las k respuestas $(Y_1, Y_2, \ldots, Y_k)$ a optimizar se pueden modelar adecuadamente por un modelo polinomial de segundo orden en términos de p factores de control $(X_1, X_2, \ldots, X_p)$, es decir, partimos de que se tienen los k modelos estimados dados por

$$\widehat{Y}_m = \hat{\beta}_{0m} + \sum_{i=1}^{p} \hat{\beta}_{im} X_i + \sum_{i=1}^{p} \hat{\beta}_{iim} X_i^2 + \sum \sum_{i<j}^{p} \hat{\beta}_{ijm} X_i X_j \qquad (1)$$

donde $m = 1, 2, \ldots, k$, los cuales se ajustaron sobre datos experimentales y cada uno explica de manera adecuada la respuesta que le corresponde. Ninguno de los métodos

hasta ahora propuestos toma en cuenta todos los aspectos relevantes al problema como son: la respuesta predicha, el error estándar, la importancia relativa de las respuestas y las especificaciones.

Por razones de espacio, aquí nos limitamos a describir el método gráfico que consideramos que no ha sido apreciado en lo que vale, al resultar bastante competitivo en relación a los métodos analíticos existentes. El método gráfico proporciona una idea intuitiva de cómo es la multirrespuesta y siempre es recomendable utilizarlo como apoyo cuando se quiere aplicar un método analítico, ya que minimiza la posibilidad de obtener una solución incorrecta o una solución que es óptimo local y que no es la mejor dentro de la región experimental. Muchas veces se llegará a que la solución gráfica supera a la solución del método analítico.

2 Descripción del método gráfico

Quizás lo primero que se ocurre al optimizar varias respuestas es superponer todas las superficies sobre la región experimental y localizar adentro de ella subregiones en las cuales todos los modelos predicen valores aceptables para las respuestas. Cuando se tienen solo dos factores de control (X_1 y X_2) es bastante fácil superponer las superficies de respuesta en su forma de contornos sobre la región experimental, que en este caso se puede dibujar como un cuadrado o círculo centrado en el origen. El problema es que cuando se tienen más de dos factores las superficies no se pueden dibujar de una sola vez sobre toda la región experimental; aunque se pueden estudiar y superponer tomando dos factores a la vez y fijando a los restantes. Por ejemplo, si son tres los factores de control la región experimental es un cubo (o esfera) centrado en el origen y se puede dibujar una superposición de las k superficies sobre cada corte del cubo, lo que implica fijar uno de los factores de control, ver figura 1. En esta figura se fija el factor X_2 en X_{20} y los contornos que vemos sobre esta lámina son en realidad un corte de los contornos originales que están en tres dimensiones. Con más de tres la situación se complica, pero se preserva la misma idea.

El método gráfico de optimización simultánea se puede aplicar en el paquete *Design Expert* el cual, a partir de los datos experimentales, ajusta los k modelos y es posible sobreponer las superficies sobre diferentes cortes de un cubo o hipercubo. Por ejemplo, para $k = 3$ en cada corte del cubo se dibujan dos curvas de nivel por cada respuesta que corresponden a sus especificaciones y el software señala la subregión donde todas las respuestas predicen valores factibles, es decir, adentro de las especificaciones. Posteriormente la región factible se va reduciendo, estrechando poco a poco las tolerancias de las variables más importantes, hasta tener una región pequeña en la que todos los modelos predicen mejores valores de las respuestas. Si las variables son igualmente importantes, sus tolerancias se estrechan en la misma proporción, y cuando una variable es más importante sus tolerancias se estrechan en mayor proporción, lo que acaba dando más peso a esta variable.

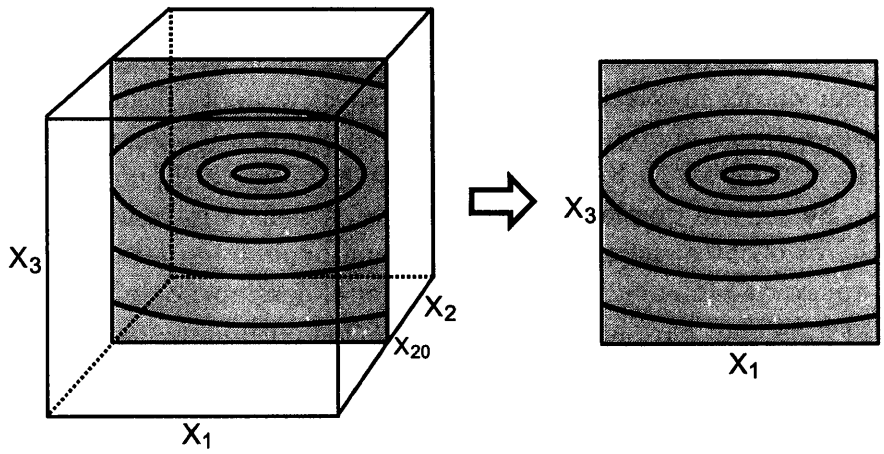


Figura 1: Aspecto de los contornos de la hipersuperficie en $X_2 = X_{20}$

Cuando esta pequeña región es la misma en los tres posibles cortes del cubo, lo que se sabe porque el valor que se fija en cada factor se mantiene a la hora de fijar otro factor, se ha encontrado un punto óptimo simultáneo que es el centro de dicha región factible, ver figura 2.

Muchos autores, ver por ejemplo, Del Castillo (1996), han criticado la dificultad de aplicar el método gráfico cuando se tienen más de dos factores de control. En este trabajo se muestra que el método es bastante fácil de aplicar hasta con tres factores; con cuatro es el doble de difícil pero todavía es posible. Además, la gran mayoría de los casos prácticos de optimización se hacen con cuatro o menos factores de control, siendo el caso de tres el más frecuente.

3 Aplicación del método gráfico

3.1 Ejemplo : Optimización de neumáticos

En esta sección se analiza el método gráfico utilizando del ejemplo de las llantas discutido por Derringer y Suich (1980), que consiste en encontrar la combinación óptima simultánea de tres ingredientes de un compuesto de las bandas de llanta (silica (X_1), silane (X_2) y sulfur (X_3)), para cuatro variables de interés cuyos nombres y especificaciones son:

$$\begin{aligned} Y_1 &> 120, \textit{Índice de abrasión} \\ Y_2 &> 1000, \textit{Módulo 200\%} \\ 400 &< Y_3 < 600, \textit{Elongación} \\ 60 &< Y_4 < 75, \textit{Dureza.} \end{aligned}$$

Si bien las dos primeras variables no tienen límite de especificación superior, desde el punto de vista práctico se considera que no hay ninguna ganancia adicional en estas propiedades si Y_1 y Y_2 toman valores mayores a 170 y 1300, respectivamente. En este sentido se pueden considerar estos últimos números como los valores objetivo (targets) de dichas variables. Por su parte las variables Y_3 y Y_4 toman como valores objetivo el punto central de su rango de tolerancia. Se supone que las cuatro variables tienen la misma importancia desde la perspectiva de los clientes, así que en todos los métodos se les dará (o tratará de dar) el mismo peso, aproximadamente. La idea es observar el desempeño de los diferentes métodos en cuanto a su habilidad para atrapar la combinación óptima de los ingredientes.
Utilizando el paquete *Design Expert* se localiza la región que mejor cumple con los requerimientos de las cuatro variables de respuesta y los resultados se muestran en la figura 2. Esta región se logra dibujando dos curvas de nivel de la superficie descrita por el modelo ajustado en cada respuesta, cuyos valores son

$$\begin{aligned} Y_1 &\rightarrow 131 - 170 \\ Y_2 &\rightarrow 1192 - 1300 \\ Y_3 &\rightarrow 464 - 536 \\ Y_4 &\rightarrow 64.8 - 70.2. \end{aligned}$$

Esto equivale a una reducción de 64% del tamaño del rango de especificación original, excepto en la variable Y_1 donde solo se reduce 22%. Esto es, como las variables son igualmente importantes se trataba de reducir sus rangos en la misma proporción para sacrificar a todas por igual. Sin embargo, no fue posible reducir más el rango de la primera variable sin perjudicar fuertemente el desempeño de las otras tres, en particular el de la tercera. En otras palabras, en este ejemplo es necesario sacrificar a la variable Y_1 en aras de un mejor desempeño de las demás variables; este hecho también lo registran los otros métodos.
Los resultados de una solución gráfica óptima bajo las condiciones arriba descritas se observan en la figura 2 y se resumen en la tabla 1. El punto óptimo simultáneo gráfico es la combinación de niveles dada por (x_{10}=-0.217, x_{20}=0.400, x_{30}=-0.725). En la Tabla 1 también se muestran los valores predichos para cada respuesta sobre el punto óptimo.

Tabla 1. Resultados de optimización por el método gráfico

$\hat{Y}_1(\mathbf{x}_0)$	$\hat{Y}_2(\mathbf{x}_0)$	$\hat{Y}_3(\mathbf{x}_0)$	$\hat{Y}_4(\mathbf{x}_0)$	x_{10}	x_{20}	x_{30}
131.61	1257.33	466.28	69.91	-0.217	0.400	-0.725

4 Discusión

En la Tabla 2 se resumen los mejores resultados para el problema de los neumáticos, obtenidos con los cuatro métodos mencionados en la introducción y reportados en De la Vara

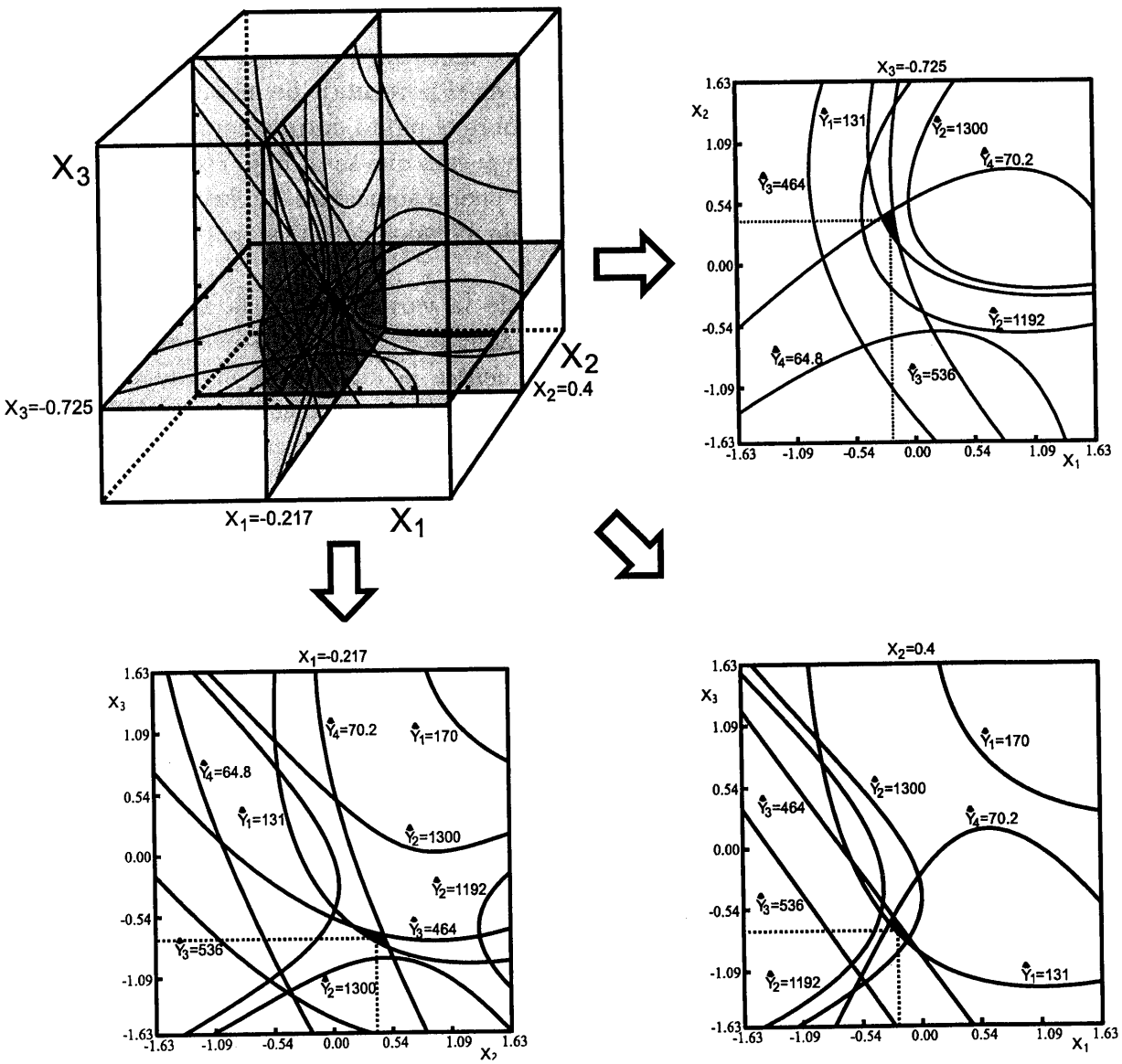


Figura 2: Optimización simultánea por método gráfico

y Domínguez (1998). Con DS nos referimos al método de Derringer y Suich (1980); KCa1 y KCa2 se refieren al método las dos funciones de distancia sugeridas por Khuri y Conlon (1981) y restringiendo la región a las condiciones de DS; AMES es la solución dada por Ames et al. (1997) de este mismo problema y AMES1 es la solución que resulta al corregir los valores nominales que utilizan los autores en la función de pérdida.

Las mejores soluciones son las obtenidas con los métodos de la función de deseabilidad (DS) y el gráfico (tabla 2), quedando muy parejas en el porcentaje de mediciones afuera de especificaciones que se espera tener en el futuro sobre el punto óptimo correspondiente. El método gráfico tiene la ventaja de ser más flexible puesto que se trabaja con las superficies de respuesta originales y en la escala original. Las peores soluciones son las llamadas KCa2 y AMES, puesto que reportan predichos demasiado alejados de los valores objetivos en las variables Y_3 y Y_2, respectivamente (tabla 2).

Ninguno de los métodos considera, además de la variabilidad y la media de $\hat{Y}_i$, su habilidad para cumplir con las tolerancias, de aquí que todas las soluciones dadas en la tabla 2 sean susceptibles de mejora si se encuentra la manera de involucrar los tres aspectos. Es decir, algunos de los métodos utilizan sólo la media, otros únicamente la media y la variabilidad, alguno más sólo la media y las especificaciones, pero ninguno las tres cosas. Como resultado de esto, en los puntos óptimos simultáneos reportados se predice un excesivo número artículos defectuosos en el futuro, en términos de algunas de las variables. La idea entonces es determinar puntos óptimos que predigan un mínimo de defectuosos en las mediciones futuras de todas las respuestas.

Tabla 2. Resultados con los métodos originales

	punto óptimo			predichos			
método	x_{10}	x_{20}	x_{30}	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$
Gráfico	-0.217	0.400	-0.725	131.61	1259.73	466.28	69.91
DS	-0.540	0.150	-0.869	129.43	1300.00	466.05	68.03
AMES	-0.460	-0.284	-0.525	122.73	1069.10	500.00	67.50
AMES1	0.060	0.536	-0.545	139.90	1325.02	422.72	70.19
KCa1	-0.179	0.864	-1.137	127.41	1334.91	487.55	71.24
KCa2	-0.187	0.279	-0.121	139.00	1250.00	419.21	70.31

En otras palabras, no basta cumplir los requerimientos, sino que, para evitar artículos defectuosos, idealmente, se debe cumplir que el intervalo de confianza para la predicción futura sobre el punto óptimo se encuentre completamente adentro de las especificaciones y no sólo parcialmente como ocurre en algunas variables. Por ejemplo, la solución DS predice un valor de 129.43 para la primera variable, pero dado su error estándar, el límite inferior del intervalo del pronóstico es de alrededor de 113, es decir, se predice un porcentaje importante de defectivos. Lo mismo pasa con la segunda variable en esta solución, que aunque toma su valor óptimo de 1300 en promedio, el error estándar de las futuras mediciones es de

alrededor de 420, por lo que también se tendrán defectuosos por este concepto. La pregunta es: ¿cómo obtener el punto óptimo simultáneo con cada método, que, además, tenga el mejor centrado del intervalo de predicción dentro de las especificaciones?

Referencias

Ames, A. E.. Mattucci, M. Stephen, M., Szonyi,G. y Hawkins, D. M. (1997). Quality Loss Functions for Optimization Across Multiple Response Surfaces. *Journal of Quality Technology* 29, pp. 339-346.

De la Vara, R. y Domínguez, J. D. (1998). Metodologías de Superficie de Multirrespuesta. *Comunicación Técnica* No. I-98-08 (PE-CIMAT).

Del Castillo, E. Multiresponse Process Optimization via Constrained Confidence Regions. *Journal of Quality Technology* 28, pp. 61-70.

Derringer, G. and Suich, R. (1980). Simultaneous Optimization of Several Response Variables. *Journal of Quality Technology* 12, pp. 214-219.

Design-Expert (1991). Stat-Ease, Version 2.02 Minneapolis, MN.

Khuri, A. y Conlon (1981). Simultaneus Optimization of Multiple Responses Represented by Polynomial Regression Functions. *Technometrics* 23, pp. 363-375.

Myers, R. H. y Montgomery, D. C. (1995). *Response Surface Metodology: Process and Product Optimization Using Designed Experiments.* John Wiley & Sons, New York.

Busqueda Tabú Mediante la Generación de una Estructura de Vecindades Aleatorias para la Identificación de Prototipos de Grupos Previamente Clasificados

Sergio De los Cobos
UAM-Iztapalapa, México

John Goddard
UAM-Iztapalapa, México

Blanca Rosa Pérez
UAM-Iztapalapa, México

Miguel Angel Gutiérrez
UAM-Azcapotzalco, México

1 Introducción

Los métodos para obtener una clasificación mediante una partición óptima de un conjunto de individuos, pertenecen a la clase de los problemas de optimización combinatoria. El problema de particionamiento se puede enunciar de manera general como: "Dada una colección de objetos, se desea clasificarlos en clases (grupos o conglomerados) bien diferenciados entre sí, de forma tal, que las clases sean lo más homogéneas (respecto a algún criterio preestablecido) entre sí". Este problema es de fácil presentación pero, en la práctica existen diferentes tipos de dificultades, tanto de índole matemática como computacional, respecto a esto último, podemos mencionar que el problema de clasificación mediante particionamiento es del tipo NP-hard. En general, se puede observar que los métodos tradicionales para obtener una clasificación mediante una partición para conjunto de datos "grandes", obtienen soluciones suboptimales. Por lo que ha llevado a muchos investigadores a utilizar técnicas "inteligentes" que llamaremos de "programación estocastica en optimización", como son entre otras: Búsqueda Tabú, Recocido Simulado, Algoritmos Genéticos, Redes Neuronales, etc. (De los Cobos et al. (1997), Murillo y Trejos (1996), Trejos (1996) y Piza y Trejos (1996)).

Lo que consideramos es la idea de enfocar la atención sobre las estrategias más promisorias e intentar encontrar una solución "buena" sin explotar todas las posibles estrategias.

En este trabajo presentamos algunas ideas de una técnica para la solución del problema de clasificación mediante particionamiento, pero con conjuntos previamente clasificados (que son utilizados para medir la eficiencia o "bondad" de la técnica propuesta) y utilizan-

do híbridos de diferentes técnicas. Este estudio se motivo por el trabajo De los Cobos et al. (1997), donde nos preguntábamos el porqué utilizando la misma métrica, pero diferentes técnicas se obtenía el mismo valor de la función objetivo, siendo los particionamientos resultantes diferentes, lo que nos llevo a plantear dos objetivos: a) ¿Cómo identificar los elementos prototipos (característicos) de una partición?, y b) ¿Cuáles son los particionamientos más robustos?.

En este trabajo se utiliza para la identificación de prototipos de clases previamente identificadas, un marco de generación de vecindades de tipo aleatorio. Lo anterior se realizó con el propósito dual de por un lado ver la bondad de la técnica Tabú para este tipo de problemas con esta metodología y como siguiente paso, el de proponer un tipo de robustez del método para cualquier tipo de clasificación. En este trabajo se presenta la metodología utilizada, así como resultados promisorios.

2 Búsqueda Tabú

La técnica de la Búsqueda Tabú es introducida en su forma actual por Glover (1989 y 1990), la cual se ha utilizado para resolver problemas de optimización de gran escala. Dicho procedimiento, está diseñado para guiar a otros métodos (o a procesos componentes) para escapar de la optimalidad local. La Búsqueda Tabú tiene como antecedentes a varios métodos diseñados para cruzar fronteras de factibilidad o de optimalidad local; sistemáticamente impone y relaja restricciones para permitir la exploración de regiones de otra manera prohibidas.

En General, se puede decir que la filosofía de la Búsqueda Tabú está basada en manejar y explotar una colección de principios para resolver problemas de manera inteligente, la cual tiene como elemento principal el uso de memoria flexible. Desde cierto punto de vista, el uso de memoria flexible involucra el proceso dual de crear y explotar estructuras para tomar ventaja de la "historia", por lo que, se combinan las actividades de adquisición y mejoramiento de la información. La Búsqueda Tabú se funda en tres puntos principales:

1. El uso de estructuras de memoria basadas en atributos diseñados para permitir criterios de evaluación e información de búsqueda histórica.

2. Un mecanismo asociado de control, mediante el empleo de estructuras de memoria, basado en el interjuego entre las condiciones que restringen y liberan al proceso de búsqueda (envuelto en las restricciones tabú y el criterio de aspiración).

3. La incorporación de funciones de memoria de diferentes lapsos de tiempo, para implantar estrategias que refuerzen la combinación de movimientos y las características de solución que históricamente se han encontrado como buenas, mientras que las estrategias de diversificación manejan la búsqueda dentro de nuevas regiones.

Las estructuras de memoria de la Búsqueda Tabú operan bajo cuatro dimensiones principales: pertenencia, frecuencia, calidad e influencia. El papel de estos elementos en la creación de procesos efectivos para la resolución de problemas son uno de los focos de atención de este trabajo.

3 Desarrollo de la técnica

Inicialmente se realizó una clasificación aleatoria de los elementos en cierto número de clases y las vecindades eran generadas de manera aleatoria respecto de esta primera bajo la consideración de tomar como elemento tabú la clasificación obtenida, posteriormente se utilizó el criterio de aspiración clásico, llevándose un registro histórico del proceso, considerando los puntos principales en que se basa la Búsqueda Tabú y que se describen en la sección anterior. Al final del proceso del método propuesto (que denotaremos en adelante como SC) y respecto de las estructuras de memoria, se utilizó la información de, en cuál clase original se encontraba cada elemento como una forma de medir la eficiencia de SC. Para tal propósito se utilizaron los conjuntos de datos clasificados que se utilizaron en Goddard et al. (1998) y obtenidos de Fisher (1938), Peterson y Barney (1952) y Yan (1993), denotados como: IRIS, PETU y YAN respectivamente, en particular se utilizaron parcialmente los datos dados en Yan (1993), para poder observar el comportamiento del SC. Se propusieron como los elementos característicos o prototipos o "agujeros negros de clase", aquellos que proporcionaban las mejores características respecto de las estructuras de memoria desarrolladas. Consideramos que el proceso para formular la representación inicial o para mejorar la representación dada, aún no es bien comprendida. Pareciera que el hallazgo de cambios deseables en la representación de un problema, depende de la experiencia acumulada en los intentos por resolverlo, por lo que consideramos que esta experiencia nos puede permitir descubrir la existencia de ciertos elementos o reglas simplificadoras.

4 Resultados

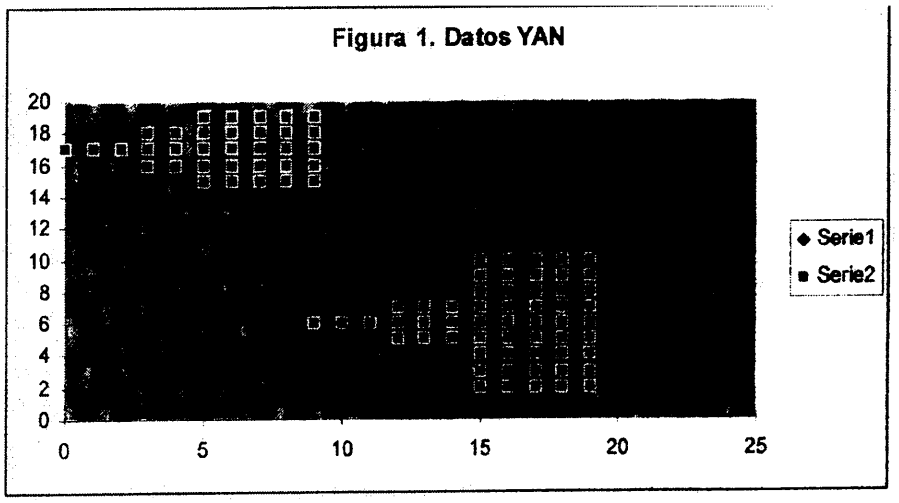
Se realizaron varias corridas sobre los tres conjuntos de datos con SC. La tabla 1 muestra algunos de los resultados obtenidos.

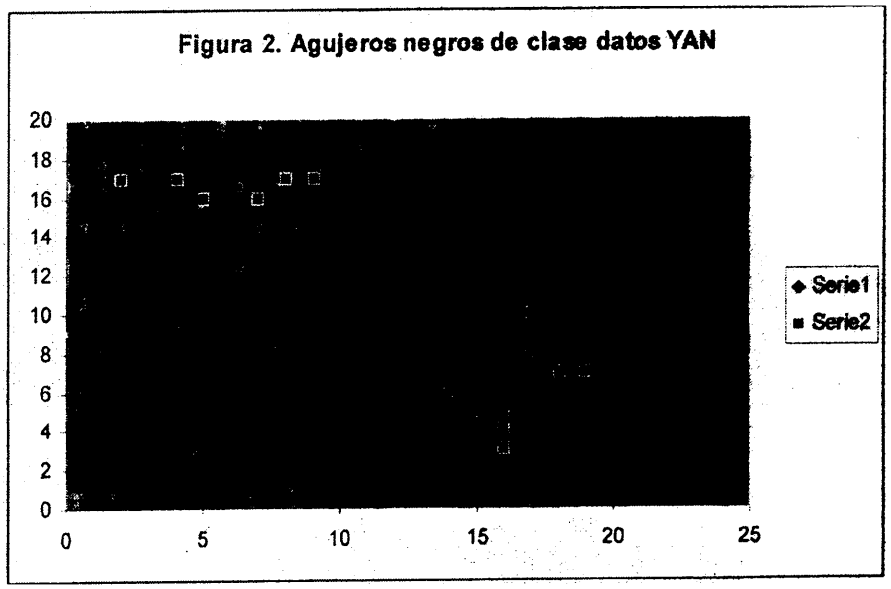
Tabla 1. Clasificación de los datos utilizando SC.

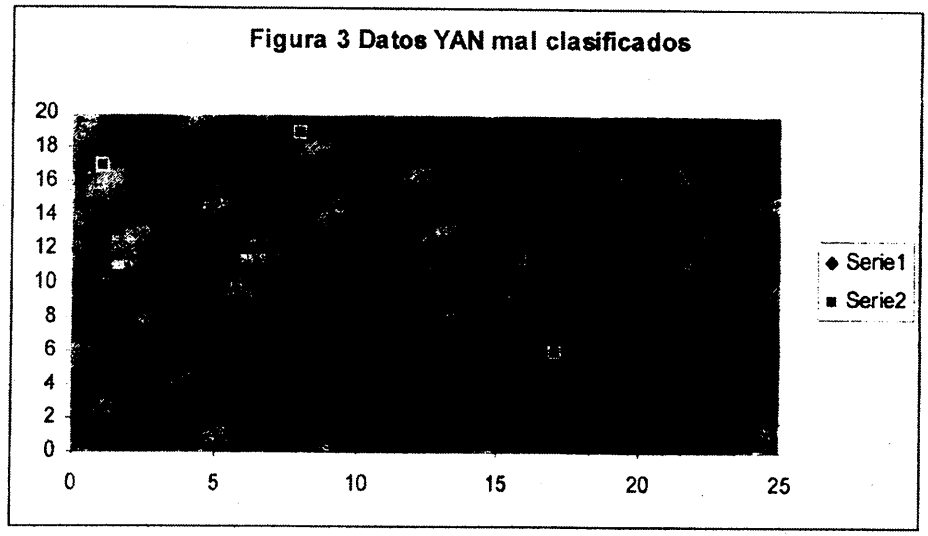
Conjunto	No. de datos	Clases Originales	Clases con SC	Eficiencia SC %
YAN	162	2	2	96.29
IRIS	150	3	3	84.66
PETU	184	2	2	96.19

Cabe mencionar que la cantidad de "parámetros o características" medidas para cada dato de los conjuntos YAN, IRIS y PETU eran 2, 4 y 4 respectivamente. Los datos mostra-

dos en la tabla 1, son los mejores que se obtuvieron después de realizar (en tres ocaciones) 2, 6 y 2 corridas respectivamente de los conjuntos de datos, realizando modificaciones en los diferentes parámetros descritos en la sección anterior. En este punto es importante mencionar que la información obtenida de los parámetros para el caso de los datos de IRIS no se utilizaron para el resto de los conjuntos de datos probados. Es interesante mencionar que en algunas de las corridas de IRIS, se encontraron una o dos clases más, pero éstas contenían a lo más dos elementos, por lo que consideramos que no eran representativas. (no es el caso de los resultados presentados). Las gráficas de diferentes conjuntos de datos se proporcionan a continuación.







5 Conclusiones

Es interesante observar que SC proporciona la cantidad exacta de clases para los conjuntos de datos utilizados por lo que se podría pensar que se trata de un método eficiente para cualquier conjunto de datos a clasificar. Puesto que el método no utiliza información privilegiada, al parecer la eficiencia de SC depende del ajuste de sus parámetros, por lo que en esta dirección se tiene que realizar más investigación. Después de observar los resultados, se aprecia que SC proporciona resultados promisorios para atacar los problemas de clasificación (automática) y de reconocimiento de patrones de manera general, así como el de encontrar los patrones (agujeros negros de clase) para optimizar otros procesos como sería entre otros, el de utilizar conjuntos mínimos de entrenamiento en redes neuronales. Finalmente si se considera que la metodología propuesta proporciona particionamientos robustos, la cantidad de elementos que deben de considerarse en los agujeros negros de clase para una colección de datos en particular, es otra línea de investigación.

Referencias

De los Cobos Silva S.G., Goddard Close J., Gutiérrez Andrade M.A. y Pérez Salvador B.R. (1997). Redes Neuronales Probabilísticas: Perspectivas en clasificación y reconocimiento de patrones. *XII Foro Nacional de Estadística*, IIMAS-UNAM.

Fisher R.A. (1936). The use of multiple measurements in taxonomics problems. *Annual Eugenics*, 7, Part III, pp. 179-188.

Glover F. (1989). Tabu Search, Part I. ORSA *Journal on Computing*, 1:3, pp. 190-206.

Glover F. (1990). Tabu Search, Part II. ORSA *Journal on Computing*, 2:1, pp. 4-31.

Goddard J., Martínez A.E. & Martínez F.M. (1998). Prototype selection for Nearest Neighbour classification. *Congreso Latinoamericano de Ingeniería Biomédica*, Mazatlán, México.

Murillo, A, Trejos J. (1996). Classification Tabou baseé en transferts. *IV Journées de la Societe Francophone de Classification*, S. Joly & G. Le Calvé(eds.), Vannes: 26.1-37.4

Peterson y Barney (1952). Control methods used in a study of the vowels. *Journal of the American Statistical Association* 24, pp.175-184

Piza E. y Trejos J. (1996). Partitionnement par recuit simulé. *IV Journées de la Societe Francophone de Classification*, S. Joly & G. Le Calvé(eds.), Vannes: 27.1-27.4

Trejos J. (1996). Un algorithme génétique de partitionnement. *IV Journées de la Societe Francophone de Classification*, S. Joly & G. Le Calvé(eds.), Vannes: 37.1-37.1

Yan H. (1993). Prototype optimization for neighbor classifiers using a two-layer percepton. *Pattern Recognition*, vol. 26, pp. 317-323.

Un Modelo de Transferencia de Conocimientos Vía Internet

José Luis García Cué y José Antonio Santizo Rincón

ISEI, Colegio de Postgraduados

1 Introducción

En 1997 se constituyó el primer modelo de educación a distancia en el Colegio de Postgraduados en el que se elaboraron textos didácticos, páginas WEB, programas tutoriales en Visual Basic 4 y se instalaron servidores WEB y FTP para los cursos de Introducción a la Estadística e Introducción a los Diseños Experimentales, se utilizó el servidor de correo electrónico (E-mail) del CP para asesorías. En septiembre del mismo año se probó el modelo en el curso de Introducción a los Diseños Experimentales, de acuerdo a los resultados obtenidos y a la investigación educativa realizada, se propusieron, cambios para una mayor interacción entre el contenido de los textos y el usuario en las páginas WEB. El presente trabajo plantea cambios en el modelo al través de nuevas herramientas computacionales, lenguajes de programación e inovación en técnicas que permite la transferencia adecuada de conocimientos a distancia en el ámbito de las ciencias agrícolas.

2 Educación a distancia

La educación a distancia es una relación formal de enseñanza, mediada por la tecnología, que rompe con las barreras del tiempo y el espacio para la transmisión de conocimientos.

La noción de educación a distancia es un concepto en constante evolución. Holmberg (1986) la define como la separación física del estudiante y el maestro en el tiempo y el espacio. Por otra parte, Lauzon y Moore (1991) la describen a partir de las relaciones entre la tecnología y los medios es decir, canales de comunicación (impresos, audios y videos), medios de transmisión (correo, computadoras) y modo de transmisión (sincrónica y asincrónica)

3 La comunicación mediada por computadora (CMC)

El término Comunicación Mediada por Computadora (CMC) se refiere a la que tienen dos personas, cada una al través de una terminal computacional, en forma sincrónica o asincrónica. Well(1992) recomienda, cuando las computadoras se interconectan con propósitos pedagógicos, usar el término "Comunicación mediada por computadora CMC" porque potencialmente implica el empleo de una variedad de servicios en redes vía Internet que incluyen: correo electrónico(E-mail), bancos de datos, bancos de imágenes, bibliotecas virtuales, clubes de usuarios por especialidad en tiempo real y asincrónico, periódicos en línea, grupos de noticias, revistas, páginas con información y boletines electrónicos (home page o html), foros de discusión, audio, video, software público y animaciones.

El origen del uso de la CMC vía Internet como medio complementario para la educación a distancia se basó en sistemas tradicionales de video-conferencias, figura 1. Estos utilizaron servicios disponibles como correo electrónico (E-mail), servidores de noticias (Usenet) y grupos de conversación (IRC) como apoyo en sus cursos.

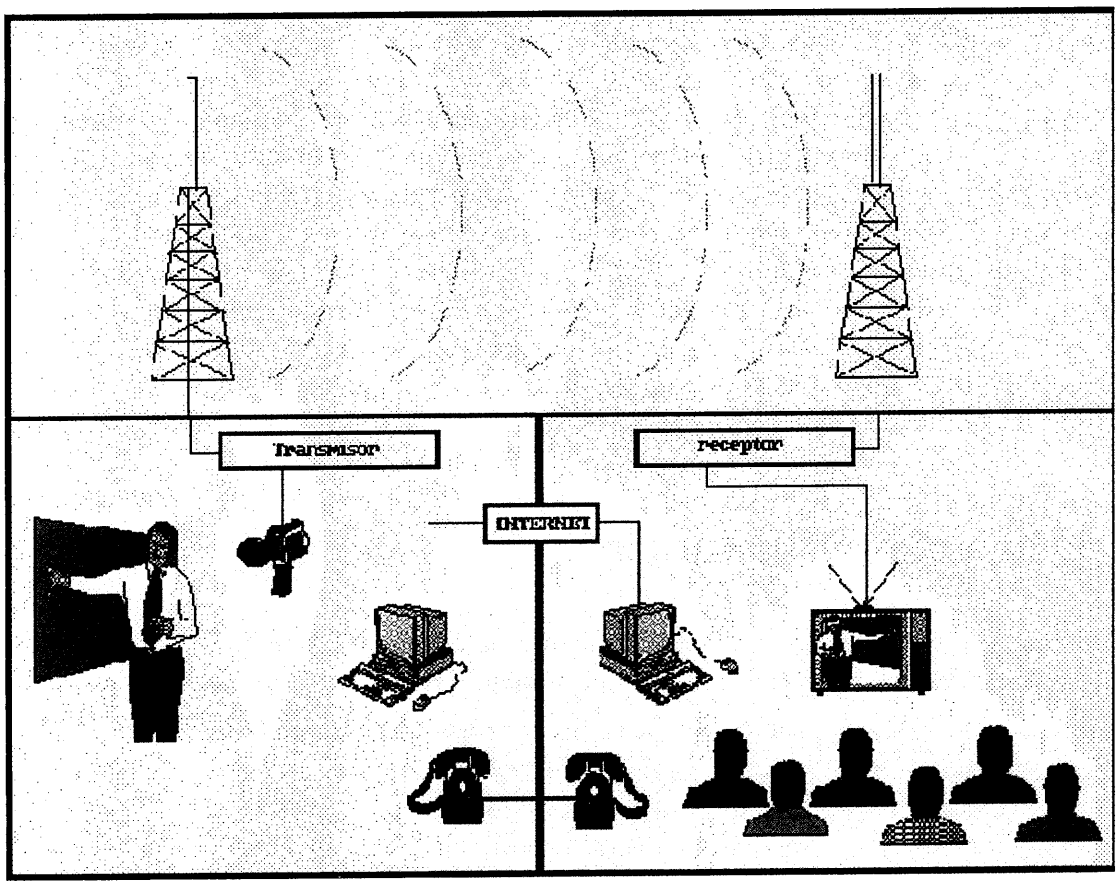


Figura 1: Video conferencias

Diversas instituciones académicas han propuesto alternativas para la transmisión de conocimientos a distancia diferentes a las video-conferencias por los altos costos de inversión de éstas y han sugerido la CMC vía Internet. Se han propuesto el uso de programas tutoriales, el correo electrónico (E-mail), conferencias en Internet (IRC), documentos escritos, video, audiográficas, páginas electrónicas apoyadas de servicios telefónicos y facsímil en todas sus combinaciones, figura 2.

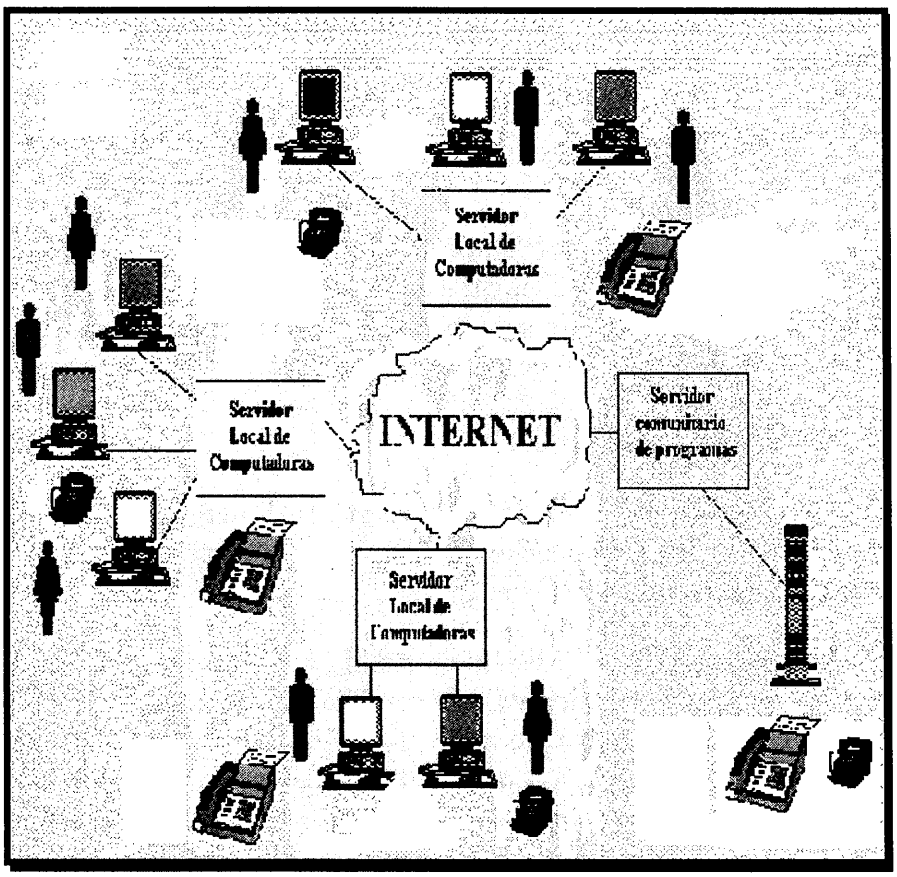


Figura 2: Comunicación Mediada por Computadora (CMC)

4 Propuesta del modelo

En esta parte se consideran los elementos del modelo de transferencia de conocimientos vía Internet.

4.1 Diseño instruccional

El modelo que se propone está orientado a impartir cursos vía Internet mediante las herramientas de Correo Electrónico (E-mail), páginas electrónicas Web y programas en Visual Basic.

4.2 Uso múltiple de tecnologías y medios de transmisión

Las tecnologías y medios de transmisión se construyen en función de los recursos físicos y tecnológicos disponibles; para el efecto en este trabajo se identifica el equipo disponible a nivel institucional y el software que se puede utilizar.

Identificación de equipo disponible.- Los equipos computacionales disponibles para este proyecto son:

- Computadora marca DELL, CPU 486 DX50, 8 Mbytes en RAM, HD de 1.0 Gbytes IDE, disco flexible de 3.5 ", memoria Cache de 256 Kbytes, monitor SVGA .28, VRAM 512 Kbytes, tarjeta de sonido Sound Blaster Pro 16 bits, CDROM marca Sanyo IDE, tarjeta de red Cabletrón E2100 con conexión par trenzado, acceso Internet con la dirección (IP) 200.23.24.32 y conexión al servidor de correo electrónico 192.100.178.1 ó colpos.colpos.mx

- Computadora Printaform, CPU 486 DX2 a 66, 16 Mbytes de RAM HD 1.2 G IDE, disco flexible 3.5", memoria cache de 512 Kbytes, monitor SVGA .28, VRAM 1 Mbyte, Tarjeta de Sonido GoldStar 16 bits, CDROM GoldStar 6000.

- Computadora marca SUN, 64 Mbytes en RAM, disco duro de 1.0 Gbytes SCSI, disco flexible de 3.5 pulgadas, memoria Cache de 256 Kbytes, monitor SVGA .28, VRAM 512 Kbytes, CDROM, tarjeta de red con conexión par trenzado, acceso Internet con la dirección (IP) 200.23.24.5 que se utiliza como servidor Web del Instituto de Socioeconomía Estadística e Informática (ISEI) del Colegio de Postgraduados.

Identificación del Software.- El software disponible para este proyecto es: Sistemas operativos MS Windows 95, MS Windows NT para PC y UNIX para SUN, lenguaje de programación Visual Basic 5.0, MS Internet Explorer V3.01 y 4.0 y Netscape Comunicator; Unix Web Server y ZBS Web Server para Windows 95 y NT.

4.3 Población a la que va dirigido el programa educativo

La población a la que va dirigida este programa educativo está conformada por estudiantes de las diferentes especialidades de maestría y doctorado del CP tanto en la sede como en los campus.

4.4 Elaboración de materiales de enseñanza

En la elaboración de materiales de enseñanza se trató de ir en concordancia a la enseñanza de cursos en la modalidad presencial y se aprovecharon los beneficios que ofrece el empleo de la tecnología computacional.

Documentos

Para la elaboración de documentos y su uso en cursos de educación a distancia se propone para el manejo del modelo, definir los puntos didácticos siguientes:

- Objetivo general del curso.
- Pre-test y post-test.

Para cada uno de los temas: Objetivos generales y específicos de cada tema; información para cubrir cada objetivo específico y ejercicios.

Elaboración de programas de cómputo para cursos a ser transmitidos vía Internet. Después de seleccionar y analizar los lenguajes de programación que van a ser utilizados, se hace un diagrama general de los módulos de instrucciones de computadora que incluyen cada uno de los programas. En esta parte trabajan en conjunto el experto del curso y el programador lo que evita confusiones.

Para este trabajo se propusieron tres diferentes lenguajes de programación:

- Visual Basic.- Para la elaboración de programas interactivos que puedan ser transmitidos vía WEB al través de Microsoft Explorer.
- HTML.- Se seleccionó para la construcción de páginas Web con información a ser consultada vía Internet.
- Java Script.- Se seleccionó para construcción de páginas Web interactivas y transmitidas vía Internet mediante Netscape e Internet Explorer.

a) Programas en Visual Basic

La estructura general de un programa, se muestra en la figura 3.

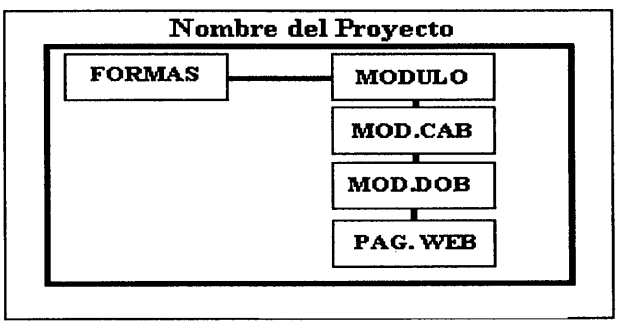


Figura 3: Estructura de los programas elaborados en Visual Basic

b) Programas de páginas electrónicas Web con HTML y Java Script

El desarrollo de los programas se basan en una estructura propuesta de acuerdo a la figura 4. El formato contiene: una página Web con dos diferentes marcos uno para el menú de opciones y otro que muestre los resultados según la selección. El menú presenta diferentes partes:

- Página Web con la teoría.
- Página Web con los ejercicios
- Página Web con un programa Interactivo para trabajar con los ejercicios.
- Página Web con instrucciones sobre el programa.
- Página Web que tiene el test del tema.

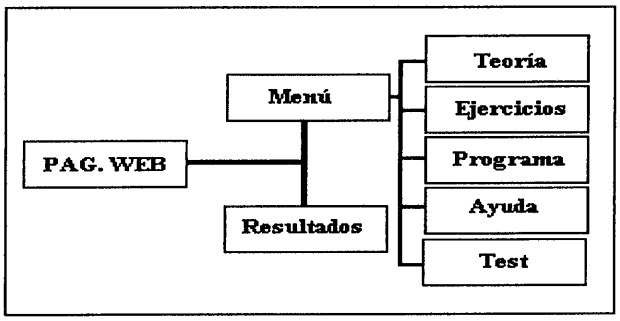


Figura 4: Estructura de los textos para educación a distancia

4.5 Interacciones establecidas en el modelo

Interacción alumno-contenido.- El alumno inscrito en el curso tiene que cumplir su programa de estudios en un tiempo establecido. El alumno puede obtener información del servidor del ISEI al través de los navegadores Microsoft Internet Explorer y Netscape.

Interacción alumno-profesor.- En forma asincrónica: En esta parte el alumno comunica sus dudas al través del correo electrónico y el profesor le contesta por este mismo medio y escribe las respuestas a las dudas más comunes dentro de una página Web. En forma sincrónica: El profesor propone un horario de atención para los alumnos y estos se comunican al través del Netscape Comunicator en conexión de uno a uno.

Interacción alumno-alumno .- Esta interacción se hace en forma asincrónica mediante el uso el correo electrónico (E-mail).

Interacción entre los sitios enlazados.- La interacción entre los sitios enlazados se hace bajo el modelo de red cliente-servidor vía Internet para los servicios de correo electrónico (E-mail), FTP y páginas Web.

5 Resultados

5.1 Instalación del software para Windows 95

El software instalado que resultó del análisis y prueba fue: Windows 95, Windows NT, UNIX, Microsoft Visual Basic 5.0, Microsoft Office 97, Netscape, Microsoft Explorer, ZBS Server para Windows 95 y Windows NT, y el UNIX Web Server para SUN.

5.2 Programas elaborados

- Programas en Visual Basic

  Los programas creados bajo el lenguaje Visual Basic fueron: 3 proyectos llamados Distfrec, Graftabz y Disbino, cada uno con sus correspondientes páginas Web y formas.

- Páginas Web y programas en Java Script.

  Se construyeron 10 páginas con instrucciones en lenguaje HTML y Java Script para hacer interactivas las páginas del tema de Función de Probabilidad. Dentro de estas páginas destacan Tablafx y Correla que calculan la E(X), E(X2), Var(X), la Correlación y la Covarianza.

6 Conclusiones

- Se investigaron nuevas herramientas basadas en los equipos de cómputo con que se disponían y se planteó la utilización de una versión modificada y superior de Visual Basic, se incluyó Java Script y lenguaje HTML.

- La propuesta del modelo de transferencia de conocimientos vía Internet obligó a profundizar los aspectos relacionados con didáctica, educación a distancia y su aplicación.

- La propuesta del modelo llevó a identificar las herramientas como correo electrónicos y páginas Web Interactivas.

- Internet Explorer 3.01 y 4.0 no leen correctamente las instrucciones de las páginas realizadas con Java Script.

- El navegador Netscape Gold 3.0 no lee las páginas Web elaboradas en Visual Basic.

Referencias

Brown, Kenion (1992). Introducción a la programación de Visual Basic Grupo Noriega Editores, México. p.504

García Cué., JL (1997). Un modelo de educación a distancia. Tesis de Maestría en Ciencias. Colegio de Postgraduados, Montecillo, México. PP 105

Holmberg, B. (1986). Growth and Structure of Distance Education. Londres, Croom Helm.

Lauzon A., G. Moore, (1991). "A Fourth Generation of Distance Education System: Integrating Computer Assisted Learning and Computer Conferencing", En The American Journal of Distance Education 3(1), pp28-39.

Microsoft (1995). Visual Basic Language Reference. Microsoft Corporation. p.1064

Moore, M. (1991). "Distance Education". En International Encyclopedia of Higher Education.

Well, R. (1992). Computer Mediated Communication for Distance Education: an International Review of Design, Teaching and Institutional Issues. (ASCDE Research Monographs, 6) State College PA, The Pennsylvania State University.

Sobre la Estimación de Densidades por Funciones Ortogonales

John Goddard
UAM-Iztapalapa, México

Sergio De los Cobos
UAM-Iztapalapa, México

Miguel Angel Gutiérrez
UAM-Azcapotzalco, México

Blanca Rosa Pérez
UAM-Iztapalapa, México

1 Introducción

Un problema que surge en la estadística no paramétrica es, dada una muestra $X_1, X_2, ..., X_N$, de una densidad de probabilidad desconocida f, encontrar un estimador para f sin suponer alguna estructura funcional.

Históricamente se han utilizado distintos métodos incluyendo: histogramas (Scott, 1992), series ortogonales (Cencov, 1962), núcleos (Parzen, 1962), secuencias de funciones deltas (Walter y Blum, 1979), y más recientemente ondillas (*wavelets* en inglés) (Vannucci, 1995).

La estimación no paramétrica es interesante porque generalmente permite la estimación de cualquier densidad, a diferencia de la estimación paramétrica que supone una forma particular de la distribución, y trata de estimar los parámetros involucrados.

Para una estimación parámetrica, no importa el número de datos disponibles si la forma de densidad escogida está equivocada, nunca va a ser la correcta. En (Tarter y Lock, 1993) se refiere a la completes de la representación para explicar este fenómeno. Desde esta perspectiva, la aproximación de densidades por medio de series ortogonales, también llamada estimadores de proyección, es particularmente atractiva.

Aplicando los resultados de los espacios de Hilbert, cualquier función en $L^2(D)$ donde $D \subseteq \Re$ puede ser representada por una serie que converge en la norma L^2, cuyos términos son de una base de funciones ortogonales.

En caso de que $D = [0,1]$ podríamos tomar senos y cosenos, pero también existen otras bases ortogonales como los polinomios de Legendre (Rees et al., 1981). De hecho, en los últimos años se han desarrollado distintas familias de ondillas que también forman bases ortogonales para diferentes subconjuntos D y que podrían ofrecer ventajas para este problema (Vannucci, 1995).

En este artículo introducimos una red neuronal que implementa un estimador no sesgado para el caso de bases ortogonales. La llamamos red neuronal por series ortogonales (RNSO).

La implementación es análoga a aquellas propuestas por Specht (1996), en la red neuronal probabilística (PNN del inglés *probabilistic neural network*) y en la red neuronal de regresión general (GRNN del inglés *general regression neural network*), que implementan un clasificador Bayesiano y una superficie de regresión general, respectivamente.

De hecho la RNSO, PNN y GRNN son "métodos basados en memoria", debido a que los datos se utilizan durante el funcionamiento del modelo. Las redes neuronales que utilizan retropropagación (Hassoun, 1995), y la red funcional de Pao, (1989) son "métodos basados en modelos", debido a que los datos se utilizan para fijar los pesos obtenidos en el entrenamiento y usualmente eliminados para el funcionamiento del modelo.

Otra diferencia importante entre RNSO y los últimos métodos mencionados es que RNSO no requiere de los valores de la densidad sino de una muestra de ella, mientras que los otros funcionan bajo una esquema de "aprendizaje supervisado" en el cual, los valores correspondientes de la función a aproximar están dados por el entrenamiento.

Para abordar esta necesidad utilizamos los valores de un histograma para realizar el entrenamiento de la red funcional. Cabe señalar que la red de Pao utiliza un entrenamiento sin capas escondidas, lo cual la hace más atractiva que retropropagación por su tiempo reducido de entrenamiento.

El trabajo está organizado de la siguiente manera: en la sección 2 revisamos la estimación no paramétrica por medio de funciones ortogonales. Posteriormente, en la sección 3, introducimos la implementación del estimador como red neuronal. La sección 4 retoma la red funcional creada por Pao, y en la sección 5 aplicamos una combinación de los métodos a un ejemplo concreto. Finalmente, presentamos algunas observaciones y conclusiones.

2 La estimación no paramétrica por medio de funciones ortogonales

La estimación no paramétrica por medio de funciones ortogonales fue introducida por Cencov (1962).

Se encuentran buenas introducciones a este tema en (Tarter y Lock, 1993) y (Silverman, 1986). Hace uso del hecho de que en $L^2(D)$, donde $D \subseteq \Re$, podemos aproximar cualquier función por medio de una base ortogonal.

Para el caso de $D = [0,1]$, sea $\{g_i\}$ una base de funciones ortonormales con respeto al producto escalar:

$$< f, h >= \int_D f(x)h(x)dx,$$

donde $f, h \in L^2(D)$. La suma $\sum c_j g_j(x)$ con $c_j =< f, g_j >$ converge puntualmente a f, si f es continua.

La forma de un estimador no sesgado para la densidad f de $L^2(D)$ es:

$$f_{K^*}(x) = \sum_{j=0}^{K} d_j g_j(x),$$

donde los coeficientes d_j están dados para una muestra de tamaño N por:

$$d_j = \frac{1}{N} \sum_{i=1}^{N} g_j(x_i).$$

Como se explica en (Tarter y Lock, 1993), se podría extender este análisis e incluir una función de peso. El presente trabajo utiliza la base de senos y cosenos, pero se podrían extender nuestros resultados al caso más general.

Un problema que surge es cómo determinar el número de términos K para el estimador. Para el caso de la base de cosenos y senos, (Tarter y Lock, 1993) sugiere la inclusión de todos los términos hasta un cierto número t que han fallado para satisfacer la siguiente desigualdad:

$$|d_j|^2 > \frac{2}{N+1}.$$

Lo más común es tomar t igual a 1 ó 2.

3 La RNSO

La RNSO es esencialmente una red neuronal diseñada para reproducir la función $f_{K^*}(x)$. La arquitectura de la red se muestra en la figura 1.

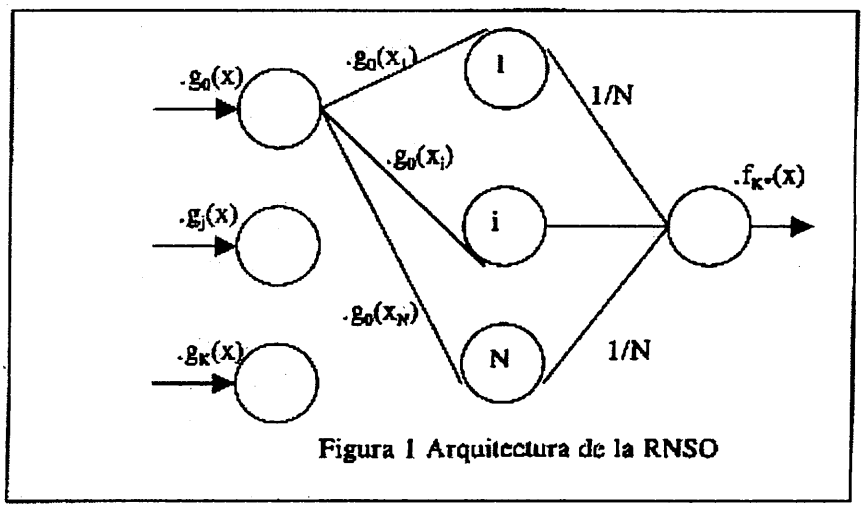


Figura 1 Arquitectura de la RNSO

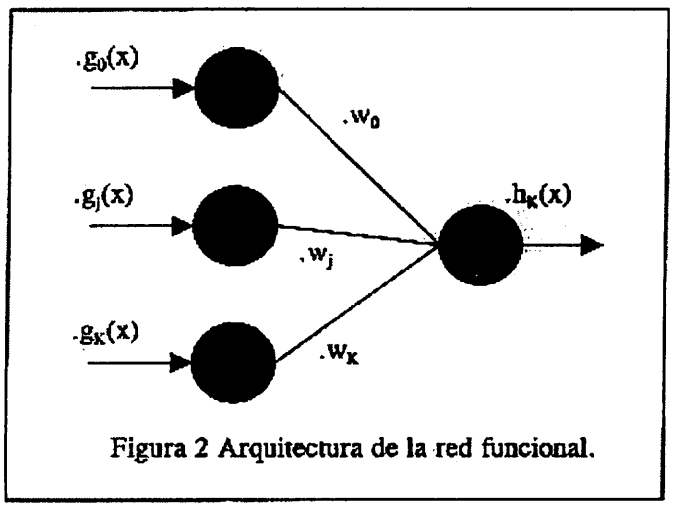


Figura 2 Arquitectura de la red funcional.

La red consiste de 3 capas: la capa de entrada, la de patrones, y la de salida. La capa de entrada tiene el mismo número de unidades que los términos en el estimador $f_{K^*}(x)$.

La capa de patrones tiene el mismo número de unidades que el tamaño de la muestra, o sea N. Finalmente, la capa de salida tiene una unidad.

Los pesos entre la capa de entrada y la capa de patrones están asignados por los valores $g_j(x_i)$. Es decir, si enumeramos las unidades en la capa de patrones, entonces para la conexión de la unidad de entrada j a la unidad patrón i asignamos $g_j(x_i)$.

Para los pesos entre la capa de patrones y la salida ponemos el valor fijo de $1/N$.

La salida de la i-ésima unidad en la capa de patrones es:

$$\sum_{j=0}^{K} g_j(x) g_j(x_i).$$

La salida final es:

$$\frac{1}{N} \sum_{i=0}^{N} \sum_{j=0}^{K} g_j(x) g_j(x_i).$$

Reordenando los términos se tiene el estimador f_{K^*}.

4 La red funcional de Pao

En Pao (1989) describe una red que la llama red funcional. En este trabajo nos interesa una versión de red que aproxime una función y puede ser definida por medio de la figura. La red consiste de dos capas, de entrada y de salida.

La salida actual de la red de la figura 2 está dada por:

$$h_K(x) = \sum_{j=0}^{K} w_j g_j(x).$$

En el caso actual, $\{g_j\}$ van a ser cosenos y senos.

La red esta entrenada con la regla delta, que minimiza el error medio cuadrado entre el valor actual de la salida y el valor deseado de la función.

El propósito del entrenamiento es el encontrar un conjunto de pesos w_j que realizan la aproximación deseada. Este tipo de entrenamiento es llamado "aprendizaje supervisado" y requiere de los valores de la función que se quiere aproximar.

En nuestro caso no se tiene estos valores por lo que se utilizan los valores de las marcas de clase del histograma generado por la muestra. Otro problema que surge es el de escoger el número de términos K para la aproximación. El criterio que utilizamos es el mismo que para RNSO.

5 Aplicación a una muestra

Hemos tomado una muestra de 500 puntos de una función gaussiana con media 0.5 y desviación estándar 0.16.

El número de términos fue determinado por el criterio mencionado en la sección 2 con $t = 2$ (en este caso el mismo número de términos fue escogido también con $t = 1$).

Para utilizar la red funcional de Pao, se requiere de valores deseados de la función y también fijar el número de entradas.

Hemos tomado los valores de la marca de clase de un histograma para la misma muestra de 500 puntos con 22 subdivisiones.

6 Observaciones y conclusiones

En este articulo, se ha hecho varios cosas. Primero, se ha definido una red, RNSO, que implementa un estimador no sesgado para el caso de bases ortogonales.

Posteriormente se ha utilizado RNSO para fijar la arquitectura de la red funcional de Pao y entrena a esta con los valores de la marca de clase de un histograma de la muestra original.

La razón ha sido de utilizar RNSO para fijar el modelo y la red funcional para ajustar los parámetros del mismo.

Finalmente se ha ilustrado este esquema con un ejemplo sencillo. Es conveniente observar que tanto RNSO como la red funcional son aproximadores universales, en el sentido de que, para las dos redes existen configuraciones de ellas que pueden aproximar cualquier función de $L^2(D)$ en la norma L^2. Sería interesante considerar otras familias de funciones ortogonales, como las ondillas, para realizar esta tarea de aproximación.

Referencias

Cencov, N.N. (1962). Evaluation of an unknown distribution density from observations. *Doklady*, 3, 1559-1562.

Hassoun, M.H. (1995). *Fundamentals of Artificial Neural Networks.* The MIT Press.

Pao, Y. (1989). *Adaptive Pattern Recognition and Neural Networks.* Addison–Wesley.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, 33, 1065-1076.

Rees C.S., Shah S.M., Stanojevic C.V., (1981). *Theory and Application of Fourier Analysis.* Dekker.

Scott, D.W. (1992). *Multivariate Density Estimation.* Wiley.

Silverman, B.W. (1986). *Density Estimation.* Chapman & Hall.

Specht, (1996). Probabilistic Neural Networks and General Regression Neural Networks. *Fuzzy Logic and Neural Network Handbook.* En Chen (Ed.) McGraw–Hill.

Tarter, M.E. y Lock, M.D. (1993). *Model-Free Curve Estimation.* Chapman & Hall.

Vannucci, M. (1995). *Nonparametric Density Estimation using Wavelets.*

Walter, G.G. y Blum, J. (1979). Probability Density Estimation using Delta Sequences. *The Annals of Statistics*, 7, n.2, 328-340.

Análisis Estadístico de un Estudio de Rendimiento, Egreso y Deserción Escolar

Arturo González Izquierdo y Belem Trejo Valdivia
CIMAT, Aguascalientes

1 Introducción

En enero de 1998, la representación de la SEP en el Estado de Coahuila llevó a cabo un estudio sobre el desempeño escolar a nivel licenciatura. Para ello se evaluó el desarrollo de estudiantes de un Instituto Tecnológico[1]a lo largo de su carrera. El objetivo principal era el de modelar el patrón de comportamiento del desempeño escolar de dichos estudiantes junto con el de evaluar el efecto que pudiera tener el tipo de escuela de procedencia así como el efecto de características individuales de dichos estudiantes. Como objetivos específicos se identificaron entre otros los siguientes:

–Describir el rendimiento de los estudiantes, con respecto a características relevantes de la escuela de procedencia.

–Identificar cuales características de la escuela de procedencia y cuales características de los individuos influyen, y en qué medida, en la permanencia de los estudiantes.

No fue posible elaborar un diseño de muestra para estudiar el fenómeno mencionado dado que la oficina de Control Escolar de tal instituto, sólo proporcionó la información correspondiente a las generaciones que ingresaron en 1992 y su seguimiento hasta el segundo semestre de 1997. A pesar de esta restricción se espera obtener una adecuada descripción del desempeño escolar debido a que no se tienen elementos para pensar que tales generaciones sean distintas a las demás.

Una característica de especial consideración fue la duración del programa de estudios de la escuela de procedencia, de las cuales es posible identificar escuelas de dos y tres años. Otra característica igualmente importante fue el tipo de sostenimiento, que divide a las escuelas en públicas y privadas. Además se incluyó en el análisis el tipo de autorización que obtuvo la escuela preparatoria para su creación. Como el tipo de autorización depende del organismo que la expide, quedan identificadas escuelas autorizadas por la SEP y escuelas autorizadas por un organismo distinto.

[1]Por razones de confidencialidad no se mencionará el nombre de la institución educativa bajo estudio.

Figura 1: Esquema de los tiempos del estudio

Por otro lado, se tienen características de los estudiantes tales como el sexo, calificaciones obtenidas en materias del área de ciencias básicas (matemáticas finitas y cálculo, dibujo, química, física, etc.); así como el tipo de examen en el que el individuo acreditó la materia (es decir, ordinario, extraordinario, etc.). También se tienen datos acerca del número de semestres que el alumno permaneció inscrito. Esta característica se midió como un conteo de semestres hasta el último en que el alumno se registró y obtuvo calificaciones en las materias correspondientes. Por último se cuenta con el estatus escolar de los individuos, el cual nos indica si el este se dio de baja definitiva, baja temporal, si es egresado o aún permanecía vigente[2] hasta el segundo semestre de diciembre de 1997.

Debido a que existen dos períodos de ingreso, enero y agosto, la generación que ingresó en agosto de 1992 no tiene un doceavo semestre registrado en la base de datos. Así, los períodos de los cuales se tiene información se pueden ilustrar de la siguiente manera.

De esta base de datos se excluyeron a los estudiantes que debido a la ubicación de la escuela de procedencia, no pertenecían al área de influencia del tecnológico. De esta forma, la base de datos quedó estructurada con 577 casos.

El análisis se enfocó a modelar tres fenómenos educativos de importancia relativos al desempeño escolar: el egreso, la deserción y el rendimiento. Este último se refiere al grado de aprovechamiento del alumno en el proceso de enseñanza- aprendizaje y es medido a través de las calificaciones asignadas por los profesores al finalizar las materias. En este caso sólo se analizan las calificaciones de las materias del área de ciencias básicas. Por egresado tendremos a aquel alumno que ha cursado y aprobado todas las materias del plan de estudios. La deserción quedó establecida por aquellos alumnos que no terminaron la carrera y que ya no eran susceptibles de reingresar. Esta se midió por medio del estatus escolar, el cual nos indicaba que si un estudiante era dado de baja definitiva, podía considerarse como desertor. En este trabajo se presentan resultados que indican cuáles características

[2]El término vigente es utilizado para identificar a aquellos alumnos que aún están inscritos en alguna carrera.

de la escuela de procedencia y del individuo influyen en su desempeño como estudiantes.

2 Metodología

Para poder analizar el rendimiento de los estudiantes se decidió tratar de detectar agrupaciones naturales respecto a las calificaciones de las materias de ciencias básicas y posteriormente realizar comparaciones entre tales grupos. Puesto que no se contó con un criterio específico de agrupación y al enfrentarnos a la gran diversidad de materias por alumno, se decidió utilizar el Análisis de Conglomerados (k- medias). Además, con esta metodología obtendríamos grupos mutuamente excluyentes y valores que nos permitieran evaluar si tales agrupaciones son suficientemente diferentes como para considerarlas como una clasificación de rendimiento.

Con relación al comportamiento de la permanencia de los estudiantes, se utilizó un modelo dentro del llamado análisis de datos de supervivencia. Este tipo de modelos resultan adecuados en situaciones en las que es necesario estudiar el comportamiento de un fenómeno que involucra eventos de interés a través del tiempo. En este caso, el número de semestres que el individuo permaneció en el Instituto como estudiante activo representa un tiempo de supervivencia de tipo discreto. Sin embargo, aquí pueden medirse dos eventos de interés: que el alumno deserte o que egrese. Esto nos lleva a estudiar por separado egreso y deserción, a través de los cuales podremos tener una idea general de la permanencia.

La deserción quedó establecida por aquellos alumnos dados de baja definitiva. Para el análisis de éste fenómeno se consideró como variable respuesta al número de semestres que el alumno permaneció inscrito. Como covariables se incluyeron las características asociadas a la escuela de procedencia y el sexo del individuo. La censura quedó definida por aquellos alumnos con un estatus distinto, es decir, estudiantes dados de baja temporal o vigentes. Aquí fue posible aplicar un modelo de riesgos proporcionales de Cox.

Al identificar que los egresados ya no son susceptibles de desertar, se dividió la base de datos para estudiar por separado a éstos del resto de estudiantes. En ellos hubo que analizar el comportamiento del egreso a través de los semestres optativos 8, 9, 10, 11, 12; por lo que se tienen como variable dependiente al número de semestres que el estudiante tardó en egresar. Se consideraron las mismas covariables que en el estudio de la deserción. Como los egresados no presentan ningún tipo de censura fue posible aplicar el modelo de riesgos proporcionales de Cox y un modelo log- lineal. Ambos nos permitieron evaluar la bondad de los resultados.

3 Resultados

El análisis de conglomerados con las calificaciones semestrales del área de ciencias básicas, sugirió la formación de tres grupos, los cuales pueden ser referidos como el formado por

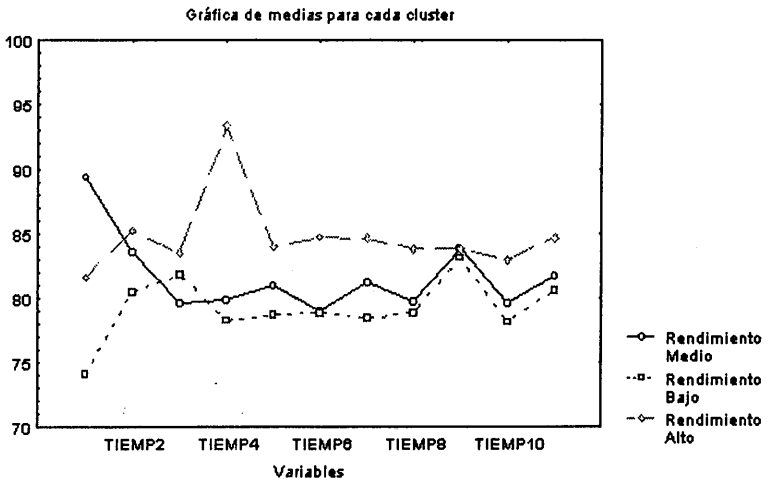


Figura 2: Comportamiento de las medias de los grupos de rendimientos, según el agrupamiento k- medias

los estudiantes de rendimiento bajo, de rendimiento medio y de rendimiento alto. Tal clasificación puede apreciarse claramente en la figura 2, donde se aprecian comportamientos diferenciados.

Sin embargo, del análisis del efecto que podrían tener las características de la escuela de procedencia, en tales grupos, se obtuvo que el rendimiento de los estudiantes no es afectado, en general, por dichas características.

En lo que se refiere al fenómeno de deserción, mediante el modelo de riesgos proporcionales de Cox, no se encontró suficiente evidencia para concluir sobre la existencia y el tipo de factores que podrían influir en la tasa de deserción. Lo cual quiere decir que la susceptibilidad de desertar se presenta de la misma manera para hombres y mujeres, y además no se ve influenciada por el tipo de escuela de la cual proviene el individuo.

Contrariamente, el ajuste estadístico de la información mostró que la influencia sobre la tasa de egreso es por parte de las variables sexo del estudiante y sostenimiento de la escuela de procedencia. El modelo ajustado para el riesgo de egreso es:

$$h(t \mid x) = h_0(t)e^{\beta x} = h_0(t)e^{-0.359SEXO+0.242SOSTENIM}$$

a partir del cual es posible conocer las curvas de supervivencia de individuos con características que afectan el egreso.

En figura 3 podemos observar que la variable sexo tiene una gran influencia en el fenómeno del egreso; como se puede notar, se presentan diferencias significativas en las curvas correspondientes a hombres y mujeres. Además es posible distinguir que las mujeres egresan con mayor rapidez que los hombres. Por otro lado, también existe diferencia entre

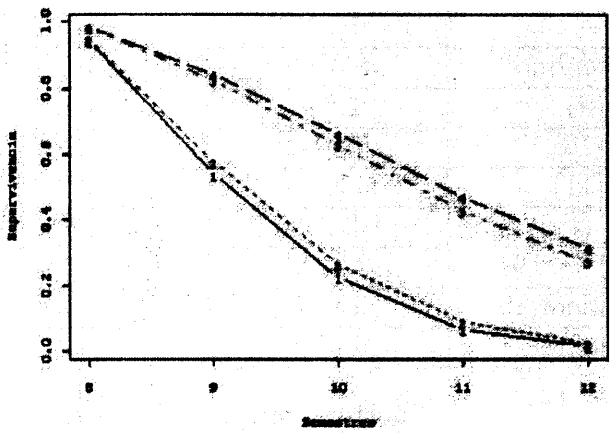


Figura 3: Curvas de supervivencia de los egresados. 1) Mujeres de escuelas privadas. 2) Mujeres de escuelas públicas. 3) Hombres de escuelas privadas. 4) Hombres de escuelas públicas.

los individuos que provienen de escuelas públicas y los de escuelas privadas, siendo estos últimos los que egresan con mayor rapidez.

Cabe mencionar que el análisis anterior es una aproximación ya que la variable bajo estudio es discreta. Esto produce que las pruebas de diagnóstico asociadas al modelo no produzcan resultados totalmente confiables. Asimismo, como las observaciones de los estudiantes egresados no presentan censura se pudo ajustar adicionalmente un modelo log-lineal para sustentar los resultados antes obtenidos con el modelo de Cox. El modelo resultante es el siguiente:

$$\log(Frec_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}\,,$$

donde $Frec_{ijk}$ es la frecuencia observada de la ijk-ésima casilla en la tabla de contingencia correspondiente al cruce de las variables número de semestres, sexo del individuo y sostenimiento de la escuela de procedencia,

μ es la media general,

α_i es el efecto de la i-ésima catergoría del sexo $(i = 0, 1)$,

β_j es el efecto de la j-ésima categoría del semestre $(j = 8, 9, 10, 11)$,

γ_k es el efecto de la k-ésima categoría del sostenimiento $(k = 0, 1)$,

$(\alpha\beta)_{ij}$ es el efecto ij de la interacción entre las variables sexo y semestre.

El ajuste del modelo anterior se realizó utilizando el paquete de cómputo GLIM, cuyo resultado se muestra en la siguiente tabla.

Parámetro	Efecto estimado	Efecto estandarizado
Media	1.040	2.3085
Sexo (masculino)	0.3365	0.575
Semestre (9)	2.054	4.3269
Semestre (10)	0.6932	1.2679
Semestre (11)	0.3365	0.57115
Sostenim. (Privada)	−0.2654	2.0167
Sex(masc)Sem(9)	0.2195	0.3547
Sex(masc)Sem(10)	1.624	2.4052
Sex(masc)Sem(11)	1.050	1.4548

Este modelo nos indica que el comportamiento del egreso es de la siguiente manera. La mayor parte de las mujeres egresan en los semestres 8 y 9. Además, los hombres egresan mayoritariamente en los semestres 10 y 11. Cabe señalar que es precisamente en el décimo semestre donde el egreso de estudiantes de sexo masculino se presenta de manera acentuada. Por otra parte, es muy baja la proporción de egresados provenientes de escuelas privadas.

A partir de los residuos de Pearson asociados al ajuste anterior y dados en la siguiente tabla, podemos tener una idea sobre la confiabilidad de estos resultados. Como puede observarse, el modelo ajustado muestra satisfactoriamente el patrón de valores observados.

Celda	Frecuencia observada	Frecuencia ajustada	Residuo de Pearson
1	2	2.830	-0.493
2	20	22.072	-0.441
3	3	5.660	-1.118
4	5	3.962	0.522
5	4	3.962	0.019
6	34	38.485	-0.723
7	44	40.183	0.602
8	21	15.847	1.294
9	3	2.170	0.563
10	19	16.928	0.504
11	7	4.341	1.276
12	2	3.038	-0.596
13	3	3.038	-0.022
14	34	29.515	0.826
15	27	30.817	-0.688
16	7	12.153	-1.478

4 Conclusiones

En resumen, el análisis estadístico nos permite concluir que:

1. El rendimiento de los estudiantes así como la tasa de deserción, en general, no están influenciados por las características del bachillerato de procedencia.

2. Las discrepancias entre los rendimientos posiblemente se deban a la influencia de otro tipo de factores que no fueron incluidos en el análisis. Por esta razón, es recomendable para futuros estudios que se obtenga información de los individuos tal como tiempo de dedicación al estudio, métodos de estudio, nivel socioeconómico, etc.

3. Del modelo de riesgos proporcionales de Cox, se desprende lo siguiente:

   (a) Las variables Sexo y Sostenimiento afectan el comportamiento del tiempo de egreso.

   (b) La variable Sexo tiene el efecto más significativo sobre el tiempo de egreso.

   (c) Las mujeres tienden a egresar más rápido que los hombres a lo largo de los semestres.

   (d) Estudiantes de escuelas privadas tienden a egresar en semestres más tempranos que los de escuelas públicas.

   (e) Además, el orden de egreso es: mujeres de escuelas privadas, mujeres de escuelas públicas, hombres de escuelas privadas, hombres de escuelas públicas.

4. Del modelo log-lineal:

   (a) Se confirman las tendencias observadas con el modelo de riesgos proporcionales.

   (b) La mayoría de las mujeres egresan en los primeros semestres (8 y 9).

   (c) La mayoría de los hombres egresa a partir del décimo semestre.

Suplementos Antioxidantes y Salud Respiratoria de los Boleros de la Ciudad de México con Relación a su Exposición al Ozono. Un Análisis Bayesiano

Leticia Gracia-Medrano y Silvia Ruiz-Velasco
IIMAS-UNAM

1 Antecedentes

El efecto de la exposición al ozono en la salud respiratoria ha sido un tema de investigación en epidemiología, diversos estudios se han llevado a cabo, particularmente en poblaciones de alto riesgo.

La exposición al ozono se ha relacionado con inflamación del tracto respiratorio que puede llevar, entre otras cosas, a decrementos en las funciones pulmonares y aumento en síntomas respiratorios. El efecto biológico del ozono se ha atribuido a su habilidad para causar oxidación o peroxidación de biomoléculas por medio de reacciones de radicales libres.

Por otro lado se conocen las propiedades de los siguientes suplementos antioxidantes:

La Vitamina E es un lípido soluble antioxidante que representa la principal defensa contra daños en la membrana de tejido humanos inducidos por oxidación.

La Betacarotena es un precursor de la vitamina A, su acción acumula tejidos en las membranas y es un purificador del radical O^{2-}

La Vitamina C contribuye a la actividad antioxidante y a la regeneración de membrana de los tejidos.

2 Estudio

Un grupo de boleros, que trabajaban en el centro de la Ciudad de México, fue reclutado para participar en un ensayo clínico cuyo objetivo era determinar el posible efecto protector de suplementos antioxidantes en la salud respiratoria. Este estudio fue dirigido por la Dra. Isabel Romieue del Instituto Nacional de Salud Pública, en 1997.

Los criterios para elegir a los integrantes del fueron:

1. Hombres entre 18 y 58 años.

2. No fumadores o fumadores moderados (menos de 10 cigarrillos diarios).

3. Trabajar a menos de 2 Km del monitor ambiental de la Merced.

A cada uno de los integrantes se les hizo un cuestionario para recabar datos acerca de aspectos socio-demográficos, de salud respiratoria y de hábitos alimenticios.

El estudio fue del tipo doble ciego y cruzado generando observaciones repetidas acerca de la salud respiratoria, estas observaciones no son necesariamente equiespaciadas dentro de cada período. Se registraron también variables exógenas como la temperatura y los niveles de contaminación; y para cada período se asignó una variable para determinar si el individuo recibía tratamiento o placebo.

La primera fase del estudio fue de marzo 13 a mayo 30 de 1997 y la segunda fase del estudio del junio 18 a agosto 8 de 1997, el periodo de "lavado" fue de al menos dos semanas, dado que no todos los boleros fueron evaluados el mismo día.

Los participantes acudían al centro de salud dos veces por semana a realizar una espirometría, al final de su día de trabajo y en ese momento se recolectaba la siguiente información:

1. la hora del día en que empezaron a trabajar.

2. la presencia de algún síntoma respiratorio durante el día.

3. si había fumado durante el día.

Como medida de exposición a los contaminantes se construyeron tres variables, por cada contaminante, en este caso ozono y bióxido de nitrógeno.

1. la exposición acumulada del día (del momento en que comenzaron a trabajar a la hora de la espirometría).

2. la exposición del día anterior (utilizando de la hora que habitualmente empieza a trabajar y la hora que habitualmente termina).

3. la exposición acumulada del mismo día y del día previo (1+2).

3 Modelo propuesto

En este trabajo consideramos como variable respuesta, la presencia de algún síntoma de los siguientes: tos, flemas,catarro, disnea, irritacion de ojos; el día de la visita al centro de salud.

Las variables explicativas consideradas fueron: la edad, si fuma o no, si fumó ese día, la temperatura mínima, y niveles de contaminación por ozono: ademas de considerar el efecto del tratamiento y de acarreo.

Si se supone que la probabilidad de que la respuesta binaria y_{ijt} sea positiva (síntoma presente, en el individuo i, en el periodo j y en la repetición t) puede ser modelada por la función

$$P(y_{ijt} = 1 \mid b_i) = \Phi(x'_{ijt}\beta + b_i), \tag{1}$$

donde b_i representa un efecto aleatorio con una distribución normal con media cero y varianza σ^2 desconocida.

La contribución del i-ésimo individuo a la verosimilitud está dada por:

$$\int \prod_{j=1}^{2} \prod_{t=1}^{n_{ij}} \Phi(x_{ijt}\beta + b_i)^{y_{ijt}}[1 - \Phi(x_{ijt}\beta + b_i)]^{1-y_{ijt}} exp(-\frac{b_i}{2\sigma^2})\sigma^{-1} db_i$$

La manera clásica de resolver el problema es integrar el efecto aleatorio y maximizar la verosimilitud.

Aun desde el punto de vista Bayesiano, la distribución aposteriori es difícil de trabajar. El modelo que se utilizó está basado en el propuesto por Albert y Chib, 1996. La idea que estos autores presentan es introducir una variable no observada (latente), z_{ijt} tal que

$$z_{ijt} \mid b_i \sim N(x_{ijt}\beta + b_i, 1)$$

y hacen que la variable respuesta observada y_{ijt} indique si la variable latente fue positiva o negativa:

$$y_{ij} = \begin{cases} 1 & \text{si } z_{it} > 0 \\ 0 & \text{si } z_{it} \leq 0 \end{cases}$$

Es claro que estas y_{ij} satisfacen la ecuación (1). De esta manera la instrumentación del algoritmo de Gibbs se simplifica. Y puede pensarse que la variable respuesta es continua y que si rebasa cierto umbral,se dicotomiza.

Los autores entonces modelan el valor esperado de las variables latentes de la siguiente manera:

$$E(z_{ijt} \mid b_i) = \beta_0 + \beta_1 x + \beta_t t_{ijt} + \beta_p p_{ijt} + \beta_c c_{ijt} + \beta_w w_{ijt} + b_i$$

Donde x representa un vector que contiene las variables que no cambian a través del estudio, t_{ijt}, p_{ijt} y c_{ijt} representan variables indicadoras del tratamiento, período y efecto de acarreo y w_{ijt} representa un vector que contiene las variables de exposición al ambiente. Para completar el modelo bayesiano se asigna una distribución a priori uniforme a β y una $GI(\nu, \delta)$ a σ^2.

Utilizando el algoritmo de Gibbs, podemos obtener muestras de la distribución conjunta de $(\{z_{ijt}\}, \beta, \{b_i\}, \sigma^2)$, simulando sucesiones de valores de las distribuciones condicionales. En este caso consideramos

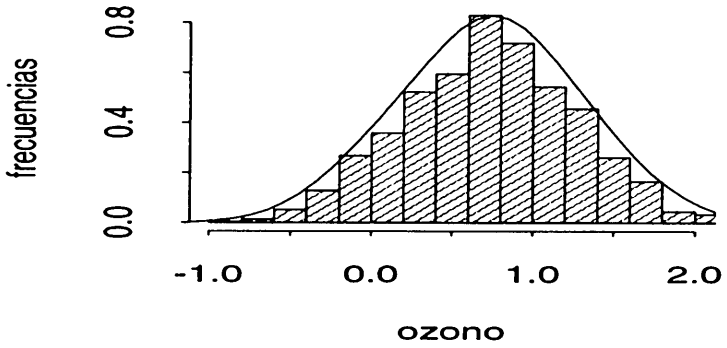


Figura 1: Coeficiente del ozono

1. La distribución de $\{z_{ijt}\}$ dados $Y, \{bi\}$, es una normal truncada,

$$\begin{cases} N_{(-\infty,0)}(x'_{ijt}\beta + b_i, 1) & \text{si } y_{ijt} = 0 \\ N_{(0,\infty)}(x'_{ijt}\beta + b_i, 1) & \text{si } y_{ijt} = 1 \end{cases}$$

2. La distribución de β dados $\{z_{ijt}\}, \{b_i\}$ es una $N_p(\hat{\beta}, B^{-1})$ con $\hat{\beta} = B^{-1}\sum_{i=1}^{n} X_i(z_i - i_t b_i)$ y $B = \sum_{i=1}^{n} X_i' X_i$.

3. La distribución de $\{b_i\}$ dados $\{z_{ijt}\}, \sigma^2$ y β que es $N(\hat{b_i}, V_i^{-1})$ con $\hat{b_i} = V_i^{-1}\sum_{ij=1}^{n_{ij}}(z_{ij} - x_{ij}\beta)$ y $V_i = n_{ij} + \sigma^{-2}$.

4. La distribución a posteriori de de la varianza de los efectos aleatorios σ^2, dados $\{z_{ijt}\}, \{b_i\}$ es una $GI(\nu + \frac{n}{2}, \delta + \frac{\sum_{i=1}^{n} b_i^2}{2})$.

3.1 El modelo elegido para los boleros de la Ciudad de México

El estudio constó con una muestra de 34 individuos, haciendo un total de 729 repeticiones.

Las variables explicativas incluidas en el modelo final fueron la media general, la variable que indicaba si fumaba o no, la edad de la persona, el tipo de tratamiento que estaba recibiendo, la medición del ozono, la interacción del ozono con el tratamiento, la temperatura mínima del día de la medición, la época, la variable indicando si existe o no efecto de acarreo, y el número de horas trabajadas.

En el algoritmo de Gibbs, la solución de máxima verosimilitud se eligió como punto inicial, y se hicieron un total de 2000 iteraciones. Para las distintas βs se hicieron gráficos

de autocorrelación para verificar su convergencia. En esta ocasión los gráficos para los coeficiente que multiplican a la variables ozono, interacción del ozono con el tratamiento y la temperatura mínima muestran ya convergencia para estas 2000 iteraciones.

Desafortunadamente el coeficiente correspondiente al tratamiento con suplementos antioxidantes y el resto de los coeficientes no alcanzaron convergencia y serán necesarias más iteraciones para poder hablar acerca de su contribución al modelo. De los tres coeficientes que alcanzaron convergencia sólo el histograma del ozono aparece centrado en un valor distinto al cero (cercano al 0.9), ver figura, significando esto que la presencia de este contaminante provoca un aumento en la probabilidad de presentar algún síntoma.

El manejo de los datos se hizo a través del código S-Plus, en una pentium, con un tiempo de cómputo aproximado de 6 horas.

Referencias

Albert, J. H. y Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *JASA Theory and Methods.* vol. 88, 422.

Predicción Vía *Bootstrap* Bayesiano

Eduardo Gutiérrez Peña
IIMAS, UNAM

1 Introducción

Considérese el siguiente problema: dada una muestra aleatoria de una distribución desconocida F, se desea hacer inferencias sobre el valor de una observación futura proveniente de dicha distribución. En este trabajo se introduce un procedimiento Bayesiano para atacar este problema. Dicho procedimiento se basa en un criterio Bayesiano de selección de modelos y hace uso de técnicas de Monte Carlo para llevar a cabo los cálculos necesarios.

2 Bootstrap: ideas básicas

Sea $X_1 = x_1, \ldots, X_n = x_n$ una muestra de observaciones independientes e idénticamente distribuidas de una función de distribución desconocida F definida sobre $\mathbb{R}$. Supongamos que se desea hacer inferencias sobre el parámetro $\theta = H(F)$, donde $H(\cdot)$ es una funcional de F; por ejemplo, $H(F) = \int x\, dF(x)$ ó $H(F) = \inf\{t \in \mathbb{R} : F(t) \geq q\}$ para algún valor $0 < q < 1$.

Sea $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$ la función de distribución empírica de la muestra $x_1, \ldots, x_n$, donde $I(\cdot)$ denota la función indicadora. Entonces, bajo ciertas condiciones bastante generales, $\hat{\theta} = H(F_n)$ es un estimador consistente de θ.

Efron (1979, 1982) introduce un método, llamado *bootstrap*, cuyo propósito es aproximar la distribución muestral del estimador $\hat{\theta}$. En términos generales, el método de Efron consiste en lo siguiente:

Para $r = 1, \ldots, B$,

1. Generar $x_1^{(r)}, \ldots, x_n^{(r)} \overset{i.i.d.}{\sim} F_n(x)$
2. Calcular $\hat{\theta}^{(r)} = H(F_n^{(r)})$, donde $F_n^{(r)}$ denota la función de distribución empírica de la muestra $x_1^{(r)}, \ldots, x_n^{(r)}$.

La muestra $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$ es usada entonces para aproximar la distribución muestral de $\hat{\theta}$. En particular, este método proporciona una aproximación numérica del error estándar de $\hat{\theta}$ en situaciones donde no es posible calcularlo de manera analítica.

Rubin (1981) discute un procedimiento que puede ser considerado como la versión Bayesiana del método de Efron. La idea de este procedimiento, denominado *bootstrap Bayesiano*, puede describirse de manera sencilla a través del siguiente algoritmo:
Para $r = 1, \ldots, B$,

1. Generar $(\omega_1^{(r)}, \ldots, \omega_n^{(r)}) \sim \text{Dir}_n(1, \ldots, 1)$, donde Dir_n denota a la distribución Dirichlet n-variada.

2. Calcular

$$\tilde{F}_n^{(r)}(x) = \sum_{i=1}^{n} \omega_i^{(r)} I(x_i \leq x).$$

3. Calcular $\theta^{(r)} = H(\tilde{F}_n^{(r)})$.

La idea es utilizar la muestra $\theta^{(1)}, \ldots, \theta^{(B)}$ para aproximar la distribución final de θ y de esta manera hacer inferencias sobre el parámetro de interés.

3 Descripción del problema

Supongamos ahora que se desea hacer inferencias sobre el valor de una observación futura $Y = X_{n+1}$ de la distribución F. El bootstrap Bayesiano sugiere entonces el siguiente algoritmo:

Para $r = 1, \ldots, B$,

1. Generar $(\omega_1^{(r)}, \ldots, \omega_n^{(r)}) \sim \text{Dir}_n(1, \ldots, 1)$.

2. Generar $y^{(r)} \sim \tilde{F}_n^{(r)}(y) = \sum_{i=1}^{n} \omega_i^{(r)} I(x_i \leq y)$.

La muestra $y^{(1)}, \ldots, y^{(B)}$ es usada entonces para aproximar la distribución predictiva final de Y. De esta forma es posible hacer inferencias aproximadas sobre el valor de la observación futura Y.

Tanto el bootstrap clásico de Efron como el bootstrap Bayesiano de Rubin son casos particulares del llamado bootstrap ponderado (ver, por ejemplo, Præstgaard y Wellner, 1993). Un inconveniente de este tipo de procedimientos es que asignan probabilidad cero al evento de que una observación futura no tome uno de los valores observados en la muestra original $x_1, \ldots, x_n$.

El procedimiento propuesto en la Sección 5 "suaviza" las probabilidades asignadas a los distintos valores de X, eliminando así el problema de los métodos de bootstrap ponderado.

4 Un Criterio Bayesiano de selección de modelos

Sea $\mathcal{M} = \{M_\lambda : \lambda \in \Lambda\}$ una colección de modelos paramétricos, donde

$$M_\lambda = \{p_\lambda(x|\theta_\lambda), p_\lambda(\theta_\lambda)\}.$$

Los modelos en $\mathcal{M}$ se eligen generalmente porque son relativamente sencillos de analizar y/o porque facilitan la comunicación de resultados, y tienen el propósito de describir el comportamiento de una variable aleatoria X.

Desde el punto de vista Bayesiano, el problema de elegir un modelo en $\mathcal{M}$ puede plantearse como un problema de decisión con los siguientes elementos (Gutiérrez-Peña y Walker, 1998):

Espacio de decisiones.

$$\mathcal{D} = \Lambda$$

Espacio de estados de la naturaleza.

$$\mathcal{F} = \{F : F \text{ es una función de distribución con soporte apropiado}\}$$

Distribución inicial.

$$F \sim \mathcal{DP}(\alpha_0, F_0).$$

donde $\mathcal{DP}(\alpha_0, F_0)$ denota un proceso Dirichlet con media F_0 y parámetro de escala α_0 (Ferguson, 1973).

Función de utilidad.

$$U(\lambda, F) = \int \log f(y;\lambda)\, dF(y),$$

donde

$$f(y;\lambda) = \int p_\lambda(y|\theta_\lambda)\, p_\lambda(\theta_\lambda)\, d\theta_\lambda$$

es la distribución predictiva final bajo el λ-ésimo modelo, dada una muestra $X_1, \ldots, X_n$ de F.

La distribución final de F está dada por

$$[F|\boldsymbol{x}] \sim \mathcal{DP}(\alpha_n, G_n),$$

donde $\alpha_n = \alpha_0 + n$ y

$$G_n = \frac{\alpha_0 F_0 + n F_n}{\alpha_0 + n}$$

(ver Ferguson, 1973).

En caso $\alpha_0 = 0$ se interpreta generalmente como no informativo. En este caso, la utilidad esperada final está dada por

$$\bar{U}(\lambda) = \int \log f(y;\lambda)\, dF_n(y) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i;\lambda).$$

El modelo óptimo corresponde entonces al valor λ_* que maximiza $\bar{U}(\lambda)$.

5 Procedimiento propuesto

Supongamos ahora que, en lugar de maximizar la utilidad esperada final $\bar{U}(\lambda)$, generamos una muestra $\tilde{F}_n^{(1)}, \ldots, \tilde{F}_n^{(B)}$ de la distribución final de F, es decir, del proceso Dirichlet $\mathcal{DP}(n, F_n)$. Esto puede hacerse fácilmente a través del siguiente algoritmo:

Para $r = 1, \ldots, B$,

1. Generar $(\omega_1^{(r)}, \ldots, \omega_n^{(r)}) \sim \mathrm{Dir}_n(1, \ldots, 1)$.
2. Definir $\tilde{F}_n^{(r)}(y) = \sum_{i=1}^{n} \omega_i^{(r)}\, I(x_i \leq y)$.

El método que proponemos consiste en:

(a) Maximizar, para cada $r = 1, \ldots, B$,

$$\bar{U}^{(r)}(\lambda) = \sum_{i=1}^{n} \omega_i^{(r)} \log f(x_i;\lambda),$$

obteniendo valores $\lambda_*^{(1)}, \ldots, \lambda_*^{(B)}$.

(b) Generar, para cada $r = 1, \ldots, B$,

$$y^{(r)} \sim f(y;\lambda_*^{(r)}).$$

Los valores $y^{(1)}, \ldots, y^{(B)}$ forman una muestra bootstrap de la distribución predictiva final de la observación futura $Y = X_{n+1}$. Cabe señalar que este procedimiento hace uso de los modelos paramétricos considerados por el analista como posibles candidatos para modelar la distribución de la variable aleatoria X. Como consecuencia, a diferencia de los métodos de bootstrap ponderado, las muestras de "observaciones futuras" no necesariamente toman valores sólo en el conjunto definido por la muestra original $x_1, \ldots, x_n$.

Si $\theta_\lambda = \theta$ es un parámetro común a todos los modelos en la clase $\mathcal{M}$ y $p_\lambda(\theta) = \delta_\lambda(\theta)$ (donde $\delta(\cdot)$ es la medida de Dirac), entonces

$$U^{(r)}(\lambda) = \log \tilde{L}(\lambda)$$

donde $\tilde{L}(\lambda)$ es la verosimilitud ponderada para λ propuesta por Newton y Raftery (1994). Estos autores utilizan la muestra $\lambda_*^{(1)}, \ldots, \lambda_*^{(B)}$ para hacer inferencias sobre θ.

Ejemplo. Los datos para este ejemplo consisten en $n = 30$ observaciones de tiempos de falla correspondientes al equipo de aire acondicionado de una muestra de aviones (Proschan, 1963).

Se consideraron los siguientes dos modelos:

Modelo 1. $M_1 = \{p_1(x|\theta_1), p_1(\theta_1)\}$,
donde

$$p_1(x|\theta_1) = \theta_1^{-1} \exp\{-x/\theta_1\} \quad \text{y} \quad p_1(\theta_1) \propto 1/\theta_1.$$

Modelo 2. $M_2 = \{p_2(x|\theta_2), p_2(\theta_2)\}$,

donde $\theta_2 = (\mu, \sigma^2)$,

$$p_2(x|\theta_2) = p_2(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\log x - \mu)^2\right\}$$

y

$$p_2(\theta_2) = p_2(\mu, \sigma^2) \propto 1/\sigma^2.$$

En este caso $\lambda \in \{1, 2\}$.

Las densidades predictivas finales para cada uno de los modelos están dadas por

$$f(y; 1) = \frac{n\,(n\bar{x})^n}{(y + n\bar{x})^{n+1}}$$

donde $\bar{x} = 59.6$, y

$$f(y; 2) = \text{log-St}\left(y \mid \bar{t}, \left[\frac{n+1}{n-1}\right] T^2, n-1\right)$$

con

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} \log x_i = 3.36, \qquad T^2 = \frac{1}{n}\sum_{i=1}^{n} (\log x_i - \bar{t})^2 = 1.74,$$

y donde log-St denota una densidad log-Student.

Las utilidades esperadas finales correspondientes son

$$\bar{U}(1) = -5.08 \quad \text{y} \quad \bar{U}(2) = -5.06,$$

por lo que el modelo óptimo es el Modelo 2.

Se obtuvo una muestra bootstrap $y^{(1)}, \ldots, y^{(B)}$ (de tamaño B=10,000) de la distribución predictiva final de Y con base en el siguiente algoritmo:

Para $r = 1, \ldots, B$,

1. Generar $(\omega_1^{(r)}, \ldots, \omega_n^{(r)}) \sim \text{Dir}_n(1, \ldots, 1)$.

2. Calcular

$$\bar{U}^{(r)}(1) = \sum_{i=1}^{n} \omega_i^{(r)} [\, n \log\{n\,(n\bar{x})\} - (n+1) \log(x_i + n\bar{x})\,]$$

y

$$\bar{U}^{(r)}(2) = \sum_{i=1}^{n} \omega_i^{(r)} \text{ log-St}\left(x_i \mid \bar{t}, \left[\frac{n+1}{n-1}\right] T^2, n-1\right)$$

3. Si $\bar{U}^{(r)}(1) \geq \bar{U}^{(r)}(2)$, generar $y^{(r)} \sim f(y;1)$.
   Si $\bar{U}^{(r)}(1) < \bar{U}^{(r)}(2)$, generar $y^{(r)} \sim f(y;2)$

En este caso es posible encontrar una expresión analítica para la distribución de la muestra bootstrap, que corresponde a la mezcla

$$\hat{f}(y) = 0.36\, f(y;1) + 0.64\, f(y;2).$$

Notemos que las utilidades esperadas finales son aparentemente muy similares entre sí, aunque no es fácil juzgar esta similitud debido a que se carece de una escala de referencia. Sin embargo, los pesos relativos asignados a cada uno de los modelos en la distribución de la muestra bootstrap sugieren que hay evidencia sustancial a favor del Modelo 2.

Referencias

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics* **7**, 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics* **1**, 209-230.

Gutiérrez-Peña, E. y Walker, S.G. (1998). A Bayesian Predictive Approach to Model Selection. *Sometido.*

Newton, M.A. y Raftery, A.E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society B* **56**, 3-48 (con discusión).

Præstgaard, J. y Wellner, J.A. (1993). Exchangeable Weighted Bootstraps of the General Empirical Process. *Annals of Probability* **21**, 2053-2086.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics* **5**, 375-383.

Rubin, D.B. (1981). The Bayesian Bootstrap. *Annals of Statistics* **9**, 130-134.

Un Modelo de Pronóstico para un Sistema de Transporte de Valores

Luis F. Hoyos Reyes y José C. Romero Cortés

UAM-Azcapotzalco

1 Introducción

Ante la creciente inseguridad las compañías de transporte de valores buscan no solo diversificar sus servicios, sino también hacerlos más eficientes.

Entre los clientes más importantes se encuentran los bancos. Un sistema de transporte de valores está integrado por dos componentes: la compañía de seguridad que transporta los valores y los bancos. En nuestro caso particular el transporte se limitará a dinero en efectivo.

Las interacciones de flujo de efectivo son de dos tipos: cuando el efectivo es transportado de la compañía al banco (dotaciones) y viceversa, el efectivo va del banco a la compañía (concentraciones).

Los bancos tratan de operar con un mínimo de efectivo que garantice operatividad es decir, que exista suficiente circulante para cubrir todas las operaciones de ventanilla, como pagos de cheques, compra y venta de divisas, por mencionar dos ejemplos.

Esta necesidad surge de la relación riesgo - monto de efectivo, que efectivamente es proporcional, a mayor monto de efectivo mayor es el riesgo. Sin embargo si el banco no tiene efectivo incurre en costos de oportunidad, ya que la capacidad del banco de retener a sus clientes y atraer más ahorradores disminuye (ver Ross y Westerfield, 1996).

Existen modelos de administración de efectivo muy estudiados como los propuestos por Baumol (1952) y Miller y Orr (1966). Ambos modelos evalúan la necesidad de disponibilidad de efectivo en términos del monto actual.

El objetivo de este trabajo consiste en proponer un modelo de regresión dinámico a partir de los montos correspondientes a las dotaciones y concentraciones por unidad de tiempo con la finalidad de estimar la dotación del futuro inmediato.

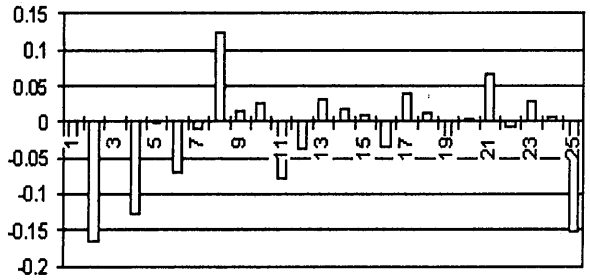


Figura 1: Autocorrelaciones de los Residuales

2 Construcción del modelo de regresión dinámica

El modelo propuesto es un modelo dinámico de regresión con una parte correspondiente a una función de transferencia entre las variables y cuyo residual presenta una forma ARMA.

Para identificar la función de transferencia para nuestras variables:

Y_t = dotación en el tiempo t

X_t = concentración en el tiempo t

Primero hubo que diferenciarlas como:

$DIFY_t = Y_t - Y_{t-k}$

$DIFX_t = X_t - X_{t-k}$

Y sobre estos se calculó la función de autocorrelaciones cruzada, presentando sólo una diferente de cero, identificándose la relación

$DIFY_t = c + \beta_1 DIFX_{t-m} + \beta_2 D + \epsilon_t$

Dicha función se ajustó y analizando la función de correlación cruzada entre los residuales y los α_t (input sin ruido blanco), el modelo resultó adecuado (ver Pindyck y Rubinfeld, 1990).

Cabe aclarar que D es una variable dummy que toma el valor de 2.3 si $X_t > 1.5$ unidades y cero en cualquier otro caso. Esto específicamente para cada sucursal.

Este modelo tiene asociado una $R^2 = 0.394$, el coeficiente de Durbin-Watson es 2.065, $S_c = 374.36$ y las variables resultaron significativas.

Analizando los residuales del modelo estimado como una serie de tiempo en base a las autocorrelaciones y autocorrelaciones parciales se identificó un modelo ARMA que agregamos al modelo anterior, y tenemos:

$DIFY_t = \alpha_0 + \alpha_1 DIFX_{t-m} + \alpha_2 D + \alpha_3 a_{t-l} + \alpha_4 DIFY_{t-l} + \alpha_5 DIFY_{t-(l+s)} + \epsilon_t$

El modelo se ajustó y los residuales asociados son esencialmente ruido blanco según se observa en las figuras 1 y 2. Además las pruebas de Box-Pierce y la de Ljung-Box (ver Box et al., 1994) resultaron no significativas. Por otra parte, este modelo tiene un coeficiente de determinación múltiple de 0.994, Durbin-Watson igual a 1.97, $S_c = 38.86$ y

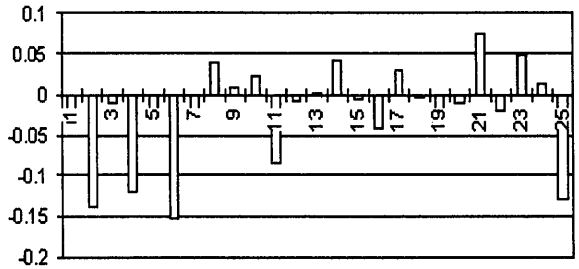


Figura 2: Autocorrelaciones Parciales Residuales

con significancia en las variables.

Debido a la gran cantidad de sucursales consideradas sólo hemos presentado la técnica en general, los lags m,l y s dependen de las operaciones de cada banco.

Cabe mencionar que en ocasiones los coeficientes estimados estuvieron cercanos a 1 para el término de media móvil, sin embargo hubo convergencia en la estimación en aproximadamente 60 iteraciones usando TSP o ETS de SAS, eliminando este término donde pudiera haber problemas de invertibilidad el modelo continuo manteniendo bondades similares a las señaladas.

El modelo además de parsimonioso resulta adecuado en muchas sucursales de diferentes regiones y bajo horizontes de tiempo grandes.

Una vez identificado, estimado y verificado, el modelo fue usado para estimar la dotación de efectivo del día siguiente, tanto puntualmente como por intervalo.

3 Conclusiones

El modelo se comporta adecuadamente bajo un punto de vista teórico: el coeficiente de determinación múltiple es 0.9943 y el estadístico de Durbin-Watson es cercano a 2 y las variables resultan significativas.

Sin embargo en la perspectiva de un problema general aún falta mucho por hacer. Es posible construir un modelo de interacciones en red con 2 tipos de nodos: sucursales bancarias y sucursales de protección de valores. Posteriormente podemos analizar para cada sucursal las condiciones que garanticen la existencia de una política óptima de operación del modelo de control de flujos de efectivo en la red.

La optimalidad la entenderíamos en el sentido de minimizar el monto de efectivo en cada sucursal con la restricción de dicho monto debe satisfacer las necesidades operativas.

Referencias

Baumol, W.S. (1952). The Transactions Demand for Cash: An Inventory Theoretic Approach. *Quaterly Journal of Economics*, **66**,108-122.

Box, G.E.P., Jenkins, G.M., y Reinsel, G.C. (1994) *Time Series Analysis.* Prentice Hall.

Jaffe, J., Ross, S. y Westerfield, R.W. (1996). *Corporate finance.* Irwin McGraw-Hill.

Miller, M.H. y Orr, D. (1966). A Model of the Demand for Money by Firms. *Quaterly Journal of Economics*, august 1966.

Pindyck, R.S. y Rubinfeld, D.L. (1990). *Econometric Models and Economic Forecasts.* Third Edition, McGraw-Hill.

La Incorporación de la Estructura de Covarianzas en la Evaluación de Servicios a Través de Escalas

Eduardo A. Izquierdo Gutiérrez y Olivia Carrillo Gamboa
ITESM, Campus Monterrey

1 Introducción

Es frecuente encontrarse con encuestas donde las respuestas se presentan como valores en una escala de números enteros. Tales encuestas se utilizan con frecuencia en la evaluación de servicios, presentando para cada área a evaluarse un número variable de preguntas acerca de los rubros o aspectos a calificar. En este trabajo se presenta una alternativa de análisis realizada en una encuesta de este tipo, la cual se aplicó en un departamento del Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM) con el objetvo de medir la calidad de los servicios que prestan.

La encuesta es aplicada a una muestra aleatoria de alumnos del instituto, los cuales evalúan la calidad de cada uno de los diferentes servicios que presta la Dirección Administrativa y de Planta Física (DAPF). Los servicios o áreas son veinte, los cuales son evaluados a través de entre cuatro y siete rubros específicos. Por ejemplo, uno de los servicios es el denominado AULAS, en el cual se evalúan los rubros limpieza, nivel de iluminación, condiciones del mobiliario, ventilación y condiciones del mobiliario. Todos estos rubros están bajo la autoridad de dicho departamento. En este trabajo se presenta el cálculo del índice de calidad (media ponderada) y el cálculo de su margen de error bajo la consideración de las interdependencias naturalmente encontradas entre las evaluaciones proporcionadas por un mismo alumno a los diferentes rubros.

Es decir, es de esperarse alguna dependencia entre las respuestas para los diferentes aspectos dentro de un misma área, debido al hecho de que, para un área con k rubros a evaluarse, los datos obtenidos forman vectores de k dimensiones, cada uno de estos comprendiendo las respuestas de alguno de los encuestados en particular y es natural suponer que las respuestas provenientes de una misma persona estén correlacionadas, lo cual es confirmado por el análisis aquí mencionado.

El artículo presente propone una manera de obtener, bajo la presencia de estructura, la media general para cada área y la varianza de esta media, lo que da lugar a la posibilidad de calcular intervalos de confianza y tamaños de muestra. Se presenta aplicada la metodología propuesta a datos provenientes de la encuesta aplicada por la DAPF, en abril

de 1998, para conocer la opinión de los alumnos y empleados acerca de la calidad de los servicios que brinda. Resaltan dos supuestos en el momento de efectuar el desarrollo y los cálculos: (a) que las respuestas dadas para diferentes preguntas por una misma persona se encuentran correlacionadas, pero no así las respuestas para dos diferentes personas, es decir, las respuestas de algún individuo A son independientes de las de algún otro individuo B, y (b) la estructura de dependencias entre las respuestas de los encuestados es la misma para todos los encuestados.

2 Indice de calidad y su margen de error

Los resultados para las encuestas aludidas presentan para cada una de la áreas a evaluar una estructura similar a la siguiente:

Encuesta No.	(pregunta No. 1)	(pregunta No. 2)	$\cdots$	(pregunta No. k)
1	r_{11}	r_{12}	$\cdots$	r_{1k}
2	r_{21}	r_{22}	$\cdots$	r_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	r_{n1}	r_{n2}	$\cdots$	r_{nk}

Donde las respuestas r_{ij} toman valores de una escala de números enteros, para el caso particular de la encuesta cuyos datos serán analizados, estos valores son $1, 2, \ldots, 7$.

Las medias de las respuestas para cada ítem serán

$$\overline{X}_1 = \frac{1}{m_1} \sum_{i=1}^{m_1} r_{i1}$$
$$\vdots$$
$$\overline{X}_k = \frac{1}{m_k} \sum_{i=1}^{m_k} r_{ik}$$

Nótese que el número de datos utilizados para el cálculo de las diferentes medias puede variar de una pregunta a otra debido a la existencia de no respuesta. La media global será entonces

$$\overline{\overline{X}} = \sum_{i=1}^{k} \frac{m_i}{m} \overline{X}_i$$

donde

$$m = \sum_{i=1}^{k} m_i$$

Las varianzas y covarianzas muestrales de las respuestas vendrán dadas por :

$$s_1^2 = \frac{1}{m_1 - 1} \sum_{i=1}^{m_1} \left(r_{i1} - \overline{X}_1\right)^2$$
$$\vdots$$
$$s_k^2 = \frac{1}{m_k - 1} \sum_{i=1}^{m_k} \left(r_{ik} - \overline{X}_k\right)^2$$

y

$$\operatorname{cov}(X_h, X_l) = \frac{1}{q_{hl} - 1} \sum_i \left(r_{ih} - \overline{X}_h\right)\left(r_{il} - \overline{X}_l\right) = s_{hl}$$

respectivamente, donde la sumatoria para la obtención de las covarianzas se extiende sobre todos los pares de respuestas r_{ih}, r_{ik} en los que ninguna de ellas es dato faltante, y q_{hl} representa el número de tales pares de respuestas utilizados en el cálculo. Las varianzas para las medias se obtienen de la manera usual:

$$s^2_{\overline{X}_1} = \frac{s_1^2}{m_1}$$
$$\vdots$$
$$s^2_{\overline{X}_k} = \frac{s_k^2}{m_k}$$

y para la covarianza entre medias,

$$\operatorname{cov}\left(\overline{X}_h, \overline{X}_l\right) = \operatorname{cov}\left(\frac{1}{m_h}\sum_{i=1}^{m_h} r_{ih}, \frac{1}{m_l}\sum_{j=1}^{m_l} r_{jl}\right) = \frac{1}{m_h m_l} \operatorname{cov}\left(\sum_{i=1}^{m_h} r_{ih}, \sum_{j=1}^{m_l} r_{jl}\right)$$

apelaremos a los dos supuestos mencionados en la introducción. Así, y mediante las reglas conocidas para la covarianza, la última ecuación dada se reduce a

$$\operatorname{cov}\left(\overline{X}_h, \overline{X}_l\right) = \frac{q_{hl}}{m_h m_l} s_{hl}$$

Provistos ya con las covarianzas entre las medias de las respuestas para cada pregunta, la varianza de la media global resulta ser

$$\operatorname{var}\left(\overline{\overline{X}}\right) = \operatorname{var}\left(\sum_{i=1}^{k} \frac{m_i}{m} \overline{X}_i\right) = \frac{1}{m^2}\left[\sum_{i=1}^{k} m_i s_1^2 + 2 \sum_{i<j} \sum q_{ij} s_{hl}\right]$$

3 Aplicación

Las fórmulas arriba expuestas se utilizaron primeramente para el cálculo del tamaño de muestra para diferentes niveles de precisión bajo un nivel de confianza del 95%. La encuesta aplicada a alumnos por la DAPF en el ITESM consta de veinte áreas a evaluar, de las cuales se seleccionaron para el análisis siete, que fueron catalogadas como las más relevantes por el personal de la dirección, a saber: aulas, servicios sanitarios, seguridad del campus, laboratorios, estacionamientos, centros de copiado y "Locatec". Cada una de estas áreas comprende entre cuatro y siete preguntas. Como ya se mencionó anteriormente, las respuestas dadas como números enteros entre uno y siete, indicando el número uno excelente y pésimo el siete. Se tomaron de base para los cálculos los datos de una encuesta previa la cual fué aplicada a 618 alumnos.

El número de respuestas efectivas para cada pregunta resultó ser muy variable de un área a otra y por el tipo de servicio en algunas el porcentaje de no respuestas era considerable, hasta 400 o más del total de los 618 cuestionarios. Es decir, los altos índices de no respuesta a la naturaleza misma del área que se evaluaba en particular por ejemplo, una de las áreas era laboratorios: no todos los alumnos de la institución hacen uso de laboratorios durante su carrera y algunos de los que sí, sólo durante algunos semestres de la misma.

Para el análisis de los datos se utilizaron los paquetes STATGRAPHICS versión 5.0 y EXCEL 95. Se procedió primero al cálculo de las covarianzas y los coeficientes de correlación entre las respuestas de las diferentes preguntas dentro de cada área y se encontró que estos últimos eran todos sensiblemente diferentes de cero, como lo esperábamos, además resultaron positivos todos. A continuación presentamos las medias globales y las varianzas asociadas a ellas para cada una de las áreas mencionadas.

Área	Media global	Varianza de la media global
Aulas	2.653236	0.001396
Servicios sanitarios	2.581634	0.002103
Seguridad del Campus	2.410072	0.001506
Laboratorios	2.801469	0.006731
Estacionamientos	2.801469	0.003398
Centros de copiado	2.635493	0.002618
"Locatec"	2.570144	0.005135

4 Determinación de los tamaños de muestra

Para la obtención de tamaños de muestra, especificada una tolerancia D para las medias globales por área, se consideraron únicamente las tres primeras áreas (aulas, servicios sanitarios y seguridad), debido a que, de entre las siete áreas críticas por su naturaleza

eran las únicas que se esperaba contestaran la totalidad de los encuestados. Dada la imposibilidad de despejar el tamaño de muestra de la fórmula usada para la varianza, se procedió a usar la fórmula modificada

$$\text{var}\left(\overline{\overline{X}}\right) = \frac{1}{km}\left[\sum_{i=1}^{k} s_i^2 + 2\sum_{i<j}\sum s_{ij}\right],$$

la cual se diferencia de la anteriormente dada en el hecho de que no contempla la presencia de datos faltantes. Para utilizar esta fórmula se recalcularon las varianzas y covarianzas, usando ahora solo aquellos registros que no presentaron datos faltantes para ninguna de las preguntas dentro de cada área. Así para una precisión D dada y $\alpha = 0.05$, tenemos

$$D = 1.96\frac{1}{\sqrt{m}}\left\{\frac{1}{k}\left[\sum_{i=j}^{k} s_i^2 + 2\sum_{i<j}\sum s_{ij}\right]\right\}^{1/2}$$

de donde despejamos m. El tamaño de muestra efectivo (el número de personas entrevistadas, n) vendrá dado por

$$n = \frac{m}{1-\phi}$$

donde ϕ es la proporción de registros con datos faltantes en esa área, estimado de los datos desponibles de la muestra anterior.

Para el caso de la encuesta particular aquí considerada, de las tres áreas tomadas para el cálculo de n, se encontró que servicios sanitarios era la que presentaba mayor variabilidad, Var(media global)=0.002183, con una proporción de registros incompletos $\phi = 0.0874$, lo que para una $D = 0.1$ nos da un n de 568.

Indices de Capacidad del Proceso para Poblaciones Asimétricas

Ana Isabel Landeros y Graciela González Farías
ITESM, Campus Monterrey

1 Introducción

Existe mucha controversia acerca del uso de los índices de capacidad del proceso (C_p) para monitorear la calidad de un producto. La controversia incluye temas como la validez de los índices y su consistencia en el monitoreo del mejoramiento del proceso, el abuso de los usuarios al utilizar los estimadores como medidas de eficiencia y el qué tan apropiado es utilizar un índice en situaciones particulares donde la distribución de la característica de calidad de interés no es normal. A pesar de esto, los industriales siguen calculando estos índices porque sus clientes lo siguen solicitando para que se les proporcione una garantía de la calidad del producto.

Para medir la capacidad del proceso, su comportamiento esperado se especifica en términos de una característica de calidad la cual a su vez puede estar en función de un valor objetivo o nominal, un rango de aceptación o tolerancias, o ambos. Kushler y Hurley (1992) proporcionaron la siguiente definición general:

Definición 1.1 *Un índice de capacidad del proceso es una función de los parámetros de la distribución de la característica de calidad y las especificaciones establecidas por los requerimientos del proceso.*

En el enfoque de cualquier discusión acerca de los C_p's la motivación original - monitoreo de la proporción esperada fuera de los límites de especificaciones - se mantendrá presente. Entonces, la pregunta que surge es ¿Por qué no investigar directamente la proporción observada fuera de los límites de especificaciones? Esta sería una medida de cálculo más simple y más comprensible para los usuarios, como señalan Kotz y Johnson (1993).

Notación a utilizarse en la definición de los índices: USL Límite de especificación superior; LSL Límite de especificación inferior; μ Media del proceso; σ Desviación Estándar del proceso;τ Valor objetivo; m Promedio de los límites de especificaciones; $\Phi(\cdot)$ Distribución normal acumulada.

El primer índice, C_p, fue introducido como el cociente entre los límites de tolerancia de un producto y la variabilidad del proceso. Se han estudiado ampliamente sus propiedades,

bajo normalidad y bajo el supuesto de observaciones idénticamente distribuídas e independientes (Pearn et al., 1993). El índice se calcula como: $C_p = \frac{USL-LSL}{6\sigma}$. Con el supuesto de que $\mu = m = \frac{1}{2}(USL + LSL)$ y suponiendo distribución normal, la proporción esperada de productos NC es $2\Phi(-3C_P)$. Cuando $C_p = 1$, la proporción esperada de producto NC es 0.27%.

Si se cambia el concepto de proporción NC y sólo se toma en cuenta el valor directo del índice, C_p fallaría al identificar que el proceso está fuera de los límites de especificaciones. El índice C_{pk} fue introducido para dar al valor de μ influencia directa en el valor del C_p y tomar en cuenta que la media del proceso μ no se encuentra centrada en los límites de especificaciones, es decir, $(\mu \neq m)$. Este índice fue definido como: $C_{pk} = \min\left(\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}\right)$

La definición original de C_{pm} es: $C_{pm} = \frac{USL-LSL}{6\sqrt{\sigma^2+(\mu-\tau)^2}}$. El propósito de este índice es medir el grado de centrado del proceso con respecto al valor objetivo τ. La ubicación de τ con respecto a los límites de especificaciones no tiene un efecto directo en el valor de C_{pm}. En este índice los límites de especificaciones solamente son utilizados para reescalar la función de pérdida (desviaciones cuadradas), (Kusheler y Hurley, 1992). Las propiedades de este índice han sido estudiadas para normalidad por diversos autores (consultar en Pearn et al., 1992).

Los índices robustos fueron diseñados para no tener que preocuparse más de la forma de la distribución y poder seguir utilizando estos índices en ausencia de simetría. El índice robusto propuesto por Pearn, Kotz & Johnson (1992), C_θ garantiza la detección de aproximadamente un 1% de NC para cualquier distribución. En su artículo selecciona el valor de $\theta = 5.15$ para reemplazar al múltiplo 6 del denominador del índice (Pearn et al., 1992).

Clements, (1989) propuso un método en el cual se utilizan unas tablas especiales para construir C_p's basados en los Sistemas de Curvas de Pearson (PS). El método de construcción está basado en el supuesto de que la distribución del proceso pueda ser representada adecuadamente por un miembro de esta Familia. La idea esencial es reemplazar el múltiplo 6 en el denominador de C_p por un número, digamos θ, tal que $P\left(\mu - \frac{1}{2}\theta_P\sigma \leq X \leq \mu + \frac{1}{2}\theta_P\sigma\right) = 0.0027$. Para un valor dado de los coeficientes de sesgo y curtosis, las tablas de las curvas de Pearson (Clements 1989) proporcionarán los valores de θ_P tal que $P\left(X \leq \mu - \frac{1}{2}\theta_\ell\sigma\right) = 0.135\% = P\left(X \geq \mu + \frac{1}{2}\theta_u\sigma\right)$. Donde θ_ℓ y θ_u son los valores de los percentiles 0.135 y 99.865, respectivamente, se toma el valor de $\theta_p = \theta_u - \theta_\ell$ en vez del 6σ que aparece en el denominador de los índices originales. Para distribuciones normales se tiene que el valor $\theta_p = 6\sigma$, obteniendose el índice original.

2 Porcentaje de producto fuera de especificaciones

Retomando el enfoque original con el que fueron diseñados los primeros índices se presenta una generalización de la relación que existe entre el porcentaje de productos fuera de

especificaciones y el valor de los C_p's. Aunque una medida más sencilla de cálculo sería simplemente expresar esta probabilidad en términos de los parámetros de la distribución y los límites de especificaciones:

$$Pnc = F_X(LSL;\theta) + 1 - F_X(USL;\theta) \tag{1}$$

Sin embargo, se presenta esta generalización para visualizar el efecto que tiene el valor de los índices en el porcentaje de productos fuera de especificaciones.

Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuídas con función de distribución acumulada $F_X(x;\ \theta)$ donde el vector θ contiene a los parámetros de la distribución continua.

Para construir la generalización del concepto, se estandariza la variable para ver como queda en términos del supuesto sobre la media de la distribución que implica cada índice de capacidad. Se utiliza la variable aleatoria y que es la variable de calidad estandarizada y se utiliza la relación $F_Y(y) = F_X\left(E(X) + y\sqrt{V(X)};\theta\right)$, donde $E(X)$ es la media de la distribución y $V(X)$, su varianza, las cuales están definidas para cada distribución particular en términos de sus momentos.

El índice C_p hace referencia indirecta sobre la media del proceso al suponer que debe estar centrada en el intervalo de especificaciones *i.e.* $E(X) = \mu = m$ (Kotz y Johnson, 1983). Al realizar el desarrollo bajo este supuesto, el cálculo del Pnc para C_p queda determinado por:

$$Pnc_p = F_X(\mu - 3C_p\sigma;\theta) + 1 - F_X(\mu + 3C_p\sigma;\theta)$$

Para el caso de la distribución normal estándar el porcentaje de producto fuera de especificaciones queda determinado por $2\Phi(-3C_p)$ y cuando $C_p = 1$, este porcentaje es 0.27%

Para el caso de C_{pk} donde no se hace referencia acerca del valor de la media del proceso, el cálculo del porcentaje queda expresado como la función de probabilidad (1), y se puede reescribir en términos de C_p y C_{pk}

$$Pnc_{pk} = F_X(-3(2C_p - C_{pk});\theta) + 1 - F_X(3C_{pk};\theta)$$

Notar que para el caso de una distribución normal estandar cuando $C_p = 1$ este porcentaje ya no es igual a 0.27%.

De igual forma se generaliza una función para C_{pm} con el supuesto de que $EX = \tau + \delta$ con $\tau = m$.

$$Pnc_{pm} = F_X\left(\mu - \left(3C_p + \frac{\delta}{\sigma}\right)\sigma;\theta\right) + 1 - F_X\left(\mu + \left(3C_P + \frac{\delta}{\sigma}\right)\sigma;\theta\right)$$

donde $\delta = (\mu - \tau)$

De estas tres ecuaciones es claro que para cada valor de un índice, éste no representará el mismo porcentaje de producto fuera de especificaciones a menos que $\mu = \tau = m$ y además cuando $C_p = 1$, el valor de Pnc dependerá del valor de los parámetros de la distribución.

3 Metodología

Partiendo de la definición general de índice de capacidad el problema de estimar índices en situaciones asimétricas presenta dos enfoques. 1) Primero llevar a cabo la identificación de la distribución de la variable de calidad de interés y la estimación de los valores de sus parámetros para establecer la relación que existe entre el comportamiento del proceso y sus requerimientos, permitiendo reflejar esta función en términos de un índice de capacidad del proceso. 2) Diseñar un índice robusto a la forma de la distribución del proceso. Como siempre, se asume que el proceso estudiado está bajo control.

En términos de lo anteriormente expuesto se especifica el método de estimación de índices bajo no-normalidad utilizando el primer enfoque. Se llevó a cabo una simulación para establecer la precisión de los estimadores de los índices de capacidad cuando la distribución de la variable de interés presenta asimetría y el comportamiento de los estimadores de los índices para muestras pequeñas con los índices estimados de la forma usual y con el método propuesto por Pearn, Kotz y Johnson (1992). Para el método propuesto por Clements se propone una mejora utilizando estimadores de percentiles con el método de Máxima Verosimilitud, asimismo se realizaron pruebas para otras familias diferentes de las de Pearson.

Para la realización de las simulaciones se asume que los datos provienen de un proceso bajo control estadístico y que se ha identificado a qué familia de distribuciones pertenecen los datos, ya sea mediante exploración gráfica o algún método de bondad de ajuste. Se generaron 1000 muestras aleatorias de tamaño 10(10)50,70,100 para tres miembros de la distribución Lognormal con $\mu = 1$ y con los valores para σ presentados en la tabla 1. Para cada población Lognormal particular se utilizaron diferentes especificaciones, ver tabla 2. Para cada población estudiada sus especificaciones se obtienen los valores de los diferentes índices.

La estimación de los parámetros de la distribución ajustada para cada caso particular se realizó con métodos numéricos implementados por William Meeker [1] en SPLUS 4.5. Se fijaron los parámetros de cada distribución para cuatro poblaciones diferentes. Cada población presenta diferentes grados de sesgo y curtosis, ver tabla 1. Las especificaciones se establecieron de manera que se considera el caso en el que el valor objetivo está centrado en el intervalo de especificaciones, lo cual favorece al índice C_{pm}. Por completez del estudio se realizó el cálculo de los otros dos índices para observar su comportamiento en el caso de que se escogiera un índice inadecuado para la situación.

[1] Dr. William Q. Meeker, Jr Professor of Statistics Iowa State University, http:// www.public.iastate.edu/~wqmeeker

Tabla 1. Valores de los parámetros

	σ	*Sesgo*	*Curtosis*
1	0.2	1.516	7.345
2	0.6	3.466	30.083
3	1.4	10.584	450.133

Tabla 2. Valores de las especificaciones.

	lsl	*usl*	m
1	2.20	4	3.1
2	1.90	10	5.95
3	1.15	15	8.075

4 Conclusiones

En el presente artículo se describe una metodología de análisis que consiste en seleccionar el índice mas adecuado a la situación dependiendo de, los valores de τ, m y μ. El procedimiento comienza con la identificación de la distribución que describa mejor el comportamiento de la variable de calidad, la estimación de los parámetros de la distribución con el método de Máxima Verosimilitud y luego se procede a calcular los índices.

Se llevó a cabo una simulación para establecer la precisión de los estimadores de los índices de capacidad presentando una modificación para el cálculo de los índices propuestos por Clements.

Se utilizó a la distribución Lognormal para la ejemplificación de estos comportamientos. Los resultados obtenidos sugieren que el tamaño de muestra $n > 70$ provee las mejores estimaciones para los casos considerados en el estudio.

Se observa que el valor de la curtosis y el sesgo afecta el valor del cálculo de los índices bajo el modelo lognormal, el comportamiento general observado es que a mayor valor de sesgo y curtosis se presenta mayor sobreestimación en los índices así como, menor precisión. El índice C_{pm}, como era de esperarse, fue el que presentó la estimación mas precisa ($n > 70$) y como en todos los casos, su variabilidad aumenta cuando la curtosis es grande (450). Otras simulaciones realizadas para los modelos: Weibull y Gamma, las cuales aunque no fueron exhaustivas presentan los mismos comportamientos para todos los índices. Para los tamaños de muestra pequeños se observó que todos los índices presentan una sobreestimación así como mucha variabilidad. El índice que presentó mayor variabilidad para las tres poblaciones consideradas en el estudio fue el de Clements. Se observa que el índice C_p presentó un sesgo positivo el cual se incrementó a medida que se incrementaba el valor del sesgo y curtosis de la población lognormal. Para todas las poblaciones consideradas en este estudio se observa que el estimador con menor variación fue C_{pm}, indicando que es fundamental seleccionar al índice apropiado, ver figura 1 para una ilustración de los comportamientos descritos.

Se contruyeron intervalos de confianza asintóticos para los índices C_p, C_{pk} y C_{pm} bajo distribuciones no normales (Chang et al., 1990), se extendieron estos resultados para los índices robustos propuestos por Johnson, Kotz y Pearn (C_θ) y se estudió su comportamiento bajo simulaciones Monte Carlo, pero no se reportan los resultados en este trabajo, ver Landeros (1998). Es necesario realizar nuevas investigaciones para determinar como se

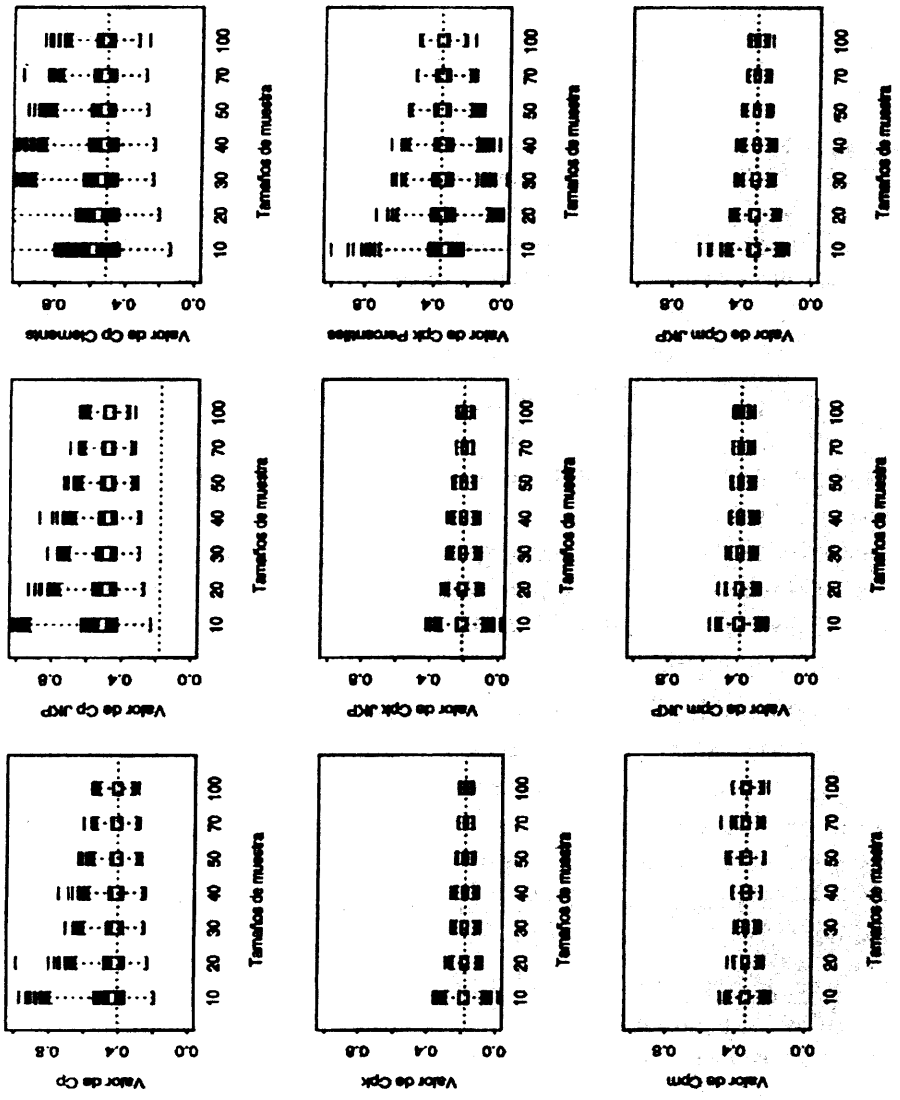


Figura 1: Resultados para los diferentes índices para la población Lognormal con $\mu = 1$ y $\sigma = 0.6$. Para cada índice se presenta el diagrama de caja, la línea blanca representa a la mediana. La línea punteada indica el valor poblacional de cada índice.

comportan los indices en el caso de que τ no se encuentra centrado en el intervalo de especificaciones. Es claro que el tamaño de muestra debe ser grande ($n > 70$) para garantizar los comportamientos deseables de los mismos, esto es, la reducción de sesgo y variabilidad así como, selecciónar el índice apropiado dependiendo de los valores de m, τ y μ.

Referencias

Chang, L.K., Xiong Z. y Zhang, D. (1990). On the asymptotic distributions of some process capability indices. *Commun. Statist. - Theory Meth.*, 19(1), 11-18.

Clements, J.A. (1989). Process Capability Calculations for Non-Normal Distributions. *Quality Progress.* 22, pp.95-100.

Kotz, S. y Johnson, N.L. (1983). *Process Capability Indices.* Chapmal and Hall, London.

Kushler, R.H., Hurley, P. (1992). Confidence Bounds for Capability Indices. *Journal of Quality Technology*, Vol. 24, No. 4, pp. 188-195.

Landeros, A. I. (1998). *Indices de Capacidad del Proceso bajo no normalidad.* Tesis del Instituto Tecnológico y de Estudios Superiores de Monterrey. Sin publicarse.

Pearn, W.L., Kotz, S. y Johnson, N.L.(1992). Distributional and Inferential Properties of Process Capability Indices. *Journal of Quality Technology,* Vol. 24, No. 4, pp 216-231.

Análisis de Capacidad del Proceso para Datos con Distribución No-Normal

Lorena López Losada

Facultad de Estadística e Informática, Universidad Veracruzana

1 Introducción

Los índices de capacidad son universalmente empleados en la relación de las empresas con los proveedores y con los clientes. Estos índices ayudan a enfatizar la necesidad de mejoras para reducir la variabilidad del proceso. También facilitan la comparación del desempeño de distintos proveedores o procesos y proporcionan una idea aproximada del porcentaje de artículos que no cumple con las especificaciones establecidas.

Existen diferentes procedimientos para monitorear la calidad del proceso de producción. Sin embargo, una vez que el proceso está bajo control estadístico surge la pregunta, "¿a qué nivel el proceso satisface los requerimientos u objetivos ingenieriles o administrativos?, o en términos más generales, la pregunta es, "¿qué tan capaz es el proceso en cuanto a producir artículos dentro de los límites de especificación?. La respuesta a esta pregunta requiere teoría estadística, pero el problema en sí mismo no es académico. El mejoramiento de calidad empresarial intenta estandarizar, y como evidencia de capacidad esto se ha convertido en un requerimiento para los proveedores y compañías que están reconsiderando el valor y la fiabilidad de únicamente resúmenes numéricos para el comportamiento de procesos complejos. Algunas organizaciones han introducido sus propios índices de capacidad, mientras otras todavía consideran restringido el uso de los índices descubriendo después que su desconsideración es un obstáculo para el mejoramiento.

Se presenta una descripción del método percentil que ha sido propuesto por Clements (1989), procedimiento que se basa en la obtención de percentiles para generar índices de capacidad en datos que tienen distribución no-normal. En la última sección se ilustra el método con un ejemplo de aplicación a un caso real en la industria.

2 Estudio de capacidad

El primer paso hacia un proceso de alta calidad es conducirlo bajo control estadístico. Díaz (1994) recomienda que antes de iniciar el control del proceso mediante algún gráfico de

control se haga un análisis exploratorio de la característica de calidad para determinar su naturaleza y poder elegir adecuadamente el tipo de gráfico. Ya que en muchas ocasiones, el conjunto de observaciones con que se cuenta para realizar un análisis de capacidad tiene una distribución sesgada y/o con observaciones extremas, lo que implica utilizar un método apropiado en la obtención de los índices de capacidad que en este caso sería para distribuciones no-normales.

2.1 Análisis inicial

Si el análisis de capacidad está basado en una distribución normal o en alguna otra distribución, una prueba de bondad de ajuste es útil para verificar el ajuste del modelo distribucional. Las pruebas estadísticas recomendadas para este propósito se basan en la función de distribución empírica como son: Kolmogorov-Smirnov, Lillifors, y la prueba de bondad de ajuste Chi-cuadrada.

La distribución ajustada también puede verificarse gráficamente utilizando los gráficos Q-plot y P-plot. Ambos gráficos pueden ser obtenidos tanto para la distribución normal como para una distribución no-normal. El Q-plot no es tan conocido como el P-plot en los procesos ingenieriles, pero estos ofrecen ventajas que ameritan su aplicación en análisis de capacidad. Chambers et al. (1983) proponen observar la distribución de la característica de calidad contra la distribución teórica ajustada a través de estos gráficos para la distribución estandarizada respectiva.

El Q-plot es un diagrama de dispersión de los valores observados de la característica de calidad contra los valores esperados (estandarizados), dando la distribución respectiva y el P-plot es un diagrama de la función de distribución acumulada observada contra la función de distribución teórica acumulada para los mismos valores. En estos diagramas de dispersión, primero se ordenan los datos en forma ascendente y si los valores observados caen sobre la línea recta, es decir, siguen la distribución teórica, entonces indica que hay un buen ajuste a los datos.

El método gráfico mas comúnmente utilizado para análisis de capacidad es el histograma. Sin embargo, muchos usuarios ignoran que la interpretación de un histograma depende de cómo están distribuidos los datos, donde el número de barras está determinado por un buen algoritmo. En todo estudio de capacidad de proceso se debe hacer el intento de conocer las fuentes de variabilidad en el proceso y una herramienta gráfica altamente efectiva para este propósito es el histograma, comparando la distribución de los datos con el modelo ajustado. El histograma acompañado además por el índice de capacidad, provee una clara visualización y una mejor comprensión del comportamiento del proceso, y además es fácilmente entendido por los operadores, ingenieros del proceso, y directivos.

Hahn y Shapiro (1967) detallan las curvas de Johnson quien describe un sistema de curvas que representa transformaciones de la curva normal estándar, y que es una alternativa al ajuste de curvas para calcular percentiles que quizá no son tan necesarias si el

objetivo sólo es estimar un índice de capacidad en datos que tienen distribución no-normal. También describe el sistema de distribuciones propuesto por Karl Pearson, que consiste de siete soluciones (de las doce originalmente enumeradas por Pearson) para una ecuación diferencial, la cual también se aproxima a un amplio rango de distribuciones de diferentes curvas.

2.2 Método percentil

Generalmente los requerimientos imponen un rango de valores aceptables que son el Límite de Especificación Superior (LES), el Límite de Especificación Inferior (LEI) y a la diferencia entre estos límites se le llama rango de especificación. El indicador más simple y conciso de capacidad de proceso es el índice C_P y está definido como la razón del rango de especificación al rango del proceso. El índice equivalente al C_P, es la razón de capacidad (C_r) que se calcula como $1/C_P$ es decir, el inverso de C_P.

La definición de los índices de capacidad para distribuciones no-normales ya ha sido desarrollada; Clements (1989) describe el procedimiento a detalle para el caso no-normal. Éste desarrollo se basa en el hecho de que para la distribución normal estándar la cantidad de 3σ es para la distancia de la media (mediana) al percentil superior 99.865 ($z_s = +3$) y la distancia del percentil inferior 0.135 ($z_i = -3$) a la media (mediana); y que tanto los límites $\pm 3\sigma$ como la media ($z_M = 0$) pueden ser remplazados por los valores correspondientes, dando los mismos percentiles 0.135 y 99.865 bajo la curva no-normal. De ésta manera, se puede estimar que el 99.73% aproximadamente de todos los productos están dentro de estos límites. Para calcular los índices de capacidad con percentiles se tienen las ecuaciones siguientes:

$C_p = (LES - LEI)/(P_S - P_I)$

$C_{pi} = (P_M - LEI)/(P_M - P_I)$ $\qquad$ $C_{pu} = (LES - P_M)/(P_S - P_M)$

$C_{pK} = Min(C_{pl}, C_{pu})$

donde

P_M representa el valor del percentil 50, $P_{0.5}$

P_I corresponde al valor del percentil inferior 0.135, $P_{0.00135}$

P_S corresponde al valor del percentil superior 99.865, $P_{0.99865}$

El comportamiento de las distribuciones continuas también puede resumirse suficientemente a través de los primeros cuatro momentos. Si al histograma de la característica de calidad se le ajusta una distribución que tenga la misma media (primer momento), varianza (segundo momento), asimetría (tercer momento) y curtosis (cuarto momento) de los datos observados, entonces se puede hacer una muy buena aproximación a la forma de la verdadera distribución; este procedimiento es más aplicable a distribuciones sesgadas. Una vez que se hace el ajuste de una distribución, se pueden calcular los percentiles esperados bajo la curva ajustada (estandarizada), y estimar la proporción de artículos producidos por el proceso que caen entre los límites de especificación.

El valor del índice C_p proporciona una idea de la variabilidad del proceso, pero siempre se deberá interpretar con precaución. Si $C_p > 1$, indica que el proceso es capaz de cumplir con las especificaciones; si $C_p < 1$ se dice que el proceso no es capaz. Lo ideal es que el C_p sea mayor que 1, ya que aquellos procesos con C_p alrededor de 1, deben ser vigilados, porque entre más descentramiento respecto del valor nominal pueden ocasionar un número elevado de productos defectuosos.

El índice C_p se puede considerar como una medida implantada en el programa de aseguramiento de calidad en una empresa; por ejemplo, en la década de los ochenta la industria japonesa adoptó un $C_p = 1.33$, mientras que otros establecieron $C_p = 2$ (Prat et. al, 1994).

3 Una aplicación

El estudio de capacidad del proceso se debe incluir como una parte esencial de un programa de mejora continua. Este tipo de estudio estadístico tiene un gran valor para diseñar programas de monitoreo y auditoría, para seleccionar y adquirir equipos, para estudios económicos, para establecer contratos con proveedores y clientes, pero sobre todo para tener un mayor conocimiento de la habilidad y capacidad de procesos y subprocesos de un sistema y organización, todo con el propósito de establecer metas realistas de mejoramiento.

Por esto último es que se pretendió instrumentar este tipo de estudios en una empresa, en la que se hizo el estudio considerando tres características de calidad para evaluar el producto final del proceso que se entrega por día a su cliente. Se obtuvieron acumulados por día para una de las características y promedios para las otras dos características.

En este trabajo sólo se reporta el análisis para una de las características y no se describe de manera específica por privacidad de la empresa. Esta característica de calidad es una variable que puede oscilar desde un límite inferior de 406 (LEI) a un límite superior de 418 (LES), no teniendo en este caso un ideal ya que esto es dependiente siempre de las condiciones que prevalezcan en un momento dado, en el sistema; como valor medio de este rango se tiene simplemente 412.

La información utilizada para realizar el análisis de la capacidad del proceso fue recopilada de registros periódicos de monitoreos de cada cuatro horas realizados en la empresa y en los puntos de salida utilizando equipo instalado para tal propósito; los datos que se utilizaron corresponden del 6 de enero al 16 de agosto de 1997.

El análisis se hizo a través del software Statistica, ya que es con el que cuenta la empresa y además contiene los procedimientos descritos en las secciones anteriores. En la primera fase se realizó un análisis exploratorio de la distribución de la característica de calidad y se identificaron valores atípicos del proceso, los cuales no se removieron porque no hubo justificación para eliminarlos de la serie de datos; se encontró que su distribución es sesgada. A través de la prueba de bondad de ajuste Chi-cuadrada se concluyó que los

Estadísticas Descriptivas

Estadísticas	N = 223 Valor
Media	411.531
Mediana	411.214
Percentil 25 (Q25)	410.357
Percentil 75 (Q75)	412.429
Valor Mínimo	407.857
Valor Máximo	421.000
Desviación Estándar	1.762
Varianza	3.104
Asimetría	1.325
Curtosis	3.533

datos tienen una distribución no-normal.

Basándose en las estadísticas descriptivas, la media del proceso es de 411.531, con un valor de 1.325 para la asimetría, lo que indica que los datos se concentran más hacia el lado izquierdo de la curva (por debajo de la media) y con el valor de 3.533 para la curtosis, que significa que la distribución es más "picuda" o elevada que la curva normal.

En el Q-plot se observa que con excepción de los puntos que caen en el extremo superior, los demás puntos caen sobre la línea ajustada.

Nuevamente se observa que los datos observados caen sobre la línea ajustada. El P-plot es consistente con el supuesto de que los datos tienen una distribución con media 411.531, varianza 3.104, asimetría 1.325 y curtosis 3.533.

El proceso tiende a producir 0.448% fuera de las especificaciones, lo cual es un número muy pequeño y representa a una sola observación. Sin embargo, el hecho de que el valor del C_p sea aproximadamente 1 implica que se debe tener cuidado con el proceso, ya que si por alguna causa se sigue descentrando, el porcentaje de defectuosos incrementará rápidamente.

En el histograma se puede observar que la distribución ajustada de la familia de las curvas de Pearson se aproxima razonablemente bien a los datos.

4 Conclusiones

El método percentil es tan sencillo de aplicar como el método bajo el supuesto de normalidad; funciona muy bien cuando se tienen valores extremos y además existe software idóneo para su aplicación. El software Statistica contiene los procedimientos para realizar análisis de capacidad en ambos casos: datos con distribución normal y no-normal.

Para este proceso, el índice de capacidad muestra la posibilidad de que el proceso puede

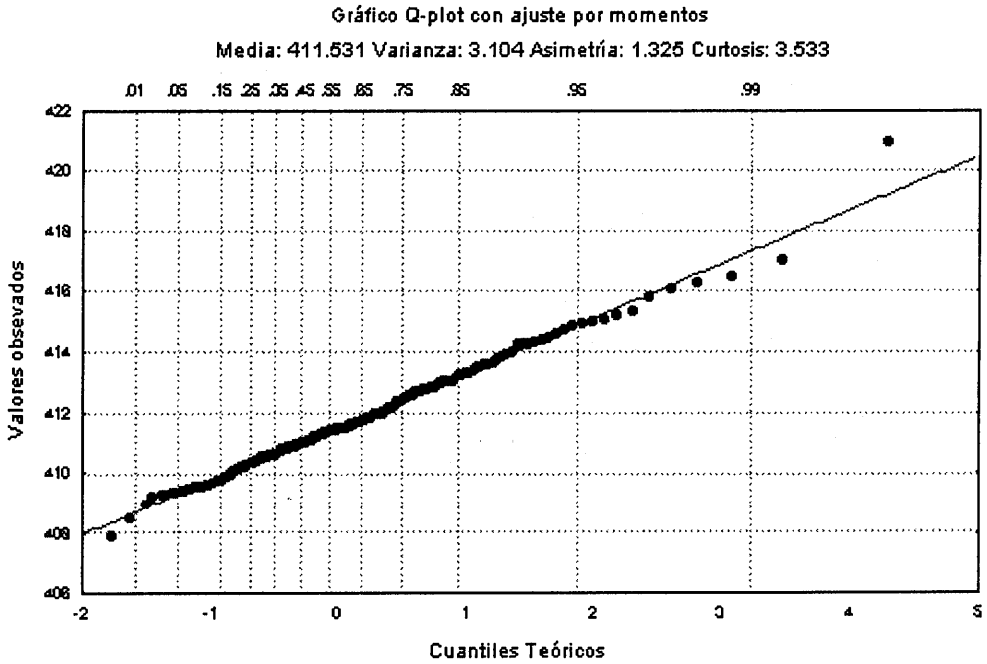


Gráfico Q-plot con ajuste por momentos
Media: 411.531 Varianza: 3.104 Asimetría: 1.325 Curtosis: 3.533
.01 .05 .15 .25 .35 .45 .55 .65 .75 .85 .95 .99
422
420
418
416
414
412
410
408
406
Valores observados
-2 -1 0 1 2 3 4 5
Cuantiles Teóricos

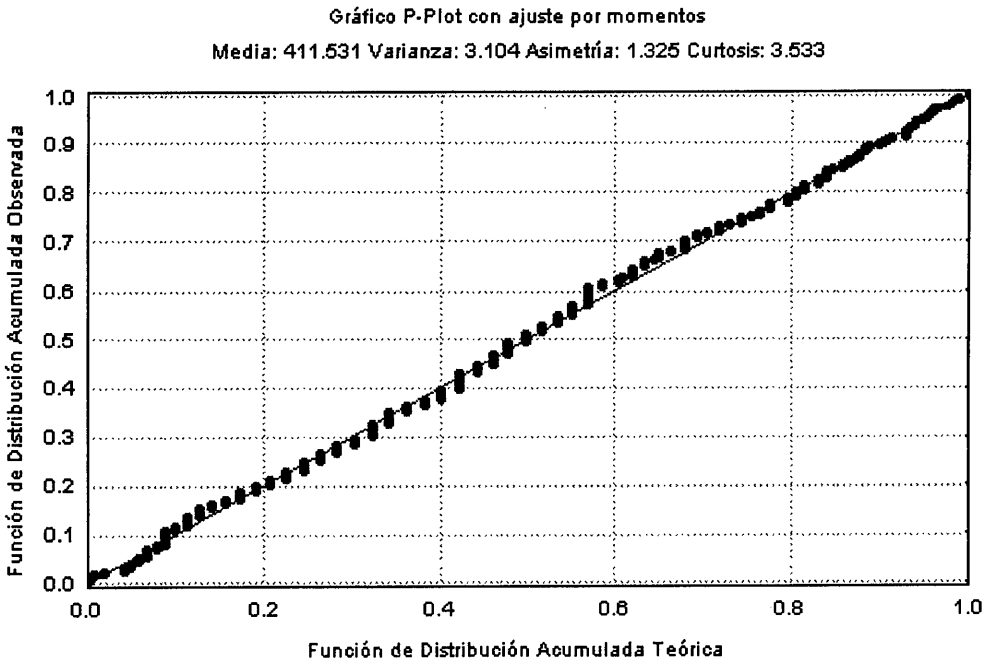
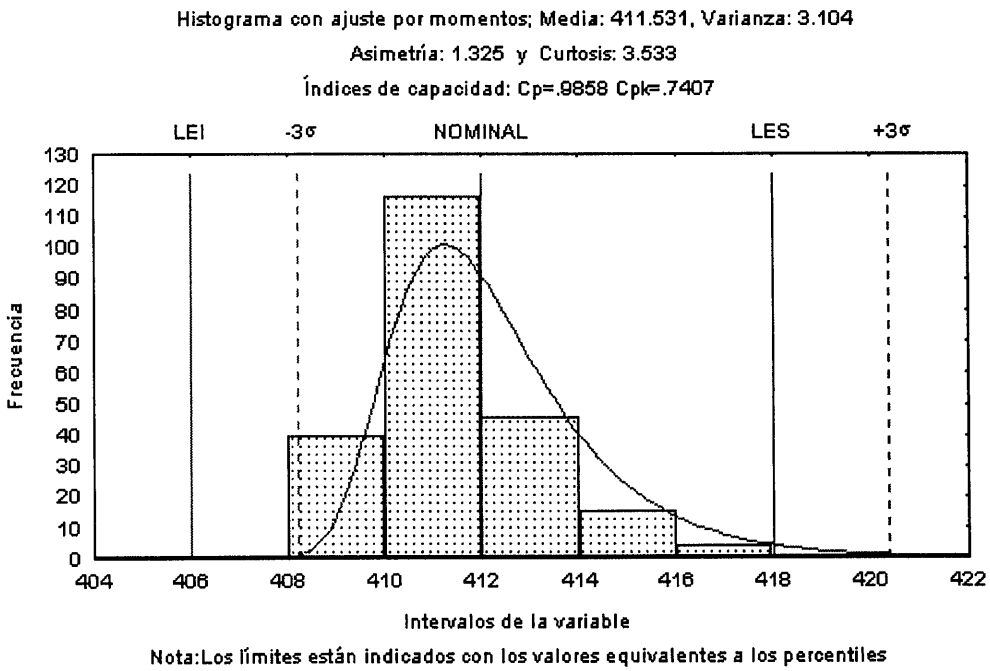


Gráfico P-Plot con ajuste por momentos
Media: 411.531 Varianza: 3.104 Asimetría: 1.325 Curtosis: 3.533
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Función de Distribución Acumulada Observada
0.0 0.2 0.4 0.6 0.8 1.0
Función de Distribución Acumulada Teórica



producir dentro de las tolerancias; es decir, el índice indica la capacidad potencial de que puede cumplir con las especificaciones.

Agradecimientos

Agradezco la valiosa colaboración del Ingeniero Pedro Alvarez Q.; así como la oportunidad de colaborar en la conducción de la asesoría estadística, al Dr. Mario Miguel Ojeda.

Referencias

Chambers, J. M., Cleveland, W. S., Kleiner, B., y Tukey, P. A. (1983). *Graphical Methods for Data Analysis.* Duxbury Press. Boston, MA.

Clemets, J. A. (1989). Process Capacility Calculations for Non-Normal Distributions, *Quality Progress* 22, pp. 95-100.

Díaz, T.C. (1994). Cartas de Control Robustas: Una Aplicación. *Memorias del IX Foro Nacional de Estadística.*

Hahn, G.J. y Shapiro, S.S. (1967). *Statistical models in engineering.* John Wiley. New York.

Prat, B. A., et al. (1994). *Métodos estadísticos: Control y Mejora de la Calidad.* Ediciones UPC. Barcelona, España.

StatSoft, Inc. (1998). *STATISTICA* Vol. IV: Industrial Statistics. USA.

Aplicación de Modelos Lineales Generalizados en Graduación y Tarificación

Evangelina Martínez y Alejandro Alegría
Departamento de Estadística, ITAM

1 Introducción

Los modelos lineales generalizados surgen como una generalización natural de los modelos lineales clásicos. En este trabajo se presentará una breve descripción de lo que es un modelo lineal generalizado, y se ejemplificará el uso de estos modelos en el problema de graduación y en la tarificación del seguro de automóviles. Los datos que se usaron en este análisis corresponden a la experiencia del sector asegurador en México para los años 1996 y 1997. Se considera que el uso de estos modelos presenta ventajas sobre los métodos tradicionales, y se presentan los resultados obtenidos. Los cálculos y estimaciones de los modelos se obtuvieron haciendo uso del paquete S-Plus.

2 Modelos lineales generalizados

Los modelos lineales generalizados (Nelder y Wedderburn, 1972) son una extensión de los modelos lineales clásicos y se definen de acuerdo a tres componentes. Por un lado, la distribución de las variables respuesta del modelo $Y_1,...,Y_n$ pertenece a la *familia exponencial.*

En segundo lugar, las variables explicativas del modelo, $X_1, X_2, ..., X_p$ ya sean covariables o factores, producen un predictor lineal $\eta^T=(\eta_1, \eta_2, ..., \eta_n)$, donde $\eta_i = \sum_{j=1}^{p} \beta_j X_{ij}$. $\boldsymbol{\beta}^T = (\beta_1, \beta_2, ..., \beta_p)$ es el vector de parámetros a estimar.

Finalmente, el predictor lineal puede expresarse como una función conocida del valor esperado de Y_i, es decir $g(\mu_i)=g[E(Y_i)]= \eta_i=\sum_{j=1}^{p} \beta_j X_{ij}$, donde $g(\cdot)$ es una función monótona y diferenciable llamada *función liga.* Cuando $g(\mu_i) = \theta_i = \eta_i$, la función recibe el nombre de *liga canónica.*

Una vez especificado el modelo de acuerdo a los tres componentes anteriores, los parámetros $\beta_1, \beta_2, ..., \beta_p$ se estiman a partir de los datos haciendo uso del método de máxima verosimilitud.

Para medir la bondad de ajuste del modelo resulta de utilidad la *devianza,* la cual mide la discrepancia entre el modelo ajustado y el modelo saturado con n parámetros. El hecho

de que la diferencia de devianzas tenga una distribución χ^2, resulta útil al momento de elegir la estructura del predictor lineal del modelo que mejor se adecue a los datos.

Los *residuos de devianza*, se definen en términos de la contribución individual de cada observación a la devianza del modelo. De acuerdo con Pierce y Schafer (1986), la distribución de los residuos de devianza es aproximadamente normal. Gráficas de residuos contra valores ajustados y covariables resultan de gran utilidad para detectar observaciones atípicas, para determinar si el modelo describe adecuadamente los efectos de las covariables en él consideradas e incluso para determinar si es necesaria la introducción de términos adicionales al predictor, o más covariables al modelo.

3 Aplicaciones en graduación

3.1 Descripción de los datos

·Formatos S.E.S.A. 3.1A y 3.2B (Asociación Mexicana de Instituciones de Seguros, 1996): total de pólizas y siniestros de la operación del seguro de vida individual del sector asegurador para 1996, respectivamente.

· Pólizas clasificadas de acuerdo a la covariable edad, x, de los 0 a los 100 años y a los siguientes factores:

Hábito(h): fumador o no fumador, y Examen(e): con examen médico o sin examen médico

Para cada factor, se introducen las siguientes variables indicadoras:

$$v = \begin{cases} 1, & \text{fumador} \\ 0, & \text{no fumador} \end{cases} \quad y \quad w = \begin{cases} 1, & \text{sin examen} \\ 0, & \text{con examen} \end{cases}$$

· 404 unidades o celdas especificadas como u={x,h,e}.
· Variable respuesta: A_u, total de muertes registradas en la celda u.
· R_u^i: expuestos iniciales al riesgo de muerte = total de pólizas por celda.
· R_u^c: expuestos centrales por celda = $R_u^i + (A_u/2)$ (Forfar, et al, 1988).
· Total de muertes registradas = 1,141.· Expuestos centrales totales = 719,311.5.
· Total de pólizas o expuestos iniciales = 718,741.

3.2 Graduación de q_u

Para la graduación de la probabilidad de muerte q_u, la variable respuesta del modelo, es decir las muertes registradas en la celda u, A_u, se asume que provienen de la distribución binomial de acuerdo a la cual $E(A_u) = R_u^i q_u$.

Después de ajustar modelos secuencialmente a partir de predictores anidados (Renshaw, 1991 y 1994a) para diferentes ligas, se eligió la liga canónica, es decir la función logit. De

acuerdo con la Tabla 1, la estructura del predictor junto con la fórmula de graduación para q_u están dadas por

$$\log\left(\frac{q_u}{1-q_u}\right) = \eta_u = \beta_0 + \beta_1 x_u + \beta_2 w_u + \beta_3 x_u^2 + \beta_4 x_u w_u, \qquad q_u = \frac{e^{\eta_u}}{1+e^{\eta_u}}.$$

En la Tabla 2 se presentan las estimaciones de los parámetros del modelo junto con el valor de la estadística t. De acuerdo con esta tabla, todos los valores son significativos y no hay evidencias para suponer sobredispersión en el modelo.

En las gráficas de los residuos de devianza con respecto a la edad y a los niveles de los factores, no se observó patrón alguno ni diferencias considerables, por lo tanto las gráficas son consistentes con la estructura del predictor adoptada, es decir, el predictor no requiere términos adicionales. En el histograma de residuos y la gráfica normal de los mismos no se observa una seria desviación al supuesto de normalidad aproximada de los residuos.

Tabla 1. Perfil de devianzas, liga logit

Variable	Devianza (D)	g.l.	Δg.l.	ΔD	$\chi^2_{\Delta gl,0.05}$	Se incluye
1	1,238.56	314	1			
+x	389.32	313	1	849.24*	3.8415	Sí
+e	340.26	312	1	49.058*	3.8415	Sí
$+x^2$	335.57	311	1	4.6908*	3.8415	Sí
+x : e	307.79	310	1	27.783*	3.8415	Sí

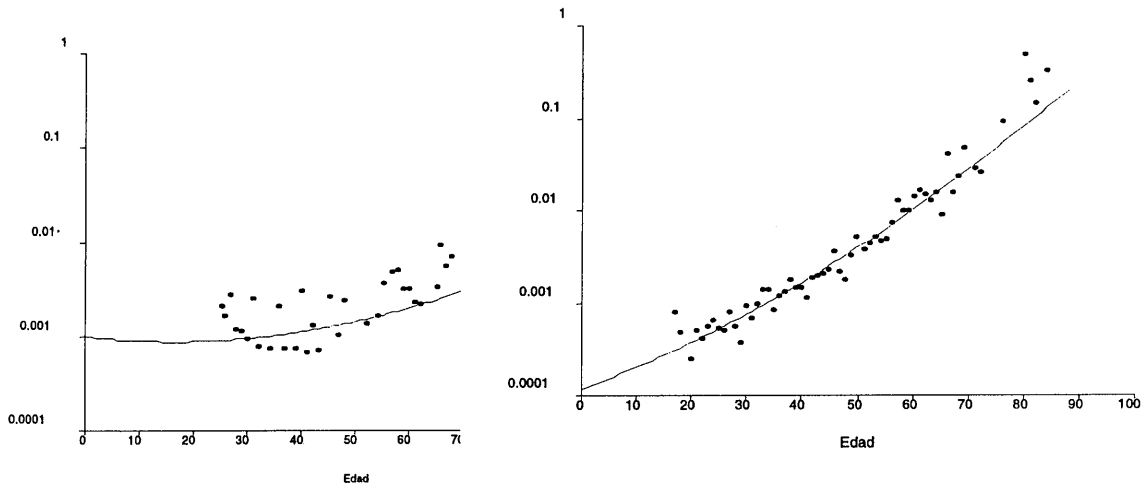
Las demás variables resultaron no significativas

Tabla 2. Parámetros estimados, graduación de q_u

$\widehat{\beta}_0 = -6.548046(-41.46499)$	$\widehat{\beta}_1 = 1.453632(2.497771)$
$\widehat{\beta}_2 = 1.020791(6.364416)$	$\widehat{\beta}_3 = 1.076227(2.706257)$
$\widehat{\beta}_4 = 3.134784(5.303061)$	$\widehat{\phi} = 0.992867$

Finalmente, por medio de simulaciones de variables Bin(,), se construyeron bandas al 95% dentro de las que se esperaría observar a los residuos del modelo ajustado, si el supuesto distribucional de la variable respuesta es razonable (Everitt, 1994). La gráfica obtenida mostró evidencia en contra de dicho supuesto.

En las Figuras se presentan las probabilidades de muerte observadas y ajustadas para las celdas con examen y sin examen, respectivamente.



3.3 Graduación de μ_u

Para la graduación de la fuerza de mortalidad se supone que la variable respuesta del modelo A_u se distribuye Poisson tal que $E(A_u) = R_u^c \mu_u = m_u$.

Se eligió a la función logaritmo como liga del modelo pues es la liga canónica. La estructura del predictor lineal y la fórmula de graduación están dadas por

$$\eta_u = \ln(R_u^c) + \beta_0 + \beta_1 x_u + \beta_2 w_u + \beta_3 x_u^2 + \beta_4 x_u w_u, \quad \mu_u = \exp\{\eta_u - \ln(R_u^c)\}.$$

Los valores observados y ajustados de la fuerza de mortalidad (en escala logarítmica) presentan un comportamisnto similar al de las Figuras.

4 Aplicaciones en tarificación en el seguro de autos

4.1 Descripción de los datos

· Pólizas emitidas entre 1996 y principios de 1998, datos proporcionados por una Institución de Seguros: total de pólizas (e_u), las reclamaciones reportadas (N_u) y el monto promedio pagado por la aseguradora (Z_u).

· Pólizas clasificados de acuerdo a la edad del asegurado, x (18.5 a 73 años) y a los siguientes factores:

Cobertura(c): i=1, Daños Materiales (DM); i=2, Responsabilidad Civil (RC); i=3, Robo Total (RT)

Marca (m): j=1, marca A; j=2, marca B; j=3, marca C; j=4, marca D; j=5, marca E
Sexo (s): k=1, femenino (F); k=2, masculino (M)
· Total de celdas especificadas de acuerdo a la edad y a los niveles de los factores =300
· Total de pólizas = 1592.· Total de reclamaciones = 761.
· Monto total superior a los $19,000,000.00.

4.2 Modelo para la frecuencia de reclamaciones

La variable respuesta del modelo N_u, el número de reclamaciones provenientes de la celda u, proviene de una distribución Poisson de acuerdo a la cual $E(N_u) = V(N_u) = m_u = e_u\lambda_u$, donde e_u son las unidades expuestas en la celda u y λ_u es el número esperado de reclamaciones dentro de la celda u (Renshaw, 1994b).

Como función liga se eligió al logaritmo, liga canónica para la distribución Poisson, es decir, $\eta_u = \ln(e_u\lambda_u) = \ln(e_u) + \ln(\lambda_u)$. El término $\ln(e_u)$ en la expresión anterior, denominado *offset*, forma parte del predictor lineal y es constante para cada celda.

Tras ajustar modelos secuencialmente con predictores anidados, el modelo elegido queda especificado como sigue:

$$\eta_u = \ln(e_u) + \beta_0 + \beta_1 a_2 + \beta_2 a_3,$$

donde a_2=1 para RC y a_2=0 e.o.c.; a_3=1 para RT y a_3=0 e.o.c.

Los valores estimados de los parámetros y sus correspondientes valores de la estadística t se presentan en la Tabla 3, de acuerdo a la cual todas las estimaciones son significativas.

Tabla 3. Parámetros estimados.

$\widehat{\beta}_0 = -0.93222(-18.89926)$	$\widehat{\beta}_2 = 0.75256(8.54735)$
$\widehat{\beta}_1 = -0.19190(-1.90176)$	$\widehat{\phi} = 0.90017$

La gráfica de los residuos de devianza contra el factor cobertura y la gráfica normal de dichos residuos resultaron consistentes con la estructura del predictor y el supuesto de normalidad de residuos, respectivamente. La gráfica normal de los residuos con bandas al 95% no apoya el supuesto distribucional del modelo.

4.3 Modelo para el monto de reclamaciones

La variable respuesta del modelo Z_u, el monto promedio de reclamaciones provenientes de la celda u, se asume que proviene de una distribución Gamma sobredispersa de acuerdo a la cual $E(Z_u) = w_u$ y $Var(Z_u) = \sigma^2 w_u^2/n_u$.

Se consideraron diferentes funciones liga, eligiéndose finalmente la función recíproco (liga canónica) y la siguiente estructura para el predictor (Tabla 4):

$$w_u^{-1} = \beta_0 + \theta_i a_i + \psi_j b_j + \delta_{ij} a_i b_j + \varepsilon_{ij} a_i b_j x_u, \quad i = 2,3, \quad j = 2,3,4,5$$

donde $b_2 = \langle_{0,\text{ e.o.c.}}^{1,\ B}$, $b_3 = \langle_{0,\text{ e.o.c.}}^{1,\ C}$, $b_4 = \langle_{0,\text{ e.o.c.}}^{1,\ D}$, $b_5 = \langle_{0,\text{ e.o.c.}}^{1,\ E}$

Tabla 4. Perfil de devianzas, liga recíproco

Variable	Devianza (D)	g.l.	Δg.l.	ΔD	$\chi^2_{\Delta gl,0.05}$	Se incluye
1	1005.6100	219				
+c	334.2545	217	2	671.3555*	5.9915	Sí
+m	309.9769	213	4	24.2776*	9.4877	Sí
+c:m	286.9016	205	8	23.0753*	15.5073	Sí
+c:m:x	249.9634	190	15	36.9382*	24.9958	Sí

Las demás variables resultaron no significativas

En el análisis de los residuos de devianza, no se observaron patrones ni diferencias significativas en los residuos con respecto a las variables explicativas del modelo, por lo que son consistentes con la estructura del predictor adoptada. La gráfica normal de los residuos resultó ser congruente con el supuesto de normalidad de los mismos, mientras que la gráfica con bandas al 95% no mostró evidencia en contra del supuesto distribucional del modelo.

Una vez estimada la frecuencia y el monto promedio de reclamaciones, la prima de riesgo para la celda u, P_u, está dada por $P_u = \widehat{\lambda}_u \widehat{w}_u$, donde $\widehat{\lambda}_u$ es la frecuencia esperada ajustada de reclamaciones y $\widehat{w}_u$ es el monto esperado ajustado de reclamaciones provenientes de la celda u.

5 Conclusiones

La graduación de la mortalidad de la experiencia de un grupo de personas, así como la tarificación en el seguro de automóviles por medio de los modelos aquí presentados, presenta ventajas frente a otros métodos, tales como la aplicación de los resultados inferenciales de dichos modelos a las estimaciones obtenidas, y la inclusión de diferentes variables explicativas, ya sea factores o covariables, a través de un predictor lineal.

Pasando a consideraciones más específicas acerca de los modelos ajustados en este trabajo, cabe comentar que a pesar de que es usual que el factor hábito (fumador o no fumador) esté considerado en modelos de graduación como los aquí presentados, el efecto de dicho factor no resultó significativo de acuerdo a los datos analizados. Por otro lado, el problema que ocasiona tener muchas celdas sin muertes registradas, podría resolverse agrupando las edades por quinquenios.

En los modelos ajustados para la tarificación del seguro de automóviles, se presentaron también algunos problemas. Por un lado, en el modelo para la frecuencia de reclamaciones

sólo se consideró el efecto del factor cobertura y la gráfica normal de los residuos con bandas no resultó como se esperaba, mientras que no todos los parámetros estimados en el modelo de los montos resultaron significativos.

Es posible que los problemas anteriores se deban en parte a la estructura y a la calidad de los datos analizados, ya que para implementar los modelos aquí presentados es importante contar con bases de datos amplias y confiables.

Es importante señalar que existen otros temas actuariales en los que se han aplicado satisfactoriamente los modelos lineales generalizados, entre los que se encuentran distribuciones de pérdida y reservas para siniestros ocurridos pero no reportados (Haberman y Renshaw, 1996).

Quedan aún muchas cosas por hacer, entre las que se encuentran la implementación de estos modelos desde el punto de vista Bayesiano y la posibilidad de modelar conjuntamente la frecuencia y el monto de reclamaciones en el seguro de automóviles.

Referencias

Asociación Mexicana de Instituciones de Seguros, A.C. (1996) *Seguro de Vida, Estadística.*

Everitt, B.S. (1994) *A Handbook of Statistical Analyses Using S-PLUS*, Chapman.

Forfar, D. O., McCutcheon, J. J. y Wilkie, A. D. (1988) On Graduation by Mathematical Formula, *Journal of the Institute of Actuaries*, **115**, 1-135.

Haberman, S. y Renshaw, A. E. (1996) Generalized Linear Models and Actuarial Science, *The Statistician*, **45**, 407-436.

Nelder, J. A. y Wedderburn, R. W. M. (1972) Generalized Linear Models, *Journal of the Royal Statistical Society*, A, **135**, 370-384.

Pierce, D. A. y Schafer, D.W. (1986). Residuals in Generalized Linear Models, *Journal of the American Statistical Association*, **81**, 977-986.

Renshaw, A. E. (1991) Actuarial Graduation Practice and Generalized Linear and Non-linear Models, *Journal of the Institute of Actuaries*, **118**, 295-312.

Renshaw, A. E. (1994a) A Comparison Between the Mortality of Smoking and Non-smoking Assured Lives in the U.K., *Journal of the Institute of Actuaries*, **121**, 561-571.

Renshaw, A. E. (1994b) Modelling the Claims Process in the Presence of Covariates, *ASTIN Bulletin*, **24**, 265-285.

Un Modelo de Interdependencia entre Respuestas Múltiples Aplicado a la Natación de Alto Rendimiento

Andrzej Matuszewski y Guillermo Bali

Acad. Polaca de Ciencias, Polonia *ITESM, Campus Cd. de México*

1 Introducción

En cualquier tipo de proceso donde intervienen factores humanos es de suma importancia disminuir la incertidumbre. La natación de alto rendimiento no constituye un caso aparte. Por esto es relevante relacionar las variables de entrenamiento y éxito dentro de este deporte, y establecer nuevos parámetros de medición que permitan un mejor aprovechamiento de los recursos humanos y económicos que hoy demanda la práctica deportiva.

En esta investigación se utilizan variables potenciales. Por estas variables entendemos preguntas en que el encuestado pueda escoger más de una respuesta entre k posibles o no está de acuerdo con ninguna. Desde cierto punto de vista esta pregunta es considerada una multifunción. Dichas multifunciones se definen a partir de los estilos en que los nadadores tienen éxito deportivo y los tipos de entrenamiento que realizan. La relación entre dichas variables potenciales se formaliza mediante el Modelo de Rasch. Desde los años 60 y después de varias aplicaciones a la psicología el modelo de Rasch confirmó su validez para el análisis de datos con un enfoque subjetivo (un amplio resumen y práctica de este modelo está en Fisher y Molenaar, (1995).

Nuestro modelo para las respuestas, no sólo incluye correlación, sino también toma en cuenta la actitud del encuestado a través de lo que hemos denominado parámetro de "potencia" deportiva. De aquí que se propongan dos formas de cálculo a partir del modelo general. Además se introduce un concepto de independencia personal que permite establecer mediante un teorema, dos condiciones equivalentes para los parámetros y funciones que intervienen en el modelo.

2 Motivación

Definamos primero las multifunciones y los eventos necesarios para introducir el modelo de los parámetros a medir.

Sea i el índice que indentifica un nadador con $i = 1, 2, ..., S$ y sea X una variable de

tipo 0-1 que representa el estilo de natación en que el nadador tiene éxitos deportivos o no. Entonces podemos definir una X-multifunción de $\{1,2,...,I\} \longrightarrow \{X_1, X_2, ..., X_S\}$, donde para toda $i < 1$, $s < S$

$$\begin{aligned} X_{its} &= 1, \text{si el nadador } i \text{ tiene éxito en el estilo } s, \\ X_{is} &= 0, \text{en otro caso.} \end{aligned}$$

Por ejemplo el nadador i puede tener éxito en 100 metros mariposa o en 200 metros libres, 100 metros pecho , etc..

Si Y es una variable de tipo 0-1 que representa el tipo de entrenamiento que utiliza o no el nadador i, podemos de igual manera definir una Y-multifunción de $\{1,2,...,I\} \longrightarrow \{Y_1, Y_2, ..., Y_T\}$, donde para toda $i < 1$, $t < T$

$$\begin{aligned} Y_{it} &= 1, \text{si el nadador } i \text{ práctica el tipo de entrenamiento } t, \\ Y_{it} &= 0, \text{en otro caso.} \end{aligned}$$

Por ejemplo el nadador i utiliza el tipo de entrenamiento aeróbico anaeróbico o velocidad pura, etc..

A partir de aquí utilizaremos los índices i, s, t con $i = 1,2,...,I$, $s = 1,2,...,S$ y $t = 1,2,...,T$ dados en la definición de las multifunciones.

Definamos ahora los siguientes eventos:

$E^+_{ist} = \{$ el evento que representa que el nadador i tiene éxitos en el estilo s y practica el entrenamiento $t\}$,

$E^-_{ist} = \{$ el evento que representa que el nadador i tiene éxitos en el estilo s, pero no practica el entrenamiento t,o viceversa$\}$, y

$E^0_{ist} = \{$ el evento que representa que el nadador i no tiene éxitos en el estilo s y no practica el entrenamiento $t\}$.

Para cada terna i, s, t, se cumple entonces que:

$$\begin{aligned} E^+_{ist} &= \{X_{is} = 1, Y_{it} = 1\} \\ E^-_{ist} &= \{X_{is} = 1, Y_{it} = 0\} \cup \{X_{is} = 0, Y_{it} = 1\} \\ E^0_{ist} &= \{X_{is} = 0, Y_{it} = 0\} \end{aligned}$$

El análisis que proponemos supone dos formas posibles de los datos. En ambos casos la fuente de información es una encuesta. Se supone que el número de personas encuestadas es I. Primero asumimos variables del tipo 0-1. Para aplicar el modelo debemos tener dos grupos de esas variables o sea dos dicotomías múltiples. En cierto sentido estos grupos son pequeñas pruebas. Pero los items de estas pruebas y esa es nuestra intención no tienen alta correlación dentro de cada una de ellas. Aparte tratamos de modelar la situación cuando ciertas variables del primer grupo tienen una correlación "personalizada" con las variables del segundo grupo.

La segunda forma equivalente de datos son respuestas para dos preguntas. Cada pregunta tiene un grupo de posibles respuestas. El encuestado puede seleccionar cualquier respuesta y esto es lo mismo que darle valor de uno a las variables de tipo 0-1 del primer grupo de datos que corresponden a esta respuesta. El encuestado puede seleccionar cualquier configuración de respuestas o decidir no escoger ninguna respuesta. Esta última opción no significa la falta de datos.

3 Modelo

Si introducimos el modelo de Rasch para los eventos E^+, E^- y E^0 entonces se cumple:

$$\ln p(E^+_{ist}) = (\theta^+_i + \beta^+_{st}) - C_{ist} \ (1)$$
$$\ln p(E^-_{ist}) = (\theta^-_i + \beta^-_{st}) - C_{ist} \ (2)$$

donde C_{ist} es una constante que depende de i, s, t y se cumple, para los eventos E^+, E^-y E^0que son mutuamente excluyentes dos a dos, que para toda i, s y t:

$$p(E^+_{ist}) + p(E^-_{ist}) + p(E^0_{ist}) = 1.$$

Las ecuaciones (1) y (2) son modelos de Rasch. Las consecuencias reales del hecho de que el modelo esté confirmado estadísticamente Fischer y Molenaar, (1995) son las siguientes:

a) Un θ^+_i grande significa que el nadador i tiene más éxitos y práctica más tipos de entrenamiento que otros nadadores en la muestra que tenemos. Por eso decimos de alguna manera que tiene gran "potencia deportiva".

b) Un θ^-_i grande en el modelo de nadadores (este debe ser mayor que cero) significa que el nadador i es muy irregular.

Lo anterior significa el nadador i práctica muchos estilos con éxitos, pero tiene muy pocos tipos de entrenamientos o viceversa. Para la primera alternativa podemos poner de ejemplo, el caso en que el nadador tiene éxitos en

cuatro estilos, pero otros tienen en tres, dos o un estilo, y práctica uno o dos tipos de entrenamiento, mientras los otros dos o más. Lo segundo significa que el nadador efectua varios tipos de entrenamiento y tiene éxito en un solo estilo o quizas en más, pero por debajo del número de éxitos en distintos estilos de otros nadadores.

c) $\beta_{st}^{-} > 0$ significa correlación negativa entre X_s y Y_t.

d) En la metodología clásica si uno quiere estimar la correlación entre dos variables del tipo 0-1: X_s y Y_t entonces

$$p(X = 0, Y = 0)$$

influye en forma monotónica a esta correlación. En el modelo que hemos considerado no hay ese tipo de influencias.

Ahora podemos plantear dos tipos de modelos. Primeramente, considerando las restricciones para los parámetros β

$$\sum_{s,t} \beta_{st}^{+} = 0 \text{ y}$$
$$\sum_{s,t} \beta_{st}^{-} = 0.$$

En este caso no hay restricciones para las θ, porque no nos interesan en el modelo de correlación.

Para el segundo modelo, que correspondería al dual del primero, se cumplen las restricciones para los parámetros θ

$$\sum_{i} \theta_{i}^{+} = 0 \text{ y}$$
$$\sum_{i} \theta_{i}^{-} = 0$$

En este caso no hay restricciones para las β, porque no nos interesan en el modelo de potencias.

4 Teorema

Los eventos E^{+}, E^{-}y E^{0} están caracterizados por ternas de números i, s, t y la distribución de probabilidad $p_{ist}(\cdot)$. Si representamos todas las posibles combinaciones de un nadador i por $E(s,t)$ entonces podemos introducir el concepto de independencia personal como:

$p\{$ Persona i ha escogido

$E(s=1,t=1) \wedge E(s=1,t=2) \wedge ...E(s=1,t=T) \wedge E(s=2,t=1) \wedge ...E(s=2,t=T) \wedge ...E(s=S,t=1) \wedge ...E(s=S,t=T)\} = p_{i11}(E(s=1,t=1)) \cdot p_{i12}(E(s=1,t=2))... \cdot p_{iST}(E(s=S,t=T))$ para cada persona i que tenemos en la muestra.

Ahora definamos dos condiciones equivalentes en las que se tiene como supuesto este tipo de independencia.

Condición 1

Sea

$$lnp_{ist}(E^+) = (\theta_i^+ + \beta_{st}^+) - C_{ist}$$
$$lnp_{ist}(E^-) = (\theta_i^- + \beta_{st}^-) - C_{ist}$$

con

$$p_{ist}(E^+) + p_{ist}(E^-) + p_{ist}(E^0) = 1$$

donde C_{ist} es una constante que depende de los parámetros θ de potencia y β de correlación y

$$lnp_{ist}(E^+) = f(\theta_i^+, \beta_{st}^+)$$
$$lnp_{ist}(E^-) = g(\theta_i^-, \beta_{st}^-)$$

son funciones de los parámetros $\theta_i^+, \theta_i^-, \beta_{st}^+, \beta_{st}^-$ respectivamente.

Condición 2

Considerando las restricciones

$$\sum_{s,t} \beta_{st}^+ = 0 \text{ y}$$
$$\sum_{s,t} \beta_{st}^- = 0$$

entonces existe una estadística suficiente y minimal (T^+, T^-) con

$T^+ = \sum_{s=1}^{S} \sum_{t=1}^{T} \chi(E(s,t) \text{ de persona } i = E^+)$ y

$T^- = \sum_{s=1}^{S} \sum_{t=1}^{T} \chi(E(s,t) \text{ de persona } i = E^-)$

donde χ es la función característica, y los estadísticos están definidos para la matriz del los $\theta_1^+, \theta_{2,...}^+ \theta_I^+, \theta_1^-, \theta_2^-, ...\theta_I^-$.

La distribución de (T^+, T^-) no depende de β_{st}^+ y β_{st}^-, es decir de la matriz de los betas.

Teorema: *Condición 1* $\Longleftrightarrow$ *Condición 2.*

La demostración es una consecuencia directa de Fischer y Molenaar, (1995).

A partir de los estadísticos mínimales y suficientes del teorema se desarrollaran los algoritmos de cálculo para la correlación entre .

Referencias

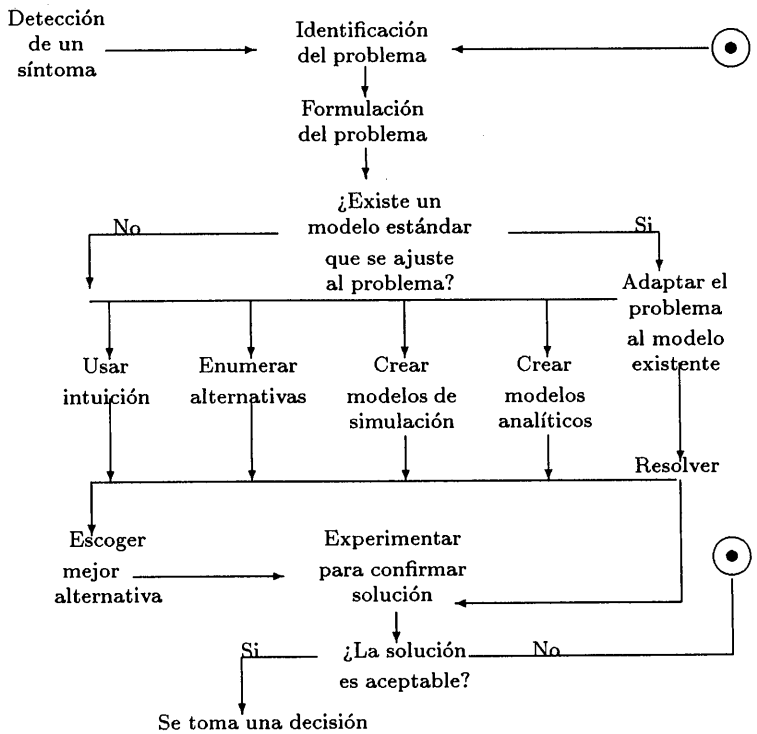
Fischer G. H. y Molenaar I. W. (eds) (1995), *Rasch models foundations, recent developments and aplications*, Springer-Verlag, New York.

Modelo de Simulación Digital para Estudiar el Fenómeno de la Desnutrición

Alicia del Rosario Nava Cardona y Blanca Rosa Pérez Salvador
DEPFI, UNAM *UAM-Iztapalapa, México*

1 Introducción

Un modelo de simulación permite representar la dinámica del fenómeno (su historia, estado actual y la proyección hacia el futuro), a bajos costos y sin alterar los sistemas reales. Con un modelo de simulación se pueden ensayar diferentes hipótesis, teniendo resultados de inmediato. En el siguiente diagrama se ubica la posición de la formulación del modelo dentro del procedimiento de solución de un problema.



La desnutrición, desde sus causas hasta sus repercusiones, es un fenómeno complejo muy difícil de analizar y de representar con un modelo analítico, por eso se propone un modelo

de simulación, para sondear algunas de sus características en función de un conjunto de variables explicativas.

En este trabajo se propone estudiar la desnutrición utilizando un modelo de simulación digital, o simulación asistido por una computadora.

2 El Modelo de simulación

La simulación dígital permite reproducir y estudiar fenómenos complejos, como el de la desnutrición, cuyos modelos son difíciles de analizar en su totalidad o en alguna de sus partes. Existen diversos paquetes diseñados para efectuar simulaciones digitales como son: DINAMO, POWERSIM Y STELLA. Los resultados presentados en este trabajo se obtuvieron con STELLA.

Para realizar la simulación, primeramente se definen los elementos que interactúan en el problema:

Se considera como **desnutrición** al déficit de nutrientes que limita el desarrollo biológico de una persona y su capacidad para el trabajo además deteriora su salud. El foco de atención del problema es el tamaño de la población desnutrida en el tiempo t, $X_1(t)$.

Un **individuo vulnerable** es aquél que no consume los alimentos necesarios, y por lo tanto no tiene una nutrición adecuada. Tiene aun un desarrollo aceptable, pero está en riesgo de desnutrición. El tamaño de la población vulnerable se representa por $X_2(t)$.

Un **individuo enfermo** es aquél que ya cuenta con una enfermedad (cualquier cuadro clínico: tos, fiebre, amibiasis, etc.). El tamaño de la población enferma se representa por $X_3(t)$.

Un **individuo nutrido** es aquel que cuenta con los nutrientes necesarios, su desarrollo biológico es adecuado. El tamaño de la población nutrida se representa por $X_4(t)$.

Se considera como el **nivel salarial** al ingreso total al núcleo familiar, menos gastos distintos a la alimentación. El nivel salarial se representa por $X_5(t)$.

Los **servicios públicos** se conforman por la infraestructura urbana con que se cuenta, como es: agua potable, drenaje, etc. La cantidad en los servicios públicos se representa por $X_6(t)$.

A. C. es el trabajo realizado por asociaciones civiles para combatir la desnutrición. El monto e impacto de este trabajo se representa por $X_7(t)$

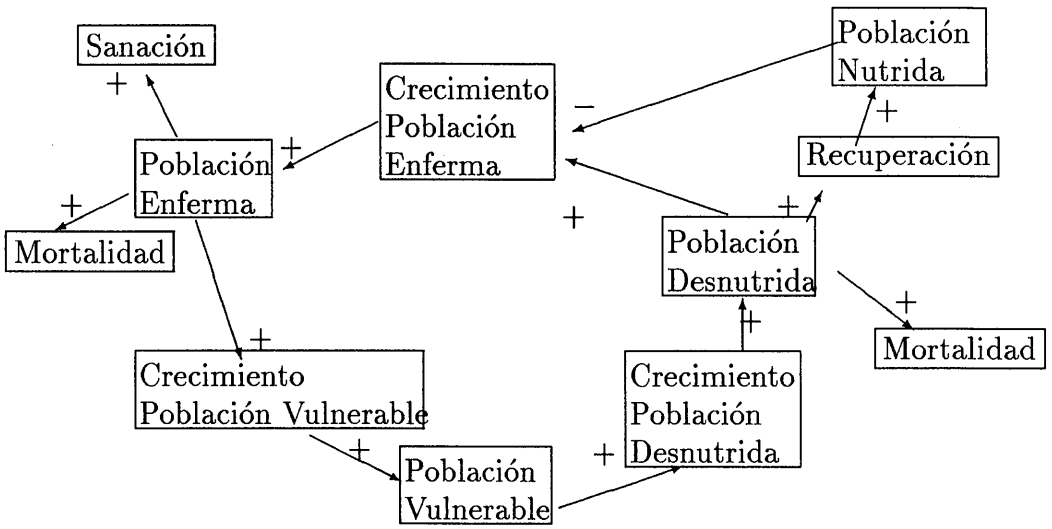
El problema se analizó en dos etapas, en la primera se consideró el círculo vicioso de la desnutrición. En este caso sólo intervienen las variables X_1, X_2, X_3 y X_4.

En la segunda etapa se introdujeron las otras variables. Éstas pueden llevar a la solución del problema y por lo tanto representa un círculo virtuoso.

2.1 El círculo vicioso

En él se describe la asociación entre la población desnutrida, la población enferma, la población vulnerable y la poblacion nutrida: a mayor porcentaje de población desnutrida se incrementa la tasa de crecimiento de la población enferma. La población enferma presenta falta de apetito, incluso puede ser que si no se les alimenta por vía intravenosa se convierte en población vulnerable que a su vez es susceptible de padecer desnutrición, el incremento de la población vulnerable provoca incremento en la tasa de población desnutrida. En cada una de estas interacciones se puede considerar que hay una probabilidad de pasar de una a otra población, con esta probabilidad se pueden generar los datos.

Las hipótesis consideradas en este modelo se esquematizan en el siguiente diagrama:



Los signos de + y de − indican cuando un elemento provoca un aumento o una disminución en otro elemento. Estas relaciones se proporcionan al paquete, quien las interpreta como un conjunto de ecuaciones recurrentes. Las ecuaciones correspondientes con el diagrama anterior son:

- $X_1(t) = X_1(t - dt) + (A_1 - B_1 - C_1)dt$ en donde:

  $A_1 = X_3(t)+$ índice de la desnutrición,

  $B_1 =$ índice de mortalidad de desnutridos y

  $C_1 =$ índice de recuperación.

- $X_3(t) = X_3(t - dt) + (A_3 - B_3 - C_3)dt$ en donde

  $A_3 = .02X_1 + .02X_2 - .2X_4+$ índice del crecimimiento de la población enferma,

  $B_3 =$ índice de sanación y

  $C_3 =$ índice de mortalidad por enfermedad.

- $X_4(t) = X_4(t - dt) + A_4dt$ en donde

  $A_4 =$ índice de recuperación.

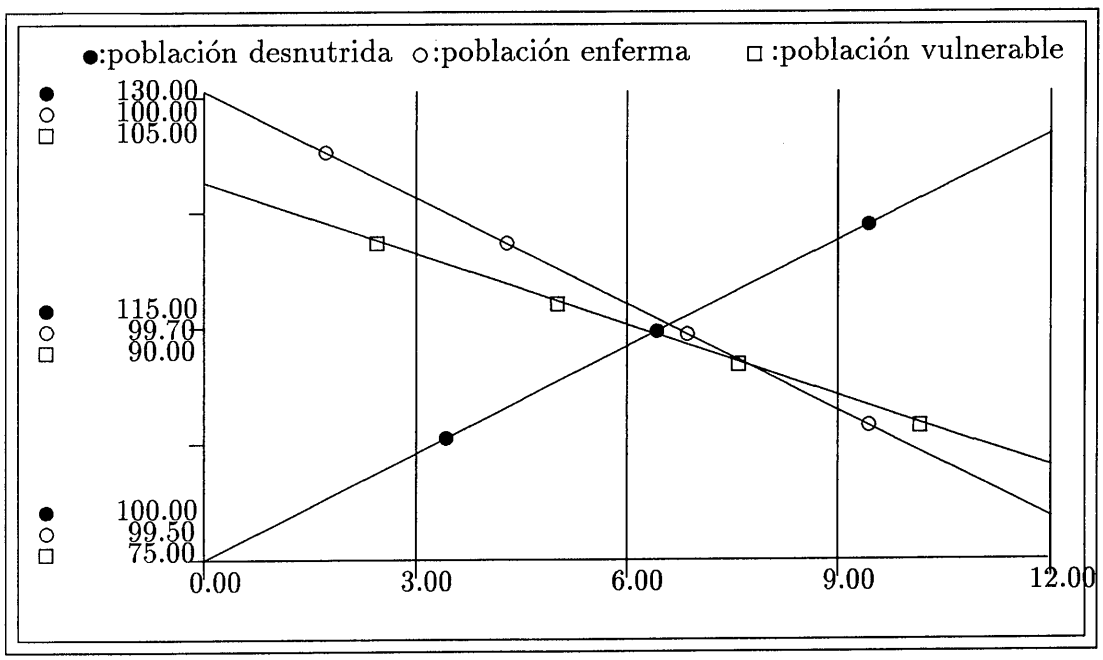
- $X_2(t) = X_2(t - dt) + (A_2 - B_2)dt$ en donde

  $A_2 = .02X_3 - .2X_4+$ índice de vulnerabilidad,

  $B_2 = .02X_3+$ índice de desnutrición.

A estas ecuaciones se les puede incorporar un elemento de azar, en los índices requeridos y como un error aditivo, de esta manera las ecuaciones anteriores son los valores esperados de estas relaciones.

La tendencia representada por las ecuaciones se presenta en la siguiente gráfica.



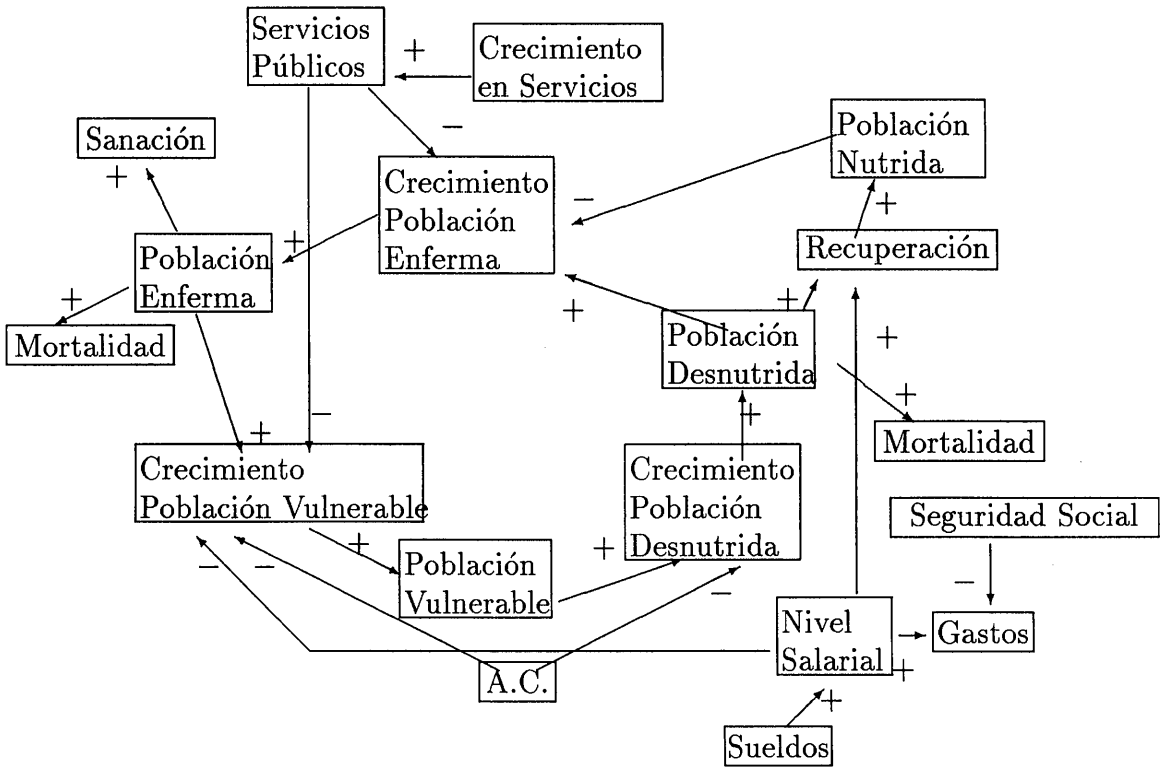
En la gráfica se observa que la tendencia de la población desnutrida aumenta, conforme pasa el tiempo. En este esquema no existe solución al problema, sino que va aumentando.

2.2 El círculo virtuoso

Ahora se agregan al modelo anterior las otras tres variables, esto tiene como consecuencia que se rompa el círculo vicioso ya descrito.

Las relaciones que se introducen son las siguientes: Al incrementar el nivel salarial de la población disminuye el crecimiento de la población vulnerable y aumenta la recuperación de la población desnutrida. Al incrementarse y mejorarse los servicios públicos baja la tasa de crecimiento de la población enferma y disminuye la tasa de crecimiento de la población vulnerable. Finalmente se incluye en el modelo el efecto de la labor realizada por las asociaciones civiles para vencer la desnutrición, al aumentar su labor disminuye la tasa de crecimiento de la población desnutrida, asimismo disminuye la tasa de crecimiento de la población vulnerable.

Las hipótesis consideradas en este modelo se esquematizan en el siguiente diagrama:



Al introducir esta nueva información al paquete se genera una serie de ecuaciones que describen el proceso, las ecuaciones son:

- $X_5(t) = X_5(t - dt) + (A_5 - B_5)dt$ en donde:

  A_5 = índice salarial + .02 empleos,

  B_5 =.02(índice de gastos de seguridad social)− .02(servicios públicos).

- $X_1(t) = X_1(t - dt) + (A_1 - B_1 - C_1)dt$ en donde:

  $A_1 = .02X_3(t)+$ índice de la desnutrición,

  $B_1 =$ índice de mortalidad de desnutridos y

  $C_1 =$ índice de recuperación$-.02X_5$.

- $X_3(t) = X_3(t - dt) + (A_3 - B_3 - C_3)dt$ en donde

  $A_3 = .02X_1 + .02X_2 - .2X_4 - .02X_6+$ índice del crecimimiento de la población enferma,

  $B_3 =$ índice de sanación y

  $C_3 =$índice de mortalidad por enfermedad.

- $X_4(t) = X_4(t - dt) + A_4dt$ en donde

  $A_4 = .02X_5+$ índice de recuperación.

- $X_2(t) = X_2(t - dt) + (A_2 - B_2)dt$ en donde

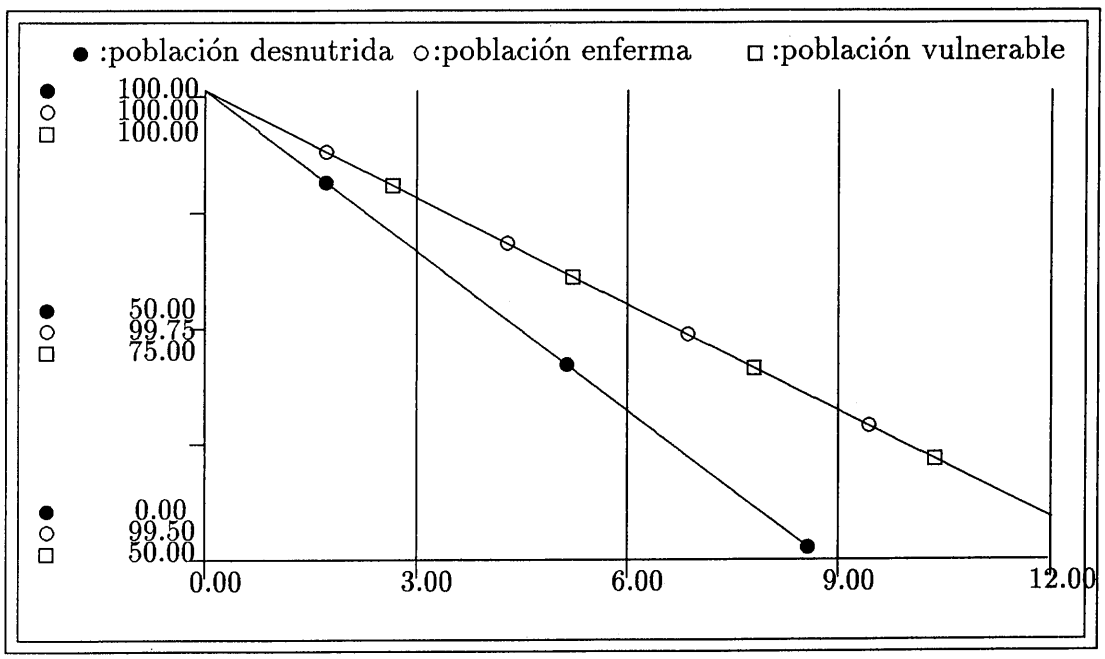
  $A_2 = .02X_3 - .2X_4 - .02X_5 - .02X_6 - .02X_7+$ índice de vulnerabilidad,

  $B_2 = .02X_3 - .02X_7+$ índice de desnutrición.

- $X_6(t) = X_6(t - dt) + A_6dt$ en donde

  $A_6 =$ índice de nivel de servicios públicos.

La gráfica resultante para este nuevo modelo es la siguiente:



Al interactuar las variables que se añaden al modelo, cambia totalmente el resultado que se observaba en el círculo vicioso Ahora es posible ver que los elementos añadidos impactan al desarrollo de la población desnutrida provocando que ésta disminuya de manera importante.

3 Conclusiones

La simulación digital permite reproducir y estudiar fenómenos complejos que aún no cuentan con un modelo que los describa. En el caso de la desnutrición se forma un círculo vicioso entre población vulnerable, población desnutrida y población enferma. Al incluir variables económicas, físicas y sociales el modelo se vuelve más complejo. Se confirman los supuestos de que estos últimos influyen en el desarrollo del problema y se vuelve un modelo propositivo de soluciones. El desarrollo de pruebas de hipótesis para modelos complejos como estos, son temas de estudio del análisis de datos estructurados.

Regresión Bayesiana: Análisis y Comparación de Modelos Lineales Generalizados

Gabriel Nuñez Antonio

ITAM

1 Introducción

Las técnicas de regresión usuales proponen alguna forma paramétrica para el predictor lineal y proceden a analizar el modelo como si éste fuera el verdadero, sin considerar la discrepancia entre la forma real y el predictor paramétrico asumido. En el presente trabajo se analiza un modelo de regresión generalizado semiparamétrico que toma en cuenta la discrepancia mencionada anteriormente y se propone un enfoque predictivo para la selección de modelos lineales generalizados desde un punto de vista Bayesiano

2 Descripción general del problema

Sea $\mathcal{M} = \{M_1, ..., M_k\}$ un conjunto de modelos paramétricos, donde

$$M_i = \{p_i(\mathbf{y}|\theta_i), p_i(\theta_i)\}; \qquad y \in \mathbf{Y},\ \theta_i \in \Theta.$$

Así, el modelo M_i se define por la verosimilitud $p_i(\mathbf{y}|\theta_i)$ y la correspondiente distribución inicial $p_i(\theta_i)$.

Problema: Seleccionar uno de los modelos en $\mathcal{M}$, dada la información inicial y una muestra de observaciones de $\mathbf{Y}$, con fines predictivos.

En la literatura existen varias formas de plantear el problema de selección y comparción de modelos. En Bernardo y Smith (1994) se discuten los llamados enfoques $\mathcal{M}$-cerrado y $\mathcal{M}$-abierto.

3 Soluciones tradicionales: enfoque $\mathcal{M}$-cerrado

3.1 Factores de Bayes

En la literatura existen (ver por ejemplo, Bernardo y Smith, 1994) técnicas basadas en los llamados factores de Bayes para tratar los problemas de selección de modelos y pruebas

de hipótesis. Sin embargo, el problema de los factores de Bayes es que éstos dependen fuertemente de la especificación de las distribuciones iniciales de los parámetros. Específicamente, los factores de Bayes no están bien definidos si se utilizan distribuciones iniciales impropias. Berger y Pericchi (1996) proponen un criterio llamado *factor de Bayes intrínseco*; por su parte O'Hagan (1995) utiliza los llamados *factores de Bayes fraccionales.* En ambos casos se pretende resolver algunas de las desventajas de los factores de Bayes.

3.2 Criterios predictivos

Otras propuestas que existen en la literatura consideran criterios predictivos para solucionar el problema de Comparación de Modelos. San Martini y Spezzaferri (1984) atacan el problema asumiendo el llamado en foque $\mathcal{M}$-cerrado y consideran una función de utilidad score-logarítmica (la cual resulta adecuada en estos casos).

En la práctica asumir el enfoque $\mathcal{M}$-cerrado es poco realista. En Bernardo y Smith (1994) se propone una solución aproximada, que involucra lo que se conoce como *validación cruzada*, para aproximar la utilidad esperada considerando el enfoque $\mathcal{M}$-abierto.

4 Propuesta semi-paramétrica

4.1 El modelo lineal normal

Supóngase que el conjunto de datos $\{(y_i,x_i),\ i = 1,...n\}$ sigue el modelo

$$y_i = \eta(x_i) + \varepsilon_i, \tag{1}$$

donde $\eta(x)$ es una función desconocida o su forma funcional es extremadamente complicada. En la práctica es frecuente aproximar $\eta(x)$ a través de una función simple $p_k(x) = \beta_0 + \beta_1 h_1(x) + ... + \beta_k h_k(x)$. Las técnicas de regresión usuales suponen que los $\{\varepsilon_i\}$ son variables aleatorias independientes, con $\varepsilon_i \sim N(0,\sigma^2)$ $(i = 1,...n)$, y analizan el modelo (1) considerando que $\eta(x)$ es igual a $p_k(x)$. Así, el modelo usual clásico ignora la discrepancia entre $\eta(\cdot)$ y $p_k(\cdot)$. En su lugar Blight & Ott (1975) proponen un modelo alternativo de la forma

$$y_i = p_k(x_i) + \delta_i + \varepsilon_i, \tag{2}$$

donde $\delta_i = \delta(x_i) = \eta(x_i) - p_k(x_i)$ es llamado el *error determínistico* del modelo. Los errores $\{\delta_i\}$ son determinísticos y difieren del componente de error aleatorio ε_i en el sentido de que $\delta_i = \delta_j$ si $x_i = x_j$, mientras que dos errores aleatorios para observaciones en el mismo punto de diseño son independientes.

El modelo (2) se puede reescribir como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \Sigma) \tag{3}$$

con $\mathbf{y} = (y_1, ..., y_n)'$, $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$, $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$, $\mathbf{X}$ la matriz con renglones $x_i' = (1, x_i, ..., x_i^k)$ y $\Sigma = \sigma^2 W$. Aquí W es una matriz diagonal de pesos con elemento j-ésimo m_j^{-1}.

4.2 Modelos lineales generalizados

Una generalización de los modelos lineales clásicos son los llamados *modelos lineales generalizados*, los cuales tienen como casos especiales a los modelos de regresión lineal, de análisis de varianza, los modelos logit, probit, loglineales, modelos de respuesta multinomial y algunos modelos para datos de supervivencia.

Los modelos lineales generalizados se pueden específicar a través de las siguentes componentes:

Componente Aleatoria. Y_1, ..., Y_n son variables aleatorias independientes cuya distribución es un miembro de la familia exponencial, con función de densidad de probabilidad dada por

$$f(y_i \mid \theta_i, \sigma^2) = b(y_i, \sigma^2/m_i) \exp(m_i[y_i\theta_i - a(\theta_i)]/\sigma^2) \tag{4}$$

con $a(\cdot)$ y $b(\cdot, \cdot)$ ciertas funciones especificas, donde los $\{m_i\}$ son pesos conocidos, asociados a cada observación. Si σ^2 se conoce, entonces (4) es un modelo que pertenece a la *familia exponencial natural* con *parámetro canónico* θ_i. El parámetro σ^2 es conocido como el *parámetro de dispersión.*

Componente sistemática. Para cada respuesta Y_i, se tiene asociado un vector de covariables $x_i = (x_{i1}, ..., x_{ir})'$, con el cual se obtiene el predictor lineal $\eta_i = \eta(x_i) = p_k(x_i) + \delta$.

Liga. Las componentes aleatoria y sistemática se relacionan vía la función liga, de tal manera que $\eta_i = g(\mu_i)$. Una liga particularmente importante se obtiene cuando $g^{-1}(\cdot) = a'(\cdot)$. En este caso $\theta_i = \eta_i$ y $g(\cdot)$ se denomina la *liga canónica.*

Especificación del Modelo Semiparamétrico

Nivel 1. Condicionalmente sobre β y δ, Y_1, ..., Y_n son variables aleatorias independientes con $y_i \sim f(y_i \mid \theta_i, \sigma^2)$ y $\theta_i = t(x_i'\beta + \delta_i)$ $(i = 1, ..., n)$. Lo cual produce la siguiente verosimilitud aproximada

$$l(y \mid \beta, \delta) \propto \exp\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n} m_i \left[y_i t(x_i'\beta + \delta_i) - a(t(x_i'\beta + \delta_i))\right]\right\}$$

Nivel 2. Condicionalmente sobre ρ^2 y λ, los parámetros β y δ son independientes y

$$\beta \sim N(b_0, B_0^{-1})$$
$$\delta \sim N(\mathbf{0}, \rho^2 \Lambda_\lambda)$$

La densidad final correspondiente a esta especificación inicial es de la forma

$$\begin{aligned} p(\beta, \delta \mid \mathbf{y}) &\propto \exp\left\{\frac{-1}{2}(\gamma - t_0)'T_0(\gamma - t_0)\right\} \\ &\quad \times \exp\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n} m_i \left[y_i t(x_i'\beta + \delta_i) - a(t(x_i'\beta + \delta_i))\right]\right\} \\ &= \exp\left\{\frac{-1}{2}(\gamma - t_0)'T_0(\gamma - t_0) + \frac{1}{\sigma^2}\sum_{i=1}^{n} m_i \left[y_i t(D_i\gamma) - a(t(D_i\gamma))\right]\right\} \quad (5) \end{aligned}$$

donde $\gamma = (\beta', \delta')'$ es un vector de $(p+n)\times 1$, con $\delta' = (\delta_1, ..., \delta_n)$ un vector de $n\times 1$. $t_0 = (b_0, \mathbf{0})'$ un vector de $(p+n)\times 1$, T_0 una matriz de $(p+n)\times(p+n)$ con $[B_0, (\rho^2\Lambda_\lambda)]$ como elementos de su diagonal. Aquí, $D_i = (x_i', 1_i')'$ un vector de $(p+n)\times 1$, con 1_i un vector de n componentes todos iguales a cero excepto en el lugar *i-ésimo* donde tiene un valor de uno.

No en todos los casos se pueden hacer inferencias analíticas exactas basadas en (5). Las inferencias clásicas para modelos lineales generalizados se basan en la estimación de parámetros y las propiedades distribucionales asintóticas de los estimadores. Sin embargo, en general la maximización de la verosimilitud requiere de métodos numéricos. Por otro lado, los métodos de Monte Carlo vía Cadenas de Markov producen una forma relativamente directa para hacer inferencias Bayesianas para una clase amplia de modelos lineales generalizados.

Dellaportas y Smith (1993) utilizan el muestreo de Gibbs en este contexto, para hacer inferencias sobre la densidad final (5); también pueden emplearse distintas versiones del algoritmo de Metropolis-Hasting (ver, por ejemplo, Smith y Roberts, 1993), tales como los algoritmos de *Caminata Aleatoria* e *Independencia.* Estos últimos son los que se utilizan en este rabajo.

4.3 Comparación de modelos lineales generalizados

Sea $\mathcal{M} = \{M_1, ..., M_k\}$ un conjunto de modelos de regresión paramétricos, donde

$$M_i = \{p_i(y|\boldsymbol{\beta}_i),\ p_i(\boldsymbol{\beta}_i)\} \qquad y \in \mathbf{R}, \quad \boldsymbol{\beta}_i \in \mathbf{R}^{q_i}.$$

En este caso

$$p_i(y|\boldsymbol{\beta}_i) = f(y \mid \theta(\boldsymbol{\beta}_i), \sigma^2)$$

y

$$p_i(\boldsymbol{\beta}_i) = N_{q_i}(\boldsymbol{\beta}_i | b_{0i}, \mathbf{B}_{0i}^{-1}),$$

donde $f(y \mid \theta(\boldsymbol{\beta}_i),\sigma^2) = b(y,\sigma^2/m)\exp(m[y\theta - a(\theta)]/\sigma^2)$ denota una familia exponencial y $\theta(\boldsymbol{\beta}_i) = t(\mathbf{h}_i(x)'\boldsymbol{\beta}_i$ define tanto al predictor lineal como a la función liga.

Problema: Seleccionar uno de los modelos en $\mathcal{M}$ con fines predictivos.

4.3.1 Solución al problema de selección de modelos lineales generalizados

El enfoque semiparamétrico para el problema de selección de modelos, considera éste como un problema de decisión con los siguientes elementos.

Espacio de decisiones:

$$\mathbf{D} = \mathcal{M}$$

Espacio de 'sucesos inciertos':

$$\mathbf{E} = \{\eta : \eta \text{ es una función suave sobre } \mathbf{R}^r\}$$

Función de utilidad sobre D xE:

$$u(M_i,\, \eta) = \int \log p_i(y_*|\mathbf{y}) f(y_*|\eta(x_*))dy_*,$$

donde

$$p_i(y_*|\mathbf{y}) = \int p_i(y_* \mid \beta_i)\; p_i(\beta_i \mid \mathbf{y})\; d\beta,$$

es la distribución predictiva final de una observación futura y_*, considerando el i-ésimo modelo. Aquí $f(y_*|\eta(x_*))$ representa un modelo de la familia exponencial.

Distribución inicial sobre E :

Nivel I. Condicional sobre $\boldsymbol{\beta}$

$$\eta(x) \sim \mathcal{N}(\mu^*_\beta(x), \Sigma^*(x,x)) \qquad \text{(Proceso Gaussiano)}$$

con

$$\mu^*_\beta(x) = \mathbf{h}(x)'\boldsymbol{\beta}$$
$$\Sigma^*(x,x) = \rho^2\Lambda_\lambda$$

es decir, dado $\boldsymbol{\beta}$

$$\eta(x) = \mathbf{h}(x)'\boldsymbol{\beta} + \boldsymbol{\delta}(x), \qquad \text{donde } \boldsymbol{\delta}(x) \sim \mathcal{N}(0, \rho^2\Lambda_\lambda).$$

Nivel II.

$$\boldsymbol{\beta} \sim N_q(b_0,^2 \mathbf{B}_0^{-1}).$$

Duración de la Diabetes (en años)	Estudio previo. Presencia de Retinopatía		Estudio actual. Presencia de Retinopatía	
	Si	No	Si	No
0-2	17	215	46	290
3-5	26	218	52	211
6-8	39	137	44	134
9-11	27	62	54	91
12-14	35	36	38	53
15-17	37	16	39	42
18-20	26	13	23	23
21+	23	15	52	32

Tabla 1. Datos de diabetes-retinopatía, Knuiman y Speed (1988)

Solución: Seleccione aquel modelo que maximice la utilidad esperada final,

$$\begin{aligned}\bar{u}(M_i) &= \mathrm{E}_{\eta|\mathbf{y}}\left[u(M_i,\,\eta)\right] \\ &= \mathrm{E}_{\eta|\mathbf{y}}\left[\int \log p_i(y_*|\mathbf{y})\; f(y_* \mid \eta(x_*)\;)\, dy_*\right]\end{aligned}$$

Así, el modelo M_i será preferido al modelo M_j si y sólo si

$$\bar{u}(M_i) > \bar{u}(M_j).$$

Para una revisión completa de este trabajo, ver Nuñez (1998).

5 Ejemplo

Los datos de la Tabla 1, tomados de Knuiman y Speed (1988), reflejan la relación entre la duración de la diabetes (medida en años) y la retinopatía (una enfermedad de los ojos).

Knuiman y Speed sugieren el uso de un modelo logístico definido por

$$\log\left(\frac{\pi_{1j}}{\pi_{2j}}\right) = \beta_1 + \beta_2 X_j + \beta_3 X_j^2 = \eta_j$$

5.1 El modelo semiparamétrico

La especificación del modelo semiparamétrico se realizó asumiendo

$$\rho^2 = 0.7, \qquad \lambda = e^{-1} \qquad \text{y}$$
$$\mu_\beta^* = \mathbf{h}(x_*)'\beta = \beta_1 + \beta_2 x + \beta_3 x^2.$$

π_{1i} observadas	Modelo Semiparamétrico	Ajuste en S-plus
0.13690	0.13239	0.09869
0.19772	0.19609	0.16601
0.24720	0.24367	0.25215
0.37241	0.37248	0.34730
0.41759	0.42155	0.43894
0.48148	0.48257	0.51731
0.50000	0.50553	0.57768
0.61905	0.63752	0.63707
Devianza	0.001054	3.60394E-02

Tabla 2. Ajuste obtenido a los datos de diabetes-retinopatía

La tabla 2 presenta el ajuste obtenido con el modelo semiparamétrico y el correspondiente obtenido con S-plus.

Modelo	M_0	M_1	M_2	M_3	M_4	Semiparamétrico
C. Aleatoria	-9.20	-5.484	-4.989	-4.984	-4.994	-4.948

Tabla 3. Utilidades esperadas finales, $\bar{u}(M_i)$.

5.2 El modelo semi-paramétrico. Comparación de modelos

Para el estudio de diabetes-retinopatía supóngase que se desean comparar los siguientes modelos:

$$
\begin{aligned}
M_0 &: \log\left(\frac{\pi_{1i}}{\pi_{2i}}\right) = \beta_{00} \\
M_1 &: \log\left(\frac{\pi_{1i}}{\pi_{2i}}\right) = \beta_{10} + \beta_{11}x \\
M_2 &: \log\left(\frac{\pi_{1i}}{\pi_{2i}}\right) = \beta_{21} + \beta_{22}x + \beta_{23}x^2 \\
M_3 &: \log\left(\frac{\pi_{1i}}{\pi_{2i}}\right) = \beta_{31} + \beta_{32}x + \beta_{33}x^2 + \beta_{34}x^3 \\
M_4 &: \log\left(\frac{\pi_{1i}}{\pi_{2i}}\right) = \beta_{41} + \beta_{42}x + \beta_{43}x^2 + \beta_{44}x^3 + \beta_{45}x^4
\end{aligned}
$$

Para efectos del ejemplo se consideraron las siguientes especificaciones:

$$
\begin{aligned}
x_* = \bar{x} = 11.75,\ \rho^2 = 0.7,\ \lambda = e^{-1} \quad &\text{y} \\
\mu^*_\beta = \mathbf{h}(x_*)'\beta = \beta_1 + \beta_2\bar{x} + \beta_3\bar{x}^2 &
\end{aligned}
$$

La Tabla 3 muestra las utilidades esperadas finales para cada uno de los modelos a comparar.

Se puede notar que el modelo M_3 es el mejor modelo y que en términos del poder predictivo los modelos M_2 y M_3 son equivalentes.

Referencias

Berger, J. O. y Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *J. Am. Statist. Assoc.* 91, 109-122.

Bernardo, J. M. y Smith, A. F. M. (1994). Bayesian Theory. Chichester: Wiley.

Blight, B. J. N. y Ott, L. (1975). A Bayesian Approach to Model Inadequacy for Polinomial Regression. *Biometrika* 62, 79-88.

Dellaportas, P. y Smith, A. F. M. (1993). Bayesian Inference for Generalised Linear and Proportional Hazard Model via Gibbs Sampling. *Appl. Statist.* 42, 443-459.

Knuiman , M. W. y Speed, T. P. (1988). Incorporating prior information into the analysis of contingency tablas. *Biometrics*, 44, 1061-1071.

Nuñez, A. G. (1998). Tesis de Maestría. Regresión Bayesiana: Análisis y Comparación de Modelos Lineales Generalizados. Dirección General de Estudios de Posgrado, UNAM.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison (with discussion). *J. of the Roy. Statist. Soc. B* 57, 99-138.

San Martini, A. y Spezzaferri, F. (1984). A Predictive Model Selection Criterion. *J. of the Roy. Statist. Soc. B* 46, 296-303.

Smith, A. F. M. y Roberts G. O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *J. of the Roy. Statist. Soc. B* 55, 3-23.

Inferencias en Mezclas Bajo Censura con Identificación Parcial

Federico O'Reilly
IIMAS-UNAM

1 Introducción

Considere una situación en la que se tienen observaciones independientes de una variable aleatoria cuya distribución es una mezcla de k poblaciones, cada una con su respectivo parámetro, ya sea vectorial o escalar. En adición, suponga que la variable aleatoria será observada sólo si su valor es menor o igual que un valor conocido de censura C y en ese caso, se sabrá de cual de las k poblaciones provino. En caso contrario, la observación sólo se sabrá haber sido superior a C y no se conocerá qué población la generó.

En este trabajo se propone un método aproximado para hacer inferencias sobre los parámetros de cada población, que resulta muy sencillo. Se ilustra con un ejemplo clásico de tiempos de falla recientemente discutido por Díaz-Francés (1998) en su tesis doctoral.

2 El Problema

Considere la función de distribución formada por una mezcla de k distribuciones $F_j(\cdot;\theta_j)$, $j = 1, \ldots, k$ dada por

$$G(\cdot;\boldsymbol{\theta},\mathbf{p}) = \sum_{j=1}^{k} p_j F_j(\cdot;\theta_j) \qquad \text{y}$$

$\mathbf{p}' = (p_1, p_2, \ldots, p_k)$ con los p_j las proporciones de la mezcla; esto es, $p_j > o$ para cada j y

$$\sum_{j=1}^{k} p_j = 1.$$

Cada observación $X \sim G$, tienen la peculiaridad de que permite identificar la distribución particular F_j de la que proviene si $X < C$. En caso contrario no se conoce el valor particular de X ni la identidad de la F_j que la produjo.

La información recabada después de hacer N observaciones independientes de G, puede resumirse en

$$\{x_{1i}\}_{i=1}^{n_1}, \{x_{2i}\}_{i=1}^{n2}, \ldots, \{x_{ki}\}_{i=1}^{n_k} \quad \text{y } r$$

en dónde, para cada $j = 1, \ldots, k$,

$x_{j1}, \ldots x_{jnj}$ son los n_j observaciones menores que C, identificadas como provientes de F_j y r es el número de observaciones (de las N) que excedieron a C.

La verosimilitud para $\boldsymbol{\theta}$ y $\mathbf{p}$, denotada por $L(\boldsymbol{\theta}, \mathbf{p})$ es proporcional a

$$\left\{ \Pi_{j=1}^{k} \Pi_{i=1}^{n_j} \frac{f_j(x_{ji}; \theta_j)}{F_j(C; \theta_j)} \right\} \Pi_{j=1}^{k} \{p_j F_j(C; \theta_j)\}^{n_j}$$

$$\{1 - G(C; \boldsymbol{\theta}, \mathbf{p})\}^r$$

siendo el primer doble producto en corchetes la densidad de todas las $\{x_{ji}\}_{i=1}^{n_j}$, $j = 1, \ldots, k$, condicional a que fueran $<^s C$, y el segundo producto y último término, la parte correspondiente a la probabilidad (multinomial) asociada a las conteos:

$$n_1, n_2, \ldots, n_k \qquad \text{y } r.$$

En Mendenhall y Hader (1958) se utiliza el esquema anterior para $k = 2$ y las F_j correspondientes a exponenciales negativas de un solo parámetro. El modelo se utiliza para analizar tiempos de falla de componentes que se observan hasta $C = 630$ horas y que al haber fallado durante el período de observación se podía identificar la falla dentro de una de dos posibles clases o poblaciones.

De las $N = 369$ observaciones, $r = 44$ excedieron al valor de 630 y por ello no quedaron registradas ni se supo cuántas de estas provenían de cada una de las dos poblaciones. Las 325 observaciones restantes corresponden a $n_1 = 107$ de la primera población y $n_2 = 218$ de la segunda. La lista de estas 325 observaciones puede verse en el artículo de Mendenhall y Hader.

En Díaz-Francés (1998), se hace un análisis de los datos estudiados por Mendenhall y Hader, pero modelando con distribuciones Weibull con parámetros de forma y escala. Se manejan los logaritmos naturales de las observaciones convirtiendo los parámetros en escala y localización en las correspondientes distribuciones de valor extremo. En adición se utiliza la parametrización, media y log(escala) y a los datos se les resta el logaritmo de C.

Se obtienen las verosimilitudes perfil para cada uno de los 5 parámetros $(\mu_j, \log \sigma_j), j = 1, 2$ y $p_1 = \mathrm{p}$ $(p_2 = 1 - \mathrm{p})$ y se estudia tanto la aproximación $\log F$ como la aproximación normal a estas verosimilitudes perfil, con el objeto de exhibir intervalos de verosimilitud y poder asignar la cobertura (confianza) aproximada.

3 Variables latentes

Si en el esquema del problema presentado se incorporan como variables artificiales (latentes) los números enteros $r_1, r_2, \ldots, r_k$, con $r_j \geq 0$ y $\sum_{j=1}^{k} r_j = r$ en que r_1 representaría al número de observaciones que excedieron a C y provenían de F_1, r_2 las que excedieron a C y provenían de F_2, etc. entonces el vector $\mathbf{p}$ de parámetros ya no sería necesario y la verosimilitud, conociendo $r_1, r_2, \ldots, r_k$, estaría dada por $L(\boldsymbol{\theta}) \propto \Pi_{j=1}^{k} L_j(\theta_j)$ que representa el producto de k verosimitudes $L_j(\theta_j) \propto$

$$\left\{\Pi_{i=1}^{n_j} \frac{f_j(x_{ji})}{F_j(C;\theta_j)}\right\} \binom{n_j + r_j}{n_j} \{F_j(C;\theta_j)\}^{n_j}$$

$$\{1 - F_j(C;\theta_j)\}^{r_j}$$

en la que el producto en los primeros corchetes es la densidad de las $\{x_{ji}\}_{i=1}^{n_j}$ condicional a que fueran $<^s C$ y el resto es la probabilidad de que de las $n_j + r_j$ observaciones de la población j, n_j hayan sido $<^s C$ y $r_j >^s C$.

O sea que "conociendo $r_1, r_2, \ldots, r_k$", se puede factorizar la verosimilitud e inferir sobre un θ_j sólo de su correspondiente L_j.

Sea $\hat{\theta}_j(r_j)$ el estimador máximo verosímil de θ_j para r_j fijo. Si se estudia la "verosimilitud perfil" para $r_1, r_2, \ldots, r_k$ dada por

$$\Pi_{j=1}^{k} L_j(\hat{\theta}_j(r_j)),$$

se puede entonces buscar las $\hat{r}_j$ que maximicen la expresión anterior, y los estimadores máximo verosímiles (absolutos) serán los $\hat{\theta}_j(\hat{r}_j)$.

Para los datos de Mendenhall y Hader, se hizo el ejercicio descrito utilizando las variables latentes r_1 y r_2 (pero $r_2 = r - r_1$ por lo que en realidad es sólo una) y se llegó a:

$$\hat{r}_1 = 2$$

La verosimilitud perfil para r_1 resultó (numéricamente) en

$$\begin{array}{lll} 1.18 & \text{para} & r_1 = 0 \\ 1.76 & \text{para} & r_1 = 1 \\ 1.83 & \text{para} & r_1 = 2 \\ 1.61 & \text{para} & r_1 = 3 \\ \vdots & \vdots & \vdots \\ 0.61 & \text{para} & r_1 = 7 \\ \vdots & \vdots & \vdots \end{array}$$

Siendo el recorrido posible para $r_1 : 0, 1, 2, \ldots, 44$. Los valores de la perfil podrían inclusive ser utilizados para exhibir una distribución sobre r_1.

El uso de la variable latente r_1, permite evaluar la verosimilitud perfil para, digamos μ_1 de $L_1(\mu_1, \log \sigma_1)$ haciendo una maximización sólo sobre un parámetro. En el análisis sin el uso de la variable latente, la verosimilitud perfil para μ_1 se tiene que obtener de $L(\mu_1, \mu_2, \log \sigma_1, \log \sigma_2, \mathrm{p})$; esto es haciendo una maximización sobre cuatro parámetros.

Para la utilización de la aproximación normal para el correspondiente estimador máximo verosimil $\hat{\mu_1}$, el uso de la variable latente permite evaluar la varianza asintótica utilizando la matriz observada de Fisher proveniente de $L_1(\mu_1, \log \sigma_1)$, en comparación con la matriz de 5×5 que se requiere si se busca una aproximación normal sin utilizar la variable latente r_1.

Las comparaciones, desde luego se vuelven mucho más dramáticas al considerar valores mayores de k.

En el apéndice, aparecen las gráficas de las verosimilitudes perfil obtenidas mediante la consideración de la variable latente r_1, para el valor utilizado $\hat{r}_1 = 2$. Dichas gráficas aparecen prácticamente "encimadas " sobre sus correspondientes verosimilitudes provenientes del uso de la aproximación normal.

En relación a los resultados reportados en Díaz-Francés (1998), existe una buena concordancia entre los estimadores máximo verosímiles reportados allí y los obtenidos considerando la variable latente r_1, habiéndose asignado el valor $\hat{r}_1 = 2$,

Estimaciones Máximo-Verosímiles (y desviaciones aproximadas)

	Díaz-Francés	Con Variable Latente
μ_1	-1.0451 (.0975)	-1.0606 (.0785)
μ_2	-0.5781 (.0636)	-0.5758 (.0607)
$\log \sigma_1$	-0.2362 (.0937)	-0.2525 (.0764)
$\log \sigma_2$	-0.1180 (.0571)	-0.1161 (.0565)
(p estimada por 0.2975)		(p implicada es 0.2954)

Se hace la observación de que en todos los casos la desviación obtenida con el uso de la variable latente es inferior, ya que supone un conocimiento adicional y por ello conduce a intervalos de confianza ligeramente más chicos que los que obtiene Díaz-Francés.

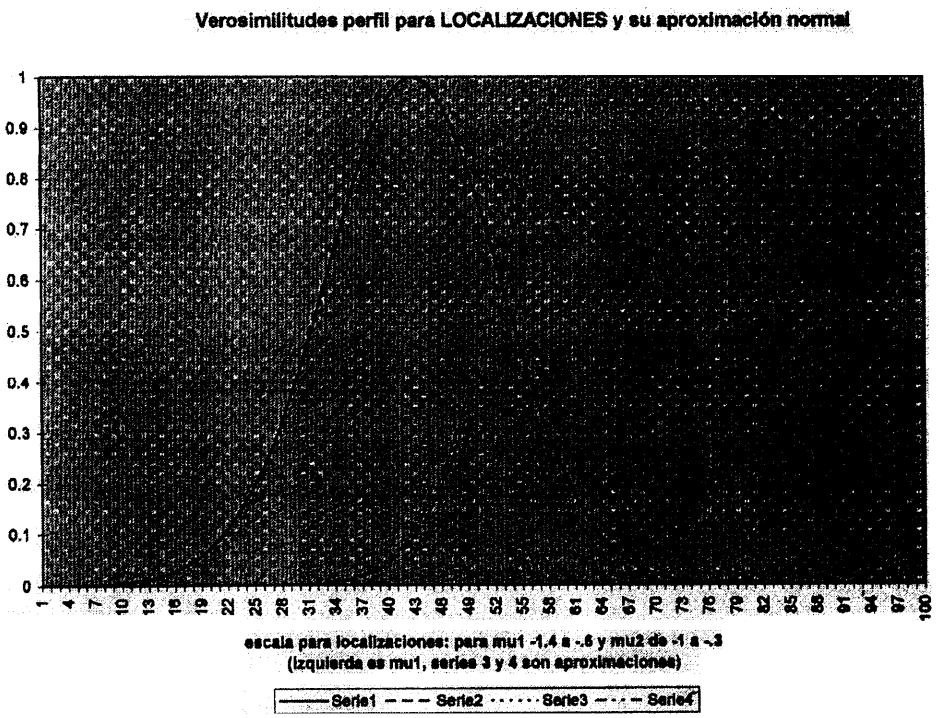
Aunque no se explicita el parámetro p cuando se modela con el uso de la variable latente r_1, puede inferirse sobre él aduciendo que el número de observaciones provenientes de la población 1, fueron 109 de las 369 (las 107 identificados más las $\hat{r}_1 = 2$ asignadas, de allí el valor implicado para p en la tabla).

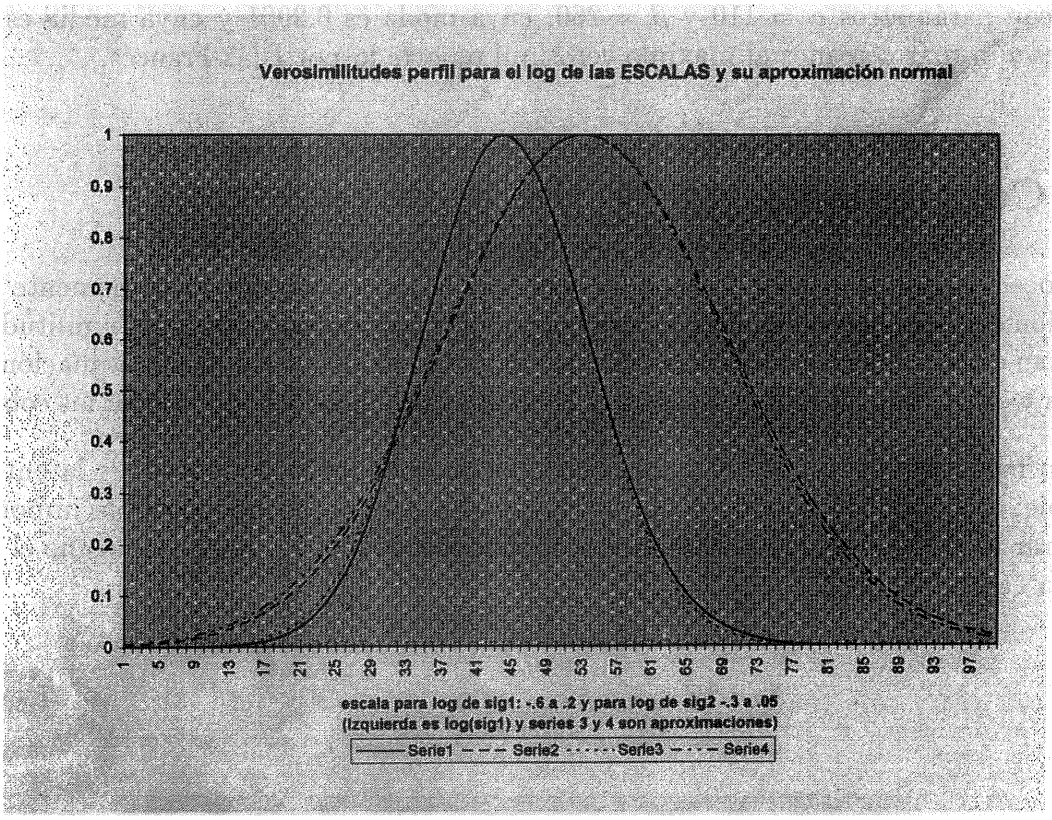
Utilizando la distribución fiducial para p, $\{1 - Bin(109; N, \mathrm{p})\}$ con $Bin(\cdot; N, \mathrm{p})$ el valor de la función de distribución binomial con parámetros N y p en $\cdot$, se tiene que ésta es una beta con parámetros $\alpha = 110$ y $\beta = 260$, cuya moda es 0.2961 y cuya media es 0.2973, valores aún más cercanos al máximo verosímil reportado por Díaz-Francés.

4 Conclusiones

El procedimiento propuesto basado en variables latentes simplifica enormemente la construcción de verosimilitudes perfil (aproximadas). ¿En qué grado estas verosimilitudes perfil resultan en una buena o mala aproximación?. En el ejemplo visto, la aproximación es muy buena y quizás sólo hagan falta algunos estudios de simulación para verificar las coberturas.

En mezclas de más de dos poblaciones, el procedimiento con variables latentes se ve factible; no siendo así el método que no las considera. Por ello, consideramos importante el elaborar un estudio comparativo para otros ejemplos de $k = 2$ poblaciones como el descrito y para casos con valores mayores de k.





Referencias

Cox, D.R. (1959). The analysis of exponentially distributed life-times with two types of failure. JRSS, Series B, V 21, pp 411-421.

Díaz Francés, E. (1998). Scientific application of maximum likelihood in multiparametric problems. PhD Dissertation under D.A. Sprott. CIMAT, México.

Mendenhall, W. and R.J. Hader (1958). Estimation of parameters of mixed exponentially distributed failure time distribuions from censored life test data. Biometrics, V 45, pp 504-520.

Estimación de Componentes de Varianza en un Modelo Partido Aplicado a un Ensayo Agronómico

Emilio Padrón Corral
UAAAN Saltillo, Coah.

Angel Martínez Garza
COLPOS, Edo. de México

Gustavo Burciaga Vera
UAAAN Saltillo, Coah.

1 Introducción

Los componentes del modelo, de acuerdo a la inferencia estadística pueden clasificarse en efectos fijos y aleatorios la combinación de estos efectos genera los llamados modelos mixtos; aquí se hará referencia a los aleatorios y de ellos nos interesa lo concerniente a sus varianzas. El objetivo del trabajo es el de estimar los componentes de varianza de un modelo partido y observar el efecto de la componente de genotipos y su interacción con localidades en un ensayo con híbridos de maíz. Por lo tanto, es necesario estudiar el comportamiento de genotipos de maíz en diferentes ambientes, y se logra el verdadero efecto de sus varianzas contando para ello con la herramienta fundamental de las componentes de varianza estimadas, en base a la técnica del análisis de varianza (ANOVA) como se aprecia en Searle (1987), también autores como Brownlee (1984), nos comenta como Bennet y Franklin elaboraron un procedimiento para obtener los valores d e esperanzas de cuadrados medios en situaciones parcialmente jerárquicas. Searle et al. (1992), comentan que las sumas de cuadrados del análisis de varianza para datos desbalanceados siguen siendo los mismos que para datos balanceados excepto que en lugar de tener n se debe tener n_i, y en lugar de N se debe tener $\sum n_i$, y el que los datos sean desbalanceados no elimina la posibilidad de obtención de estimadas negativas de σ_τ^2 (componentes de varianza para tratamientos) en el análisis de varianza. La teoría desarrollada en este trabajo se aplicó a un experimento de campo titulado "Aptitud Combinatoria de Líneas S_2 en Cruza con tres Probadores Contrastantes". Esta investigación forma parte del programa de mejoramiento genético del Instituto Mexicano del Maíz (Mario E. Castro Gil) de la U.A.A.A.N. y consta de dos localidades, Celaya, Guanajuato y Torreón, Coahuila, respectivamente; además de 78 híbridos tanto experimentales como comerciales, utilizados estos últimos como testigos.

2 Desarrollo

2.1 Descripción del problema

Dado un modelo estadístico con varios factores agronómicos e interacción se descompondrá como sigue: Genotipos se partirá en cruzas y éstas en líneas dentro de probador uno, dos y tres; también se obtiene la información de probadores, testigos, el contraste de cruzas contra testigo y todos estos efectos se interactuarán con localidades; de cada uno de ellos se desarrollarán las sumas de cuadrados y se obtendrán las esperanzas de cuadrados medios para conocer sus correspondientes componentes de varianza estimadas y por lo tanto, saber en cuanto están contribuyendo de acuerdo a los resultados del experimento de campo.

2.2 Metodología

Se aplicará la técnica de la esperanza de cuadrados medios en el modelo que a continuación se presenta:

$$Y_{ijk} = \mu + L_k + R_{j/k} + G_i + (LG)_{ki} + E_{ijk}$$

donde

$$\begin{aligned} i &= 1, 2, 3, \ldots, t \quad \text{genotipos} \\ j &= 1, 2, 3, \ldots, r \quad \text{repeticiones} \\ k &= 1, 2, 3, \ldots, l \quad \text{localidades} \end{aligned}$$

Y_{ijk} : Variable aleatoria observable de la k-ésima localidad en la j-ésima repetición del i-ésimo genotipo

μ : Media general

L_k : Efecto de la k-ésima localidad

$R_{j/k}$: Efecto de la j-ésima repetición dentro de la k-ésima localidad

G_i : Efecto del i-ésimo genotipo

$(LG)_{ki}$: Efecto conjunto de la k-ésima localidad y del i-ésimo genotipo

E_{ijk} : Componente aleatoria asociada con la k-ésima localidad en la j-ésima repetición del i-ésimo genotipo

Por lo tanto de acuerdo a dicho modelo, el cuadrado medio de la interacción Localidad×Genotipo es el apropiado cuadrado medio del error para probar genotipos. Además se asume que las esperanzas de efectos son cero, es decir,

$$E[L_k] = E[R_{j/k} = R/L] = E[G_i] = E[(LG)_{ki}] = E[E_{ijk}] = 0$$

También se supone que las esperanzas de productos cruzados de los diferentes efectos son cero, se tiene además que

$$E[L_k^2] = \sigma_L^2 \qquad E[R/L]^2 = \sigma_{r/L}^2 \qquad E[G_i^2] = \sigma_G^2 \qquad E[(LG)_{ki}^2] = \sigma_{LG}^2 \qquad E[E_{ijk}^2] = \sigma_e^2$$

2.3 Resultados

A continuación se obtienen las esperanzas de cuadrados medios de cada uno de los efectos, pero en el apéndice se describe el desarrollo de algunos de ellos para un mejor entendimiento del tema.

$$\begin{aligned}
E[CM(Loc)] &= rt\sigma_L^2 + r\sigma_{LG}^2 + t\sigma_{r/L}^2 + \sigma_e^2 \\
E[CM(Rep/Loc)] &= t\sigma_{r/L}^2 + \sigma_e^2 \\
E[CM(Gen)] &= rl\sigma_G^2 + r\sigma_{LG}^2 + \sigma_e^2 \\
E[CM(Cruza)] &= rl\sigma_C^2 + r\sigma_{LC}^2 + \sigma_e^2 \\
E[CM(Lin/p_1)] &= rl\sigma_{l/p_1}^2 + r\sigma_{L(l/p_1)}^2 + \sigma_e^2 \\
E[CM(Lin/p_2)] &= rl\sigma_{l/p_2}^2 + r\sigma_{L(l/p_2)}^2 + \sigma_e^2 \\
E[CMLin/p_3] &= rl\sigma_{l/p_3}^2 + r\sigma_{Ll/p_3}^2 + \sigma_e^2 \\
E[CM(Prob)] &= rl\sigma_P^2 + r\sigma_{LP}^2 + \sigma_e^2 \\
E[CM(Tes)] &= rl\sigma_T^2 + r\sigma_{LT}^2 + \sigma_e^2 \\
E[CM(Cruza \text{ vs } Test)] &= rl(\sigma_C^2 + \sigma_T^2 - \sigma_G^2) + r(\sigma_{LC}^2 + \sigma_{LT}^2 - \sigma_{LG}^2) + \sigma_e^2 \\
E[CM(Gen \times Loc)] &= r\sigma_{LG}^2 + \sigma_e^2 \\
E[CM(Cruza \times Loc)] &= r\sigma_{LC}^2 + \sigma_e^2 \\
E[CM((l/p_1) \times Loc)] &= r\sigma_{L(l/p_1)}^2 + \sigma_e^2 \\
E[CM((l/p_2) \times Loc)] &= r\sigma_{L(l/p_2)}^2 + \sigma_e^2 \\
E[CM((l/p_3) \times Loc)] &= r\sigma_{L(l/p_3)}^2 + \sigma_e^2
\end{aligned}$$

$$E[CM(Prob \times Loc)] = r\sigma^2_{LP} + \sigma^2_e$$

$$E[CM(Tes \times Loc)] = r\sigma^2_{LT} + \sigma^2_e$$

$$E[CM((Cruza \text{ vs } Tes) \times Loc)] = r(\sigma^2_{LC} + \sigma^2_{LT} - \sigma^2_{LG}) + \sigma^2_e$$

$$E[CM(Error)] = \sigma^2_e$$

En seguida se presentan cada uno de los valores de las componentes de varianza estimadas (C.V.E) de efectos principales e interacciones, así como de sus particiones para la variable rendimiento de mazorca en ton/ha^{-1} del experimento de Cuellar Chavez (1998) como se aprecia en la Tabla 1.

Tabla 1. Estimados de Componentes de Varianza por Fuente de Variación.

F.V.	**C.V.E.**
localidad	41.83450
rep/Loc	0.04463
Genotipos	4.61825
cruzas	1.75575
lin/p1	0.07325
lin/p2	0.23475
lin/p3	0.49250
Probadores	46.93500
Testigos	1.10700
Cruzas vs testigos	228.78675
Gen * Loc	2.25100
cruzas * Loc	1.52650
(lin/p1) * Loc	0.00900
(lin/p2) * Loc	0.01400
(lin/p3) * Loc	0.01900
Probadores * Loc	52.17175
testigos * Loc	1.07800
(cruza vs test) * Loc	60.44950
error	2.11200
total	451.26787

En este trabajo se obtuvo no sólo la magnitud relativa de las varianzas, sino también los porcentajes de la suma de las varianzas de cada uno de ellos, como se observa en la Tabla 2.

Tabla 2. Porcentajes de Componentes de Varianza para los efectos considerados en el modelo.

σ^2_L	9.27043%
$\sigma^2_{r/L}$	0.009889%
σ^2_G	1.023394%
σ^2_C	0.389070%
σ^2_{l/p_1}	0.016232%
σ^2_{l/p_2}	0.052020%
σ^2_{l/p_3}	0.1091369%
σ^2_P	10.400696%
σ^2_T	0.2453088%
$\sigma^2_{C \text{ VS } T}$	50.698657%
σ^2_{G*L}	0.4988168%
σ^2_{C*L}	0.3382691%
$\sigma^2_{(l/p_1)*L}$	0.0019943%
$\sigma^2_{(l/p_2)*L}$	0.0031023%
$\sigma^2_{(l/p_3)*L}$	0.0042103%
σ^2_{P*L}	11.561148%
σ^2_{T*L}	0.2388825%
$\sigma^2_{(C \text{ VS } T)*L}$	13.39548%
σ^2_e	0.4680147%

3 Conclusiones

En lo que respecta a este trabajo, se obtuvieron los grados de libertad incluidos en la Tabla 1 y el estadístico de prueba apropiado para probar las hipótesis nulas correspondientes del experimento de (Cuellar 1998), además se observa que fueron los efectos de localidad los que más contribuyeron en la respuesta (Tabla 2). Y esto lo que significa es que la localidad de Celaya, fué más rendidora que la de Torreón, (Cuellar 1998). Como se observa en los

cuadros anteriores, los estimadores de las componentes de varianza permiten conocer las fuentes de variación así como la magnitud relativa de los efectos en el modelo. Es evidente que la principal fuente de variación fue debido a las localidades (Tablas 1 y 2). Sin embargo, el fitomejorador requiere de la estimación de la varianza de los genotipos para predecir el comportamiento en ensayos posteriores.

Apéndice

Con el objeto de dar una idea más enfocada de lo que se esta haciendo, aquí se presentan desarrollos de las esperanzas de cuadrados medios de algunos de los efectos estimados, tales como los de localidad y repetición dentro de localidad.

Como ya se mencionó, para encontrar la esperanza de los cuadrados medios, se inicia obteniendo la esperanza de la suma de cuadrados del efecto correspondiente y después se divide dicha esperanza entre sus respectivos grados de libertad.

$$\begin{aligned} E[SC(LOC)] &= \frac{E[\sum_k Y_{..k}^2]}{rt} - \frac{E[Y_{...}^2]}{rlt} \\ &= rtl\mu^2 + rtl\sigma_L^2 + tl\sigma_{r/L}^2 + rl\sigma_G^2 + rl\sigma_{LG}^2 + l\sigma_e^2 \\ &\quad -(rtl\mu^2 + rt\sigma_L^2 + t\sigma_{r/L}^2 + rl\sigma_G^2 + r\sigma_{LG}^2 + \sigma_e^2) \\ &= rt(l-1)\sigma_L^2 + r(l-1)\sigma_{LG}^2 + t(l-1)\sigma_{r/L}^2 + (l-1)\sigma_e^2 \end{aligned}$$

$$\begin{aligned} E[\frac{SC(Loc)}{l-1}] &= E[CM(Loc)] \\ &= rt\sigma_L^2 + r\sigma_{LG}^2 + t\sigma_{r/L}^2 + \sigma_e^2 \end{aligned}$$

$$\begin{aligned} E[SC(Rep/Loc)] &= \frac{E[\sum_{jk} Y_{.jk}^2]}{t} - \frac{E[\sum_k Y_{..k}^2]}{tr} \\ &= rtl\mu^2 + rtl\sigma_L^2 + rtl\sigma_{r/L}^2 + rl\sigma_G^2 + rl\sigma_{LG}^2 + rl\sigma_e^2 \\ &\quad -(rtl\mu^2 + rtl\sigma_L^2 + tl\sigma_{r/L}^2 + rl\sigma_G^2 + rl\sigma_{LG}^2 + l\sigma_e^2) \\ &= tl(r-1)\sigma_{r/L}^2 + l(r-1)\sigma_e^2 \end{aligned}$$

$$\begin{aligned} E[\frac{SC(Rep/Loc)}{(r-1)l}] &= E[CM(Rep/Loc)] \\ &= t\sigma_{r/L}^2 + \sigma_e^2 \end{aligned}$$

Referencias

Brownlee, K.A. (1984). *Statistical Theory and Methodology, In Science and Engineering.* Second edition, Robert E. Krieger Publishing Company, Inc.

Cuellar, Ch.R. (1998). *Aptitud Combinatoria de Líneas S_2 en Cruza con tres Probadores Contrastantes.* Tesis licenciatura U.A.A.A.N.

Searle, S.R. (1987). *Linear Models for Unbalanced Data.* John Wiley and Sons, Inc. N.Y.

Searle, S.R.; Casella,G. and McCulloch, CH.E. (1992) *Variance Components.* John Wiley and Sons, Inc. N.Y.

Aplicación de Técnicas de Remuestreo para el Cálculo de Varianza en un Muestreo Complejo

Catalina Palmer Arrache Guillermina Eslava Gómez

Ignacio Méndez Ramírez

IIMAS-UNAM

1 Introducción

En México son muchas las organizaciones gubernamentales y privadas que levantan encuestas. Por razones operativas, o por la existencia de varios dominio de estudio, suele suceder que se aplican esquemas de muestreo complejos. Sin embargo, la realidad es que no siempre se llegan a estimar las varianzas de los estimadors puntuales, y en muchos casos, se analizan los datos sin considerar el diseño, lo cual lleva a intervalos de confianza erróneos, y a la planeación mal fundamentada de encuestas posteriores.

Cabe señalar los aspectos que distinguen a un diseño complejo, de acuerdo a Wolter (1985, pag. 2):

1. El grado de complejidad del diseño muestral
2. El grado de complejidad del estimador o estimadores
3. La existneica de características o variables múltiples que son de interés
4. El uso que se de a los datos de la encuesta (descriptivo y analítico)
5. El tamaño o escala de la encuesta

Aunados a los aspectos mencionados, se pueden presentar problemas por no respuesta y post-estratificación que hacen más compleja la obtención de los estimadores. Algunos métodos de remuestreo han probado ser fáciles de aplicar aún bajo estas condiciones, en particular, el Jackknife. No obstante, la aplicación de técnicas de remuestreo es especialmente útil en los casos en que interesan estimadores no lineales. A continuación se exponen 2 de las técnicas de remuestreo más importantes, una guía de decisión para usar un método u otro y una aplicación.

2 Repeticiones balanceadas ("Balanced Half-Sampling")

El método se aplica al caso en que se tienen L estratos con dos unidades primarias de muestreo en cada uno, es decir, el tamaño de muestra en cada estrato es 2, $n_h = 2$. De tal forma, si se decidiera tomar al azar un solo elemento por estrato, habría 2^L posibles formas de hacerlo.

La técnica de Repeticiones Balanceadas consiste en escoger un número menor a 2^L muestras, pero éstas no pueden ser cualesquiera muestras. Se considera,

$$\delta_h^{(i)} = \left\{ \begin{array}{c} 1 \text{ si la unidad } (h,1) \text{ está en la muestra } i\text{-ésima} \\ -1 \text{ si la unidad } (h,2) \text{ está en la muestra } i\text{-ésima.} \end{array} \right\}$$

Se seleccionan k muestras con balance ortogonal completo, lo que se cumple cuando se satisfacen:

$$\sum_{j=1}^{k} \delta_h^{(j)} \delta_{h\prime}^{(j)} = 0 \text{ (para } 1 < h < h\prime \leq L). \tag{1}$$

y

$$\sum_{j=1}^{k} \delta_h^{(j)} = 0. \ (h = 1, 2, \cdots, L). \tag{2}$$

La determinación de muestras con balance ortogonal completo se logra a través de las matrices Hadamard, y se simplifica a buscar k, tal que $k > L$; k múltiplo de 4.

El método se generaliza a cualquier estimador, ajustando los factores de expansión w_{hi}, cada réplica $(t = 1, \cdots, k)$, y la varianza se estima mediante:

$$v_k^c(\hat{\Theta}) = \sum_{j=1}^{k} \frac{(\hat{\Theta}_{(j)}^c - \hat{\Theta})}{k(1-\lambda)}, \quad \text{o bien,} \quad \bar{v}_k(\hat{\Theta}) = \frac{v_k^c(\hat{\Theta}) + v_k(\hat{\Theta})}{2}$$

Se advierte que las soluciones de muestras ortogonales no son únicas, de donde $v_k^c(\hat{\Theta})$ se consigue del complemento de la matriz Hadamard utilizada.

Krewski y Rao (1981) demostraron resultados asintóticos que permiten hacer inferencia basada en la varianza obtenida de esta manera.

3 Jackknife

Aunque los inicios del Jackknife se dieron en el marco de una población infinita, hoy en día ha probado ser robusto y consistente al aplicarse a esquemas de muestreo complejos, para calcular estadísticas que son funciones suaves de la muestra.

En un muestreo aleatorio simple, el Jackknife consiste en obtener tantos estimadores como sea el tamaño de muestra, eliminando todos los datos asociados a un conglomerado primario en un estrato. Rao (1997) sintetiza el procedimiento para llevar a cabo el Jackknife mediante un ajuste de los factores de expansión. En cada iteración, suponga que se elimina el conglomerado l del estrato k (en un estratificado bietápico) y sean w_{hij}^{kl} los factores de expansión ajustados como se indica a continuación:

$$w_{hij}^{kl} = \left\{ \begin{array}{rl} w_{hij}\ (h \neq k) & \text{Para las unidades que no están en el estrato donde se está eliminando un conglomerado.} \\ w_{hij}\frac{n_h}{n_h-1}\ (i \neq l) & \text{Para las unidades en el estrato } k \text{ pero que no son del conglomerado } l. \\ 0\ (i = l) & \text{Para las unidades del conglomerado que se elimina.} \end{array} \right\}$$

Se obtiene el estimador $\hat{\Theta}_{kl}$ de la misma forma que $\hat{\Theta}$ (quitando el conglomerado l-ésimo del estrato k), pero con los nuevos factores de expansión. Luego la varianza de $\hat{\Theta}$ se estima como

$$v_J(\hat{\Theta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{\Theta}_{(hi)} - \hat{\Theta})^2 \tag{3}$$

Al igual que en el caso de Repeticiones Balanceadas, Krewski y Rao (1981) sentaron las bases para poder hacer inferencia basada en la varianza Jackknife, cuando se trata de estadísticas suaves.

Se sabe que en el caso de m.a.s. el estimador Jackknife de la varianza de la mediana o de otra estadística no suave, no es consistente. Se sospecha que en el caso de muestreos complejos, no exista el mismo problema pues se elimina todo un conglomerado a la vez. Sobre ésto, no se han reportado resultados análiticos y hacen falta ejemplos empíricos al respecto (Rao, 1997).

4 Aplicación a la ENAL'96

La Encuesta Nacional de Alimentación y Nutrición en el Medio Rural 1996 (ENAL-96) fue realizada por el Instituto Nacional de Nutrición "Salvador Zubirán" (INNSZ) y el

apoyo de otras instituciones del gobierno. Esta perseguía el llegar a definir la magnitud y características de los problemas nutricionales en la población de niños de cinco años y menores que viven en comunidades rurales (que tienen entre 500 y 2500 habitantes) en los 31 estados de la República Mexicana (no se incluyó el D.F.).

Se realizó un diseño estratificado bietápico por conglomerados en cada estado. Los estratos se determinaron buscando regiones lo más homogéneas posibles, de acuerdo a criterios socieconómicos, geográficos, etnográficos y de continuidad territorial, entre otros. Como se da a entender, la unidad primaria de muestreo estuvo conformada por las localidades, y las unidades secundarias por las viviendas de cada localidad. En cada estrato, se seleccionó, mediante m.a.s. sin remplazo, un mínimo de dos localidades, y en cada localidad se seleccionaron de la misma forma, 50 familias, aunque no hubo respuesta en todas.

La Organización Mundial de la Salud (OMS) ha establecido tres medidas básicas para evaluar un estado nutricional. Estas son: **peso para la edad, talla para la edad** y **peso para la talla**. El cálculo de dichos indicadores se desprende de las medidas de peso y talla y la ubicación del individuo en la población de referencia que le corresponde de acuerdo a su edad y sexo, o sexo y talla, según sea el caso. Se construyen estadísticas de manera similar a como se estandariza una variable normal, por lo que éstas se conocen como los "score z" de peso para la edad (PEDZ), talla para la edad (TEDZ) y peso para la talla (PETZ). Específicamente,

$$PEDZ = \frac{\text{Peso del individuo - Mediana del peso en la población de referencia } \mathbf{A}}{\text{Desviación Estándar (del peso) en la población de referencia } \mathbf{A}}$$

$$TEDZ = \frac{\text{Talla del individuo - Mediana de la talla en la población de referencia } \mathbf{A}}{\text{Desviación Estándar (de talla) en la población de referencia } \mathbf{A}}$$

$$PETZ = \frac{\text{Peso del individuo - Mediana del peso en la población de referencia } \mathbf{B}}{\text{Desviación Estándar (del peso) en la población de referencia } \mathbf{B}}$$

La población de referencia **A** es aquella que se compone de individuos conla misma edad y sexo del sujeto a quien se está evaluando, mientras que los que forman parate de la población de referencia **B** tienen su misma talla y sexo. Las distribuciones de estas poblaciones son facilitadas por la OMS (1983); obviamente sus tallas corresponden a las estimaciones que resultaron de levantamientos de información, a nivel internacional.

Para los expertos en nutrición, la estimación de estos índices por estado no fue suficiente para esclarecer los distintos problemas nutricionales. De ahí que se creara un sistema de categorización, de acuerdo a la combinación de distintos niveles de los tres indicadores. que arrojó 5 grupus que poseen una adecuada interpretación en cuanto a estados nutricionales:

Normales (y cerca de lo normal) : Niños con una nutrición adecuada en el presente y pasado (Los criterios originales se hicieron un poco mas flexibles para evitar confusiones por etapa de crecimiento y un posbile sesgo de los parámetros internacionales).

Bajitos Gorditos : sujetos que tuvieron desnutrición en etapas anteriores pero actualmente están sobrenutridos. Probablemente los niños de este grupo tienen una dieta no balanceada. Es posible que en este grupo también estén incluídos menores con problemas fisiológicos que causan su gordura, pero no se cuenta con elementos para distinguirlos.

Mal para su edad : Los niños que han sufrido desnutrición en etapas anteriores de su vida o en la gestación pero actualmente reciben una nutrición adecuada

Mal para su talla : Se compone de niños que están bien de estatura o son más altos de lo normal, pero su peso no es adecuado para su talla y a veces tampoco para su edad. En palabras comunes, sonalatos o de estatura normal pero flacos. Se entiende que los sujetos incluídos en este grupo están sufriendo desnutrición al momento de la encuesta.

Mal para la edad y talla : Los menores aquí clasificados están sufriendo desnutrición actualmente y la sufrieron en etapas anterioes de su vida.

Se decidió utilizar un estimador de razón combinado a nivel *estado*, y un estimador separado a nivel *Nacional*. El estimador de razón combinado para la proporción de niños menores de seis años que viven en comunidades rurales y son clasificados en la categoría c, en un Estado E, de acuerdo al diseño muestral de la encuesta está dado por:

$$\hat{R}_E^c = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{h_i}} FLOC_h FVIV_{hi} Fcasa_{hij} y_{hij}^c}{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{h_i}} FLOC_h FVIV_{hi} x_{hij}} \tag{5}$$

FLOC, FVIV, son los factores de expansión por localidad y vivienda:

$$FLOC = \frac{\text{locs. en estrato}}{\text{Locs. encuestadas en estr.}}$$

$$FVIV = \frac{\text{viviendas en loc.}}{\text{viv. encuestadas en loc.}}$$

$F\ casa$ es el factor de expansión de los niños dentro de cada casa, y se obtiene de forma similar a los anteriores:

$$F\ casa = \frac{\text{Niños de 5 años y menores en la casa}}{\text{Niños encuestados en la casa}}$$

y^c_{hij} es el número de niños encuestados en la casa, clasificados en la categoría c, y x_{hij} es el total de menores de 6 años que residen en la casa.

Las varianzas se obtuvieron por Jackknife y por linearización (fórmulas). Se muestran algunos resultados numéricos en las Tablas 1 y 2.

Tabla 1: Resultados a nivel Nacional de las estimadores Jackknife y el de fórmula.

Grupo	Est. Jack.	Est. Fórmula	Var(Jack.)	Var(Fórm.)
Normales y casi N.	0.5267	0.5270	2.73E-05	2.75E-05
Bajitos Gorditos	0.2385	0.2384	1.84E-05	1.71E-05
Mal para edad	0.1405	0.1403	1.34E-05	1.24E-05
Mal para talla	0.0706	0.0705	7.42E-06	5.59E-06
Mal edad y talla	0.0238	0.0236	1.81E-06	1.72E-06

5 Discusión y conclusiones

Aún cuando en muchos libros de texto se presentan estimadores de la varianza de un estimador de razón, debe estar claro que tales fórmulas son una aproximación obtenida por linearización. El lector que nunca se ha enfrentado al análisis de una encuesta de gran cobertura, puede pensar que el cálculo de varianza no es tanto problema, pero la realidad de los Centros que trabajan haciendo encuestas dice lo contrario. Es especialmente difícil en la práctica asegurar que el trabajo de los computólogos en verdad se apegan al cálculo necesario, además de que se da una gran diversidad de estimadores de interés. Por otra parte, los métodos de remuestreo también son útiles si se entra en una etapa anlítica con análisis multivariados.

La experiencia de esta aplicación mostró que el Jackknife produjo estimadores puntuales y de varianza muy cercnos a la linearización, o a los estimadores obtenidos por fórmula. Sin embargo resulta mucho más rápido y confiable el hacer un programa basado en Jackknife que la aplicación mediante fórmula. Existen programas que calculan varianzas de muestreos complejos por linearización, como lo es el PcCARP, sin embargo, cuando hay problemas de no- respuesta y post-estratificación éstos no suelen ser convenientes. Además algunos centros de encuestas consideran que no es sencillo ajustar sus bases de datos a los requerimientos de los programas. Agencias gubernamentales de otros países como Canadá han preferido hacer su paquetería (llamada JACJVAR) basada en Jackknife. Una buena alternativa parece ser el WESVAR, que va a ser comercializado próximamente.

Un cálculo de varianza adecuado resulta necesario para la planificación de futuras encuestas, en lo que radica la importancia de la búsqueda de métodos de solución bien

fundamentados pero a la vez accesibles a los usuarios. En México es importante el comenzar a difuendir estas técnicas, para mejorar la calidad de las encuestas, las inferencias que de ellas se arrojen, y su utilidad en la planeación de programas de diversa índole.

Referencias

Krewski, D. y Rao, J.N.K. (1981). Inference from Stratified Samples: Proporties of the Linearization, Jackknife and Balanced Repeated Replication Methods. *The Annals of Statistics*, 9, 1010-1019.

Rao, J.N.K. (1997). Developments in Sample Survey Theory: An Appraisal. *The Canadian Journal of Statistics*, 25, 1-21.

Wolter, K.M. (1985). *Introduction to Variance Estimation.* Springer-Verlag. New York.

Intervalos de Confianza de la Encuesta sobre Migración en la Frontera Norte de México (EMIF). Nota Metodológica, Fases II y III

Rafael Pérez Abreu C. y Ignacio Méndez Gómez-Humaran
CIMAT, Aguascalientes *El Colegio de la Frontera Norte*

1 Introducción

El Colegio de la Frontera Norte en conjunto con El Consejo Nacional de Población y Vivienda y La Secretaría del Trabajo y Previsión Social han realizado desde el año de 1993 la Encuesta sobre Migración en la Frontera Norte de México. El objetivo de esta encuesta es caracterizar al migrante y al fenómeno de la migración desde su perspectiva social, midiendo el flujo migratorio en ambos sentidos (Sur-Norte y Norte-Sur) a lo largo de las localidades mexicanas de frontera con los Estados Unidos de Norte América.

Población Objetivo.

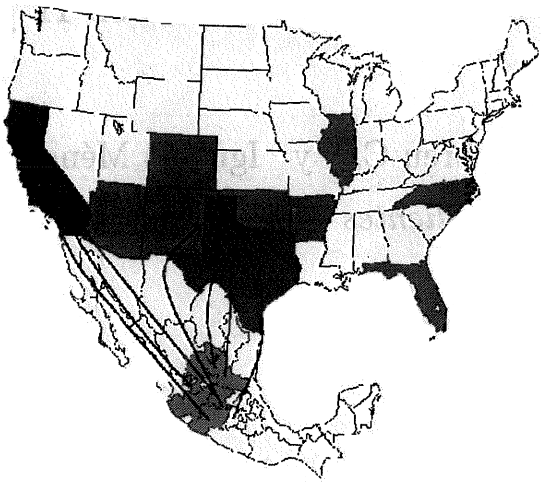
La población objetivo bajo estudio son los individuos mayores de 12 años de edad, no nacidos en Estados Unidos (mexicanos y algunos centro americanos), no residentes de la ciudad de aplicación de la entrevista, que cuya estancia en la zona fronteriza es para trabajar o buscar trabajo, para visitar familiares o amigos o de negocios.

Dependiendo de la dirección de las unidades de observación, ya sean sur-norte, o unidades en dirección norte-sur, la definición puede variar ligeramente; si hablamos en forma estricta. Sin embargo, en esencia son los mismos individuos del fenómeno de migración internacional, es decir, migrantes mexicanos hacia los Estados Unidos de Norte América.

Marco Muestral.

Debido a la misma naturaleza del estudio, el diseño muestral requiere de la elaboración de un marco muestral adecuado a las circunstancias (ad hoc). El marco muestral se construyó previo al trabajo de campo considerándose dos dimensiones: la primera dimensión es el espacio geográfico, el cual se refiere a las diferentes áreas geográficas (localidades fronterizas), en donde el flujo de migrantes internacionales puede ser observado, y la segunda dimensión se refiere al intervalo de tiempo cuando puede ser observado el fenómeno.

De todas las posibles combinaciones de estas dos dimensiones se obtiene un marco muestral en la dimensión espacio-tiempo como sigue:



Dimensión Espacio.

La dimensión espacio consiste de tres Regiones:

1. - Región Este: Formada por Matamoros, Nuevo Laredo y Piedras Negras.
2. - Región Centro: Formada por Ciudad Juárez y Nogales.
3. - Región Oeste: Formada por Mexicali y Tijuana.

A su vez las Regiones están divididas en Localidades (las cuales son las ciudades de muestreo), quienes a su vez están divididas en zonas de muestreo (que son las zonas dentro de las ciudades como tales como: terminal aérea, central camionera, etc.), quienes a su vez están divididas en puntos de muestreo (puntos específicos dentro de las zonas de muestreo).

Dimensión Tiempo.

La unidad de tiempo sobre la que se desea hacer inferencia es el trimestre. Este se compone de 92 días, en los cuales se puede observar el flujo.

Dimensión Espacio-Tiempo.

Como se mencionó anteriormente, todas las posibles combinaciones de Espacio-x-Tiempo nos proporcionan un marco muestral de puntos de muestreo Espacio-Tiempo.

Tipo de Muestreo.

Antes de seleccionar los puntos de muestreo Espacio-Tiempo, se realizó un conteo de individuos en los puntos de muestreo Espacio-Tiempo, es decir, en los diferentes puntos de muestreo geográficos y en los diferentes días de la semana (dimensión tiempo). Dicha información nos proporciona la distribución porcentual del flujo de individuos en el marco muestral Espacio-Tiempo.

El tipo de muestreo primeramente fue estratificado sobre cada una de las regiones, es decir, podemos considerar a cada región como un diseño muestral por separado.

Selección de la muestra.

Dentro de cada región el tipo de muestreo fue polietápico, seleccionando sólo un elemento (o grupo) en cada una de las etapas en forma proporcional al flujo de individuos observado, hasta llegar al punto muestra Espacio-Tiempo.

Estimación de las principales variables.

En la Fase I de este estudio se construyeron los intervalos de confianza haciendo uso del paquete estadístico Clusters. Sin embargo, debido a las dificultades que éste presenta para su ejecución y compilación se buscaron otras alteratívas para las Fases II y III del proyecto. A continuación se describe el procedimiento utilizado para calcular los intervalos de Confianza de las Fases II y III de la Encuesta sobre migración en la Frontera Norte de México (EMIF), así como algunas diferencias con las estimaciones de la Fase I.

Diferencias de Estimación con la Fase I.

El procedimiento utilizado para la construcción de los intervalos de confianza de la fase II y III de la EMIF difieren de la Fase I en el siguiente aspecto:

Para la Fase I se utilizó el paquete de cómputo estadístico llamado Clusters, mientras que para la Fase II y III se utilizó el paquete de cómputo estadístico WesVarPC (1997) de la WesStat, Inc.

La razones de este cambio obedecen a la siguiente ventaja que WesVarPC tiene sobre el paquete Clusters: al comparar estos software entre sí, se tiene: Mientras que Clusters es un paquete de cómputo poco amigable, desarrollado en 1986 en Fortran 77 para PC, el cual funciona a través de un archivo de comandos por lotes, con una sintaxis rígida, además, con un mínimo de descripciones sobre los errores de programación y con tan solo un breve manual, el paquete WesVarPC es un programa desarrollado en medio Windows (amigable) con múltiples facilidades y descripciones de los procedimientos. Adicionalmente, WesVarPC tiene un modulo de creación de tablas dinámicas compatible con hojas de calculo Excel, el cual permite transferir los resultados directamente sobre un formato de tabulados a Excel, tal y como aparecen al final de este reporte.

Presentación de WesVarPC.

WesVarPC es un paquete de cómputo estadístico para PC desarrollado en 1997 por la compañía WesStat, Inc. Este software puede realizar estimaciones de medias, porcentajes, razones, razón de diferencias y logaritmo de razón de momios. Estas estimaciones se basan en datos muestrales procedentes de muestreos con procedimientos y diseños de encuestas complejas. Este software es flexible, ya que soporta un amplio rango de diseños de muestra complejos, incluyendo muestreos polietápicos con probabilidad de selección diferente y muestreos estratificados entre otros.

WesVarPC puede realizar cálculos de errores muestrales, varianzas e intervalos, de confianza, etc., para estimadores de encuestas especificados por el usuario.

La especificación del diseño de muestra en WesVarPC se realiza describiendo la estructura del diseño muestral, detallando las etapas del muestreo y la ponderación de cada unidad muestral, definida esta última como el inverso de la probabilidad de selección.

2 Diseño de muestra

La estructura global de la muestra para las fases II y III de la EMIF fueron:

Para cada Trimestre del año (1,2,3,4) y para cada subgrupo de población de la encuesta (Procedentes del Sur, Frontera, EUA y Deportados) de la encuesta, la estructura del diseño muestral es la siguiente:

Etapas Corte Grupos/Conglomerados
Región de muestreo Geográfico Estratos
Ciudad de muestreo Geográfico Unidad Primera de Selección
Zona de muestreo Geográfico
Punto de muestreo Geográfico
Turno de muestreo Temporal
Día de la semana Temporal

Estimación de la Varianza.

A continuación se describe una breve descripción del procedimiento utilizado por WesVarPC para el cálculo de la varianza.

WesVarPC puede utilizar varios métodos para calcular los errores muestrales, el método utilizado en este muestreo polietápico por conglomerados fue el procedimiento conocido como Jackknife. La idea principal del método consiste en construir réplicas y calcular el parámetro estimado de interés de la muestra completa, así como de cada una de las submuestras. La cantidad de variación de los estimadores de la submuestra se utiliza para estimar la varianza de la muestra completa. El estimador de Varianza, $Var(\hat{\theta})$, generalmente toma la forma:

$$Var\left(\hat{\theta}\right) = c \cdot \sum_{k=1}^{G} \left(\hat{\theta}_{(k)} - \hat{\theta}\right)^2$$

Donde:

θ Es el parámetro de interés.

$\hat{\theta}$ Es el estimador de q basado en la muestra completa.

$\hat{\theta}_{(k)}$ Es la k-ésima estimador de q basado en las observaciones incluidas en la k-ésima réplica.

G es el número total de réplicas formadas.

c es una constante que depende del método de replicación.

$Var(\hat{\theta})$ es el estimado de la varianza de $\hat{\theta}$.

Ventajas y Desventajas del Método.

La principal ventaja del método es que no es necesario ningún tipo de supuestos en lo que se refiere a la distribución del estimador, dado que se construye una distribución empírica del estimador, por lo que se considera una técnica no paramétrica con respecto a la distribución del estimador. La principal desventaja, desde el punto de vista de la precisión, es que el método Jackknife tiende a ser conservador en el sentido de que el valor esperado de la varianza es más grande que la varianza verdadera de la población. (Ver. Efron, 1982), The Jackknife, The Bootstrap and other resampling plans, edited by Society for Industrial and applied Mathematics [SIAM] CBMS 38 by Arrowsmith, Ltd, Bristol England).

En términos prácticos el significado del párrafo anterior es que: las estimaciones presentadas para la fase II y III de la EMIF aunque corresponden a un nivel del 95% de confianza, en realidad este nivel mayor al 95%. Varios ejercicios comprobatorios bajo el supuesto de que la EMIF es una muestra polietápica con probabilidades de selección proporcional (ppt) al número de elementos de las unidades geográficas y temporales, muestran que los intervalos de confianza calculados al 95% vía el procedimiento Jackknife corresponden a un nivel de confianza superior al 99% si se supone muestreo ppt. y una distribución normal del estimador.

INTERVALOS DE CONFIANZA LA POBLACION SEGUN CARACTERISTICAS SOCIOECONOMICAS DE LOS MIGRANTES PROCEDENTES DEL SUR (14-dic. 94 al 13-dic.-95)

CARAC. SOCIOECONOMICAS	TRIMESTRE 1			TRIMESTRE 2			TRIMESTRE 3			TRIMESTRE 4			TOTAL		
PORCENTAJES	ESTIMADO	INFERIOR	SUPERIOR	ESTIMADO	INFERIOR	SUPERIOR	ESTIMADO	INFERIOR	SUPERIOR	ESTIMADO	INFERIOR	SUPERIOR	ESTIMADO	INFERIOR	SUPERIOR
ESCOLARIDAD PROMEDIO															
REGION 1	6.9	6.1	7.7	6.6	6.0	7.3	7.1	6.1	8.1	7.5	6.9	8.1	6.9	6.6	7.3
REGION 2	6.1	5.4	6.8	7.3	6.6	8.0	7.1	6.5	7.8	8.1	6.3	9.8	7.1	6.6	7.5
REGION 3	7.3	6.6	8.0	6.3	6.1	6.6	7.2	6.7	7.8	7.2	6.6	7.8	7.0	6.7	7.3
TOTAL	7.0	6.5	7.4	6.6	6.2	7.0	7.2	6.8	7.5	7.4	7.0	7.9	7.0	6.8	7.2
EDAD PROMEDIO															
REGION 1	28.7	23.8	33.6	30.5	28.9	32.1	29.5	27.7	31.3	30.4	28.4	32.4	29.7	28.4	31.1
REGION 2	27.2	25.2	29.3	29.6	27.4	31.8	26.3	25.2	27.3	29.0	26.5	31.6	27.7	26.7	28.6
REGION 3	27.8	26.7	28.9	29.3	27.1	31.5	28.1	27.0	29.3	28.5	26.2	30.9	28.4	27.7	29.1
TOTAL	27.8	26.8	28.9	29.7	28.4	30.9	27.8	27.0	28.6	29.0	27.4	30.5	28.5	28.0	29.0
# CRUCES PROM.PARA TRABAJAR EN E.U.															
REGION 1	3.0	1.8	4.1	2.4	1.5	3.4	1.9	0.7	3.0	1.3	0.7	2.0	2.3	1.7	2.8
REGION 2	1.8	0.8	2.7	2.3	1.4	3.2	0.9	0.7	1.1	0.7	0.3	1.2	1.4	1.1	1.8
REGION 3	1.9	1.3	2.6	2.0	1.5	2.6	1.2	0.7	1.6	0.8	0.3	1.3	1.6	1.2	1.9
TOTAL	2.1	1.6	2.6	2.2	1.8	2.6	1.2	0.9	1.5	0.9	0.5	1.2	1.7	1.4	1.9
TOTAL DE MIGRANTES															
REGION 1	69,347	16,835	121,859	80,403	29,677	131,129	63,130	23,100	103,159	38,707	14,083	63,331	251,587	170,150	333,023
REGION 2	68,115	20,756	115,474	66,990	9,915	124,065	108,680	47,868	169,493	40,482	11,361	69,603	284,267	189,869	378,666
REGION 3	196,650	32,252	361,048	167,398	5,958	328,837	139,127	13,973	264,281	119,866	289	239,442	623,040	344,760	901,321
TOTAL	334,111	158,644	509,578	314,791	139,517	490,065	310,937	170,352	451,523	199,055	75,274	322,835	1,158,894	881,093	1,436,696

NOTAS

1.- La región 1 (Este) Comprende las ciudades de Piedras Negras, Nuevo Laredo Matamoros y Reynosa
La región 2 (Centro) Comprende las ciudades de Nogales, y Cd. Juárez.
La región 3 (Oeste) Comprende las ciudades de Tijuana y Mexicali.

2- La Zona Norte comprende los Estados de Baja California, Baja California Sur, Coahuila, Chihuahua, Nuevo León, Sinaloa, Sonora y Tamaulipas
La Zona Tradicional comprende los estados de Aguascalientes, Colima, Durango, Guanajuato, Jalisco, Michoacán, Nayarit, San Luis Potosí y Zacatecas.
La Zona Centro comprende los estados del Distrito Federal, Guerrero, Hidalgo, México, Morelos, Oaxaca, Puebla, Querétaro y Tlaxcala.
La Zona Sureste comprende los estados de Campeche, Chiapas, Quintana Roo, Tabasco, Veracruz y Yucatán

3.- Trimestre 1 del 14 de diciembre de 1994 al 13 de marzo de 1995.
Trimestre 2 del 14 de marzo de1995 al 13 de junio de 1995.
Trimestre 3 del 14 de junio de 1995 al 13 de septiembre de 1995.
Trimestre 4 del 14 de septiembre de 1995 al 13 de diciembre de 1995.

FUENTE: El Colegio de la Frontera Norte, Consejo Nacional de Población y Vivienda, Secretaría del Trabajo y Previsión Social.
Encuesta sobre Migración en la Frontera Norte de México del 14 de diciembre de 1994 al 13 de diciembre de 1995.

INTERVALOS DE CONFIANZA LA POBLACION SEGUN CARACTERISTICAS SOCIOECONOMICAS DE LOS MIGRANTES PROCEDENTES DEL SUR (14-dic. 94 al 13-dic.-95)

CARAC. SOCIOECONOMICAS	TRIMESTRE 1			TRIMESTRE 2			TRIMESTRE 3			TRIMESTRE 4			TOTAL		
PORCENTAJES	ESTIMADO	INFERIOR	SUPERIOR	ESTIMADO	INFERIOR	SUPERIOR	ESTIMADO	INFERIOR	SUPERIOR	ESTIMADO	INFERIOR	SUPERIOR	ESTIMADO	INFERIOR	SUPERIOR
ESCOLARIDAD PROMEDIO															
REGION 1	6.9	6.1	7.7	6.6	6.0	7.3	7.1	6.1	8.1	7.5	6.9	8.1	6.9	6.6	7.3
REGION 2	6.1	5.4	6.8	7.3	6.6	8.0	7.1	6.5	7.8	8.1	6.3	9.8	7.1	6.6	7.5
REGION 3	7.3	6.6	8.0	6.3	6.1	6.6	7.2	6.7	7.8	7.2	6.6	7.8	7.0	6.7	7.3
TOTAL	7.0	6.5	7.4	6.6	6.2	7.0	7.2	6.8	7.5	7.4	7.0	7.9	7.0	6.8	7.2
EDAD PROMEDIO															
REGION 1	28.7	23.8	33.6	30.5	28.9	32.1	29.5	27.7	31.3	30.4	28.4	32.4	29.7	28.4	31.1
REGION 2	27.2	25.2	29.3	29.6	27.4	31.8	26.3	25.2	27.3	29.0	26.5	31.6	27.7	26.7	28.6
REGION 3	27.8	26.7	28.9	29.3	27.1	31.5	28.1	27.0	29.3	28.5	26.2	30.9	28.4	27.7	29.1
TOTAL	27.8	26.8	28.9	29.7	28.4	30.9	27.8	27.0	28.6	29.0	27.4	30.5	28.5	28.0	29.0
# CRUCES PROM.PARA TRABAJAR EN E.U.															
REGION 1	3.0	1.8	4.1	2.4	1.5	3.4	1.9	0.7	3.0	1.3	0.7	2.0	2.3	1.7	2.8
REGION 2	1.8	0.8	2.7	2.3	1.4	3.2	0.9	0.7	1.1	0.7	0.3	1.2	1.4	1.1	1.8
REGION 3	1.9	1.3	2.6	2.0	1.5	2.6	1.2	0.7	1.6	0.8	0.3	1.3	1.6	1.2	1.9
TOTAL	2.1	1.6	2.6	2.2	1.8	2.6	1.2	0.9	1.5	0.9	0.5	1.2	1.7	1.4	1.9
TOTAL DE MIGRANTES															
REGION 1	69,347	16,835	121,859	80,403	29,677	131,129	63,130	23,100	103,159	38,707	14,083	63,331	251,587	170,150	333,023
REGION 2	68,115	20,756	115,474	66,990	9,915	124,065	108,680	47,868	169,493	40,482	11,361	69,603	284,267	189,869	378,666
REGION 3	196,650	32,252	361,048	167,398	5,958	328,837	139,127	13,973	264,281	119,866	289	239,442	623,040	344,760	901,321
TOTAL	334,111	158,644	509,578	314,791	139,517	490,065	310,937	170,352	451,523	199,055	75,274	322,835	1,158,894	881,093	1,436,696

NOTAS

1.- La región 1 (Este) Comprende las ciudades de Piedras Negras, Nuevo Laredo Matamoros y Reynosa

La región 2 (Centro) Comprende las ciudades de Nogales, y Cd. Juárez.

La región 3 (Oeste) Comprende las ciudades de Tijuana y Mexicali.

2- La Zona Norte comprende los Estados de Baja California, Baja California Sur, Coahuila, Chihuahua, Nuevo León, Sinaloa, Sonora y Tamaulipas

La Zona Tradicional comprende los estados de Aguascalientes, Colima, Durango, Guanajuato, Jalisco, Michoacán, Nayarit, San Luis Potosí y Zacatecas.

La Zona Centro comprende los estados del Distrito Federal, Guerrero, Hidalgo, México, Morelos, Oaxaca, Puebla, Querétaro y Tlaxcala.

La Zona Sureste comprende los estados de Campeche, Chiapas, Quintana Roo, Tabasco, Veracruz y Yucatán

3 - Trimestre 1 del 14 de diciembre de 1994 al 13 de marzo de 1995.

Trimestre 2 del 14 de marzo de1995 al 13 de junio de 1995.

Trimestre 3 del 14 de junio de 1995 al 13 de septiembre de 1995.

Trimestre 4 del 14 de septiembre de 1995 al 13 de diciembre de 1995.

FUENTE: El Colegio de la Frontera Norte, Consejo Nacional de Población y Vivienda, Secretaría del Trabajo y Previsión Social.

Encuesta sobre Migración en la Frontera Norte de México del 14 de diciembre de 1994 al 13 de diciembre de 1995.

3 Resultados

Los cuadros anteriores son algunos de los intervalos de confianza obtenidos a través de esta metodología:

Referencias

Efron, (1982). *The Jackknife, The Bootstrap and other resampling plans*, edited by Society for Industrial and applied Mathematics. Bristol England

Comparación del Modelo Beta-Binomial con Métodos Alternativos para el Estudio de Preferencias sobre dos Opciones

Gustavo Ramírez Valverde y Candelario Méndez Olán

Colegio de Posgraduados, Montecillo, Texcoco.

1 Introducción

En Entomología, particularmente sobre el comportamiento de insectos, es de interés estudiar la preferencia de insectos sobre la elección entre dos opciones, para lo cual, los insectos son aislados unos de otros para que elijan libremente una de las dos opciones o atrayentes. Frecuentemente se da la situación de que algunas condiciones ambientales afectan a la fisiología del insecto y finalmente su respuesta en el experimento. Ante esto se diseña el experimento en bloques, por ejemplo en días distintos (cada día conforma un bloque). En caso de que exista efecto de bloque, las condiciones para cada bloque son diferentes, dando lugar a que la probabilidad de éxito P varíe en cada bloque, esto genera el fenomeno llamado sobre-dispersión. McCullagh y Nelder (1989) afirman que la sobre-dispersión está presente cuando la varianza de la respuesta de interés excede a la varianza predicha bajo el modelo supuesto.

Para probar la hipótesis de no preferencia $H_0 : P_1 = P_2$, se han usado pruebas como la t pareada, Wilcoxon con signo y Monte Carlo, a pesar de que no cumplen con los supuestos básicos para sus aplicaciones. La prueba Monte Carlo requiere que P_1 sea igual a $\frac{1}{2}$ en cada bloque, al igual que la Wilcoxon con signo para que la distribución de las diferencias de los pares (X_i, Y_i) sea simétrica. Si P_1aumenta entonces $P_2 = 1 - P_1$ disminuirá, es decir, existe interacción bloque y tratamiento (sobre-dispersión) violando el supuesto de la t pareada de no interacción bloque y tratamiento.

En este trabajo, se propone el modelo Beta-Binomial para modelar sobre-dispersión y se construye la prueba de razón de verosimilitudes generalizada para probar la hipótesis de no preferencia. A través de un estudio de simulación se observa el comportamiento del modelo propuesto en tamaño y potencia, y se compara con las pruebas estadísticas más usadas en las investigaciones de este tipo en Entomología.

2 Modelo Beta-Binomial

El modelo Beta-Binomial ha sido utilizado para modelar sobre-dispersión en la distribución binomial (Williams, 1975). El modelo Beta-Binomial consiste en suponer que la probabilidad (p) de éxito en cada binomial (bloque) varía, esto es, p la probabilidad de éxito es una realización de una variable aleatoria P y que esta variable tiene distribución beta.

Si Y_i es el número de éxitos en el bloque i, entonces Y_i dado $P = p$ tendrá una distribución binomial con parámetros n_i y $P = p$, donde n_i es el número de unidades experimentales en el bloque i. Esto es,

$$\mathrm{p}[Y_i|P=p] = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) p^{y_i} (1-p)_i^{n_i - y_i} \quad , y_i = \mathrm{U}, 1, ..., n_i$$

Si Y_i dado P tiene una distribución binomial con parámetros n_i y P, y además P tiene una distribución Beta con parámetros a y b entonces la distribución marginal de Y_i es

$$\mathrm{f}(y_i) = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \frac{\prod_{r=0}^{y_i-1} [\pi + r\theta] \prod_{r=0}^{n_i-y_i-1} [1-\pi+r\theta]}{\prod_{r=0}^{n_i-1} [1+r\theta]},$$

donde π= a/(a + b) y θ= 1 / (a + b).

A la distribución anterior se le conoce como Beta-Binomial. La hipótesis nula de no preferencia a probar corresponde a $\mathrm{H}_0 : \pi = 1/2$ vs $\mathrm{H}_0 : \pi \neq 1/2$ ya que π es el valor esperado de P y esto significa que en promedio las p's tienen un valor de $\frac{1}{2}$ no existiendo preferencia por ninguna opción. Para probar esta hipótesis se propone usar la prueba asintótica de la razón de verosimilitud generalizada para la cual es necesaria la función de logverosimilitud de la Beta-Binomial dada por:

$$\mathrm{l(n,}\theta) = \sum_{i=1}^{m} \left\{ \ln \left(\begin{array}{c} n_i \\ y_i \end{array} \right) + \sum_{r=0}^{y_i-1} \ln(\pi + r\theta) + \sum_{r=0}^{n_i-y_i-1} \ln(1 - \pi + r\theta) - \sum_{r=0}^{n_i-1} \ln(\pi + r\theta) \right\},$$

cuyas primeras derivadas parciales con respecto a π y θ son:

$$\frac{\partial l(\pi,\theta)}{\partial \pi} = \sum_{i=1}^{m} \left\{ \sum_{r=0}^{y_i-1} \frac{1}{\pi + r\theta} - \sum_{r=0}^{n_i-y_i-1} \frac{r}{1-\pi+r\theta} \right\}$$

y

$$\frac{\partial l(\pi,\theta)}{\partial \theta} = \sum_{i=1}^{m} \left\{ \sum_{r=0}^{y_i-1} \frac{r}{\pi+r\theta} - \sum_{r=0}^{n_i-y_i-1} \frac{r}{1-\pi+r\theta} - \sum_{r=0}^{n_i-1} \frac{r}{1+r\theta} \right\},$$

de donde se nota que estas ecuaciones no son lineales con respecto a π y θ. Para una aproximación numérica a la solución de ambas ecuaciones respecto de y se utiliza el método de Newton-Raphson.

El estadístico de prueba de la razón de verosimilitud es - 2 ln λ ~χ^2, donde λ = L(Ω_o) /L(Ω) (cociente de verosimilitudes de la Beta-Binomial), con

$$\Omega_o = \{(\pi, \theta)| = \pi = \tfrac{1}{2} \text{ y } \theta \geq 0\}$$

bajo H_0 y

$$\Omega = \{(\pi,\theta)|0<\pi < \tfrac{1}{2} \text{ y } \theta \geq 0 \}$$

bajo el espacio paramétrico irrestricto. Si es el valor estimado entonces la regla de decisión de la prueba de razón de verosimilitud establece que se rechace la hipótesis H_0 : = $\frac{1}{2}$ al nivel de significancia α si y sólo si

$$-2\ln\widehat{\lambda} = -2(lo - l) \geq \chi^2_{1-\alpha} \tag{1}$$

donde $\chi^2_{1-\alpha}$ es el cuantil 1-α de la distribución χ^2 con 1 grado de libertad y, l_0 y l son las logverosimilitudes evaluadas en los conjuntos Ω_0 y Ω, respectivamente.

3 Estudio de simulación

Para analizar el comportamiento del modelo Beta-Binomial y compararlo con las pruebas más usadas por los entomólogos, se realizó un estudio de simulación para estimar el tamaño y potencia de cada una de las pruebas, el cual se describe a continuación.

Se generaron 1000 muestras pseudo-aleatorias de tamaño n, con m repeticiones cada una, de la distribución Beta-Binomial para cada una de las combinaciones posibles de los factores:

1. Tamaño de la Beta-Binomial. Se ensayaron tres diferentes tamaños:

   a) n = 10 b) n = 20 c) n = 40.

2. Número de ensayos Bernoullis en cada binomial. Se ensayaron tres diferentes números de repeticiones: a) m = 5 b) m = 10 c) m = 20.

3. Tamaño del efecto: a) momio = 1.0 b) momio = 1.5 c) momio = 2.0 d) momio = 2.5.

4. Magnitud de sobre-dispersión: a) = 0.0 b) = 0.1 c) = 0.2 d) = 0.3 e) = 0.4.

   Para el caso donde el momio es igual a 1.0, esto significa que no hay preferencia por ninguna opción, es decir, P = $\frac{1}{2}$ para cada opción y, cuando θ es igual a 0 significa que no hay sobre-dispersión.

5. Las pruebas alternativas estudiadas fueron: a) t pareada b) Wilcoxon con signo c) Monte Carlo.

   El nivel de significancia usado en cada simulación para cada una de las combinaciones de los factores en estudio fue α= 0.05.

4 Resultados y discusión

En la Tabla 1 se observa una tendencia general de crecimiento en los tamaños de prueba estimados, $\widehat{\alpha}$, de cada una de las pruebas conforme la sobre-dispersión, θ, aumenta; se tiene el mismo comportamiento de crecimiento cuando m aumenta con n y θ fijos. En general, este comportamiento ocurre para los demás casos estudiados.

Tabla 1. Tamaños de prueba estimados para las pruebas estudiadas con tamaño de binomial n=10 y repeticiones m=5, 10 y 20.

T	R	Prueba	θ				
(n)	(m)		0	0.1	0.2	0.3	0.4
10	5	Beta-Binomial	.009	.025	.038	.046	.044
		t apareada	.002	.009	.021	.017	.028
		Wilcoxon con signo	.124	.127	.130	.135	.112
		Monte Carlo	.025	.107	.179	.208	.263
10	10	Beta-Binomial	.023	.048	.053	.056	.071
		t apareada	.014	.034	.036	.047	.060
		Wilcoxon con signo	.015	.037	.031	.038	.051
		Monte Carlo	.038	.103	.191	.227	.273
10	20	Beta-Binomial	.023	.049	.056	.081	.059
		t apareada	.017	.045	.050	.072	.045
		Wilcoxon con signo	.017	.042	.045	.067	.040
		Monte Carlo	.023	.109	.158	.238	.266

T = Tamaño de la binomial; R = Repeticiones de cada binomial.

Cuando no hay sobre-dispersión, $\theta = 0$, es evidente que la Monte Carlo tiene tamaños de prueba más cercanos al nivel de significancia usado (0.05); en cambio, las otras pruebas son muy conservadoras. Sin embargo, cuando existe sobre-dispersión este método se comporta de una manera muy liberal.

Nótese que si existe sobre-dispersión, $\theta > 0$, la prueba t es muy conservadora con n = 10 y m = 5 pero que tiene tamaños de prueba razonables al aumentar m. Similarmente, la Wilcoxon con signo es muy liberal cuando n = 10 y m = 5, pero mejora sus tamaños de prueba al aumentar m.

El modelo Beta-Binomial en presencia de sobre-dispersión tiene tamaños de prueba estimados aceptables. Salvo en algunos casos, este modelo tiende a ser liberal, por ejemplo, con n = 10 y m = 20 tiene un $\widehat{\alpha}$= 0.081 para θ= 0.3.

En el Cuadro 2 se nota cómo para cada momio fijo, las potencias de las pruebas tienden a disminuir conforme θ aumenta, y para cada θ fijo las potencias tienden a crecer cuando el momio crece. Nótese que la prueba Monte Carlo proporciona las potencias más grandes para cada combinación de momio y de θ, pero de la Tabla 1 se observó que este procedimiento sólo tuvo tamaños de prueba aceptables en ausencia de sobre-dispersión. De este modo, en ausencia de sobre-dispersión la prueba Monte Carlo es la más adecuada para ambos criterios: tamaño de prueba y potencia de prueba. En general, este comportamiento ocurre para los demás casos estudiados.

Tabla 2. Potencias estimadas para las pruebas estudiadas con tamaño de binomial n=10 y número de repeticiones m=20.

Momio	Prueba	θ				
		0	0.1	0.2	0.3	0.4
1.5	Beta-Binomial	.722	.528	.428	.369	.353
	t apareada	.700	.505	.406	.335	.327
	Wilcoxon con signo	.670	.488	.401	.324	.322
	Monte Carlo	.744	.688	.667	.673	.645
2.0	Beta-Binomial	.993	.929	.841	.774	.713
	t apareada	.992	.917	.825	.758	.678
	Wilcoxon con signo	.991	.909	.823	.749	.671
	Monte Carlo	.995	.966	.959	.932	.913
2.5	Beta-Binomial	1	.995	.980	.947	.904
	t apareada	1	.992	.977	.941	.888
	Wilcoxon con signo	1	.992	.973	.933	.886
	Monte Carlo	1	.999	.997	.990	.984

Nótese que las potencias de las pruebas Beta-Binomial, t pareada y Wilcoxon con signo son bastante buenas, teniendo la Beta-Binomial las potencias más grandes de las tres, pero para efectos muy grandes (momio = 2.5) la diferencia entre las potencias tiende a disminuir. Recuérdese que del Cuadro 1 se observó que los tamaños de prueba estimados para el modelo Beta-Binomial fueron los mejores que las pruebas restantes en presencia de sobre-dispersión. Por lo tanto, el modelo Beta-Binomial resulta ser más adecuado para

ambos criterios: tamaño de prueba y potencia de prueba. En general, este comportamiento ocurre para los demás casos estudiados.

5 Conclusiones

De acuerdo con los resultados, se concluye que en ausencia de sobre-dispersión la prueba Monte Carlo resultó ser mejor en ambos criterios: tamaño de prueba y poder de prueba. Sin embargo, en presencia de sobre-dispersión el modelo Beta-Binomial fue mejor para los criterios ya mencionados.

Referencias

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, 2nd ed., 506 p.

Williams, D. A. (1975). The Analysis of Binary Responses from *Toxicological Experiments Involving Reproduction and Teratogenicity.* Biometrics 31, 949 - 952.

La Regresión Isotónica Aplicada al Monitoreo de la Media de un Proceso

José G. Ríos Alejandro
Depto. de Matemáticas
ITESM Campus Monterrey

Jesús S. Arreola Risa
Depto. de Ingeniería Industrial
ITESM Campus Monterrey

Joseph J. Pignatiello Jr.
Dep. of Industrial Engineering
Florida State University

1 Introducción

Las cartas de control estadístico de procesos son importantes herramientas aplicadas en procesos de manufactura para detectar comportamientos atípicos o no aleatorios (status de fuera de contol) de estos procesos. Algunos comportamientos atípicos que se ven frecuentemente en las cartas de control son: ciclos, tendencias y saltos (Grant y Leavenworth 1980) . La importancia de detectar estos comportamientos atípicos fue puntualizada por el Western Electric Handbook (1956), el cual sugiere un conjunto de criterios de corridas (Run Rules) para detectar comportamientos o patrones no aleatorios del proceso.

El estudio usual de desempeño de las cartas de control se basa en la habilidad para detectar cambios tipo escalón en el parámetro del proceso. Sin embargo, en algunos procesos el cambio del parámetro es gradual, siguiendo una tendencia lineal, ésto es debido al desajuste de la máquina o a la fatiga del operario. Algunos autores han desarrollado cartas de control diseñadas para detectar tendencias lineales en el parámetro del proceso: Sweet (1988), Coleman (1989), Flaig (1991), Hackl y Ledolter (1992), Wasserman y Sudjianto (1992), Chan y Li (1994). Sin embargo, algunos reportes no porporcionan estudios comparativos de desempeño con cartas de control ya conocidas (CUSUM, EWMA) o no las superan.

En este trabajo se presenta una carta de control (carta RI) basada en la regresión isotónica, la cual está diseñada para detectar un comportamiento monotónico en la media de un proceso. Se presenta un estudio comparativo de desempeño con la carta CUSUM, utilizando como medida de desempeño el PLC el cual es estimado mediante simulación. Es deseable que el PLC sea grande cuando el porceso está bajo control y sea lo más pequeño

posible cuando el proceso está fuera de control.

2 Suposiciones acerca del proceso

Se supone que la característica de calidad del proceso que se está monitoreando es una variable aleatoria que se distribuye independientemente siguiendo una distribución normal. Cuando el proceso está bajo control, la media y la varianza son parámetros conocidos. Cuando el proceso está fuera de control, la media sigue una tendencia monotónica permaneciendo la varianza sin cambio. Luego, no hay pérdida de generalidad si se supone que la característica de calidad sigue una distribución normal estándar cuando el proceso está bajo control.

Se supone que la tendencia monotónica de la media del proceso inicia en un número de muestra desconocido (ψ) llamado punto de cambio. Entonces, el cambio monotónico de la media iniciando en ψ se expresa simbólicamente como: $\mu_1 = \mu_2 = \cdots = \mu_\psi = \mu_0 \leq \mu_{\psi+1} \leq \mu_{\psi+2} \leq \cdots \leq \mu_T$. Donde T es el número actual de muestras tomadas, μ_k es la media del proceso en la muestra k, y μ_0 es la media (conocida) del proceso bajo control.

En este artículo se presenta una carta de control de dos lados que llamaremos carta RI, basada en la regresión isotónica. La carta de control es de dos lados porque está diseñada para detectar tendencias monotónicas de la media para ambos casos; no decreciente y no creciente. Observe que el cambio tipo tendencia lineal y el cambio tipo salto o escalón son casos especiales del cambio tipo monotónico.

3 El estadístico de la carta RI

La carta RI está basada en la prueba generalizada de razón de verosimilitudes, donde las estimaciones de la media del proceso en la muestra k (μ_k) se obtiene mediante regresión isotónica. La regresión isotónica es una técnica aplicada en la estimación de las medias de una variable aleatoria (en este caso la media del proceso) bajo la restricción monotónica; $0 \leq \mu_1 \leq \mu_2 \leq \cdots \leq \mu_T$ (ver Robertson y otros (1988)). Donde T es el número actual de muestras tomadas.

Suponer que se tienen T observaciones: $x_1, x_2, \cdots, x_T$. Ríos (1997) demostró que el estadístico de la carta RI para dos lados es;

$$S_{RI} = \max\{S_{RI}^+, S_{RI}^-\} \tag{1}$$

donde $$S_{RI}^+ = \sum_{k=1}^{T}(m_k)^2 \tag{2}$$

y el conjunto $\{m_k\}$ es la regresión isotónica de $\{x_i\}$ sobre la región $\mathbb{k} = \{\mathbf{u} \in R^T \mid 0 \leq u_1 \leq u_2 \leq \cdots \leq u_T\}$.

$$S_{RI}^- = \sum_{k=1}^{T}(m_k)^2 \tag{3}$$

y el conjunto $\{m_k\}$ es en este caso la regresión isotónica de $\{x_i\}$ en la región $\mathbb{k} = \{\mathbf{u} \in R^T \mid 0 \geq u_1 \geq u_2 \geq \cdots \geq u_T\}$. La carta RI da la señal de alarma cuando $S_{RI} > h$, para cierto valor de h el cual es determinado de acuerdo al PLC deseado.

Para calcular el estadístico S_{RI}^+, primero se obtiene la regresión isotónica de las observaciones para el orden simple (es decir, en la región $\{\mathbf{u} \in R^T \mid u_1 \leq u_2 \leq \cdots \leq u_T\}$) aplicando el algoritmo de promedios de violaciones consecutivas PAVA, por sus siglas en inglés Pooling-Adjacent-Violators Algorithm (ver Robertson y otros (1988)). Después de obtener la regresión isotónica mediante el algoritmo PAVA, los valores negativos que hayan resultado se hacen cero. Finalmente se aplica la ecuación (2) a las estimaciones m_k's obtenidas. El estadístico S_{RI}^- se obtiene aplicando el procedimiento de S_{RI}^+ a los negativos de las observaciones, es decir a; $-x_1, -x_2, \cdots, -x_T$. Observe que S_{RI}^+ detecta comportamiento no decreciente, y S_{RI}^- detecta comportamiento no creciente en la media del proceso.

Ejemplo. Este ejemplo ficticio se ilustra el cálculo del estadístico de la carta RI. Suponer que un proceso produce barras de vidrio con longitud nominal 18 cm. Se sabe que esta longitud es una variable aleatoria con ditribución normal, con media 18 cm y desviación estándar 1.90 cm. Periódicamente se obtiene una muestra de tamaño 4 para monitorear el porceso y se registra la media de la muestra. Entonces, la media muestral es una variable aleatoria con distribución normal, con media 18 cm y desviación estándar 0.95 cm.

La tabla 1 presenta los resultados de 25 muestras del proceso. La segunda columna presenta la media de cada muestra, la tercera columna presenta el valor estándar de la media muestral obtenida con la fórmula $\overline{X_i}$estándar $= (\overline{X_i} - 18)/0.95$. Finalmente, en la última columna aparece el valor del estadístico de la carta RI.

A continuación se ilustra el cálculo del estadístico de la carta RI para la muestra 5, entonces $T = 5$. Las observaciones son {0.337, -1.326, -0.537, -0.042, 0.168} y los pesos iniciales para cada observación es 1. Ahora aplicamos el algoritmo PAVA para obtener la regresión isotónica de las observaciones para el orden simple ($u_1 \leq u_2 \leq \cdots \leq u_T$). Vemos que la primera y segunda observación violan el orden simple, luego se pormedian estas observaciones obteniendo $[0.337(1) - 1.326(1)]/(1+1) = -0.4945$ con peso 2 para esta "nueva observación". Las "nuevas observaciones" son; -0.4945(2), -0.537(1), -0.042(1), 0.168(1) con sus pesos correspondientes en paréntesis. Vemos nuevamente que la primera y segunda observación violan el orden simple, entonces se promedian estas observaciones; $[-0.4945(2) - 0.537(1)]/(2+1) = -0.5087$ con peso 3. Las "nuevas observaciones" son; -0.5087(3), -0.042(1), 0.168(1). Ahora, ya no hay violación al orden simple, entonces la regresión isotónica de las observaciones para el orden simple es $\{-0.5087, -0.5087, -0.5087, -0.042, 0.168\}$. Entonces, la regresión isotónica sobre $\mathbb{k} = \{\mathbf{u} \in R^T \mid 0 \leq u_1 \leq u_2 \leq \cdots \leq u_T\}$ se obtiene haciendo cero los valores negativos, obteniendo $\{0, 0, 0, 0, 0.168\}$ que son las m_k's de la ecuación (2). Finalmente $S_{RI}^+ = 0^2 + 0^2 + 0^2 + 0^2 + 0.168^2 = 0.028224$.

El estadístico S_{RI}^- se obtiene con el mismo proceso anterior aplicado a $\{-x_k\}$, obteniendo $S_{RI}^- = 0.75429$, finalmente $S_{RI} = \max\{0.028224, 0.75429\} = 0.75429$. Si S_{RI} no es mayor que h, sigue el proceso trabajando, de lo contrario se da la señal de alarma.

Tabla 1. Cálculo del estadístico de la carta RI (** no calculado).

muestra #	$\overline{X_i}$	$\overline{X_i}$ estándar	estadístico RI
1	18.32	0.337	**
2	16.74	-1.326	1.758
3	17.49	-0.537	1.735
4	17.96	-0.042	1.210
5	18.16	0.168	0.754
6	18.15	0.158	0.499
7	19.42	1.495	2.288
8	18.07	0.074	1.284
9	17.89	-0.116	0.757
10	19.47	1.547	3.150
11	20.80	2.947	11.835
12	19.19	1.253	11.970
13	18.42	0.442	10.333
14	17.50	-0.526	7.171
15	18.99	1.042	8.250
16	18.39	0.411	7.991
17	19.71	1.800	11.231
18	18.43	0.453	10.529
19	18.59	0.621	10.737

4 Estudio comparativo de desempeño

En esta sección se presenta un estudio comparativo de desempeño de la carta RI con la carta CUSUM. La carta CUSUM para dos lados opera mediante los estadísticos $C_i^+ = \max\{0, (z_i - k) + C_{i-1}^+\}$, $C_i^- = \max\{0, (-z_i - k) + C_{i-1}^-\}$ y $C_0^+ = C_0^- = 0$. La carta CUSUM da la señal de alarma cuando $C_i^+ > h$ o $C_i^- > h$, donde h se determina de acuerdo al PLC deseado (Hawkins y Olwell (1998)).

Las cartas RI y CUSUM se calibraron a un PLC bajo control de 230. Es decir, en promedio las cartas de control dan la señal de alarma en 230 observaciones estando el proceso bajo control. Para la carta RI, el valor de h correspondiente se obtuvo mediante simulación con 20000 corridas. Una corrida es el número de observaciones tomadas hasta que la carta de control da la señal de alarma, entonces de registraron 20000 corridas y su

media aritmética se utilizó como estimador del PLC. Para la carta RI se obtuvo un PLC de 230.3 con desviación estándar de 1.9763 cuando $h = 12.825$. Con este valor de h se llevaron a cabo todas las simulaciones de la carta RI.

El valor de h para la carta CUSUM se obtuvo usando un programa desarrollado para este fin por Hawkins y Olwell (1998), obteniéndose un PLC de 230 con $h = 4.3076$. Con este valor de h se llevaron a cabo las simulaciones de la carta CUSUM.

El PLC en estado estable se estimó mediante simulación de acuerdo al procedimiento "Estado estable cíclico" sugerido por Crosier (1986), utilizando 10000 corridas. Para la carta RI se utilizó un punto de cambio igual a 100, y para la carta CUSUM un punto de cambio 25. La tabla 2 muestra las estimaciones del PLC en estado estable para el cambio tipo tendencia lineal.

Tabla 2. Estimaciones del PLC bajo un cambio tipo tendencia lineal.

m	Carta RI	Carta CUSUM
-0.2000	**8.2**(0.0216)	**8.0**(0.0208)
-0.1500	**9.8**(0.0262)	**9.5**(0.0253)
-0.1000	**12.6**(0.0352)	**12.1**(0.0341)
-0.0550	**17.8**(0.0546)	**17.3**(0.0540)
-0.0100	**47.1**(0.1886)	**49.3**(0.2078)
-0.0055	**64.3**(0.2846)	**69.3**(0.3250)
-0.0010	**132.7**(0.8191)	**151.4**(1.0263)
0.0000	**230.3**(1.9763)	**230**
0.0010	**133.1**(0.8267)	**150.3**(1.0289)
0.0055	**64.6**(0.2864)	**69.4**(0.3266)
0.0100	**47.2**(0.1870)	**49.8**(0.2061)
0.0550	**18.0**(0.0548)	**17.5**(0.0548)
0.1000	**12.6**(0.0352)	**12.2**(0.0343)
0.1500	**9.8**(0.0263)	**9.5**(0.0257)
0.2000	**8.2**(0.0220)	**8.0**(0.0205)

La columna m de la tabla 2 indica el valor de la pendiente (en unidades de desviación estándar) que sigue la tendencia lineal de la media del proceso, y en paréntesis aparece la desviación estándar de las longitudes de corridas. De la tabla 2 vemos que la carta RI tiene mejor desempeño que la carta CUSUM para el rango de pendientes $|m| \leq 0.01$. También de la tabla 2 se puede ver que la carta RI prácticamente tiene un desempeño similar a la carta CUSUM para $|m| \geq 0.055$.

La tabla 3 muestra las estimaciones del PLC para cambios tipo escalón. La columna δ muestra el incremento o decremento de la media del proceso en unidades de desviación estándar. Entre paréntesis aparece la desviación estándar de las longitudes de corridas.

De la tabla 3 vemos que la carta RI tiene mejor desempeño que la carta CUSUM para el rango de $|\delta| < 0.5$, y un desempeño similar a la carta CUSUM para $|\delta| \geq 0.75$.

De las tablas 2 y 3 podemos concluir que la carta de control RI es una buena alternativa para monitorear la media de un proceso, sobre todo si se desea detectar cambios tipo tendencia lineal o tipo escalón muy pequeñas. Para cambios relativamente grandes, el desempeño de la carta RI es prácticamente el mismo que ofrece la carta CUSUM. Además, la carta RI está diseñada para detectar cambios tipo monotónicos, de los cuales los cambios tipo tendencia lineal y tipo escalón con casos particulares. Finalmente, la tabla 4 presenta algunos valores de h con sus correspondientes PLC bajo control para dar una idea de los valores adecuados de h de acuerdo a algún valor deseado del PLC bajo control.

Tabla 3. Estimaciones del PLC bajo un cambio tipo escalón.

δ	Carta RI	Carta CUSUM
-4.00	**1.2**(0.0045)	**1.6**(0.0051)
-3.00	**1.8**(0.0077)	**2.1**(0.0066)
-2.50	**2.3**(0.0106)	**2.5**(0.0085)
-2.00	**3.1**(0.0155)	**3.2**(0.0126)
-1.50	**4.9**(0.0264)	**4.6**(0.0212)
-1.00	**9.0**(0.0533)	**8.3**(0.0497)
-0.75	**14.2**(0.0883)	**13.3**(0.0945)
-0.50	**25.6**(0.1708)	**28.2**(0.2426)
-0.25	**64.2**(0.4855)	**87.8**(0.8427)
-0.10	**144.8**(1.2461)	**180.9**(1.7816)
0.00	**230.3**(1.9763)	**230**
0.10	**143.0**(1.2192)	**182.4**(1.8012)
0.25	**64.2**(0.4874)	**88.5**(0.8610)
0.50	**25.9**(0.1734)	**28.5**(0.2445)
0.75	**14.1**(0.0877)	**13.6**(0.0991)
1.00	**9.0**(0.0520)	**8.3**(0.0493)
1.50	**4.9**(0.0261)	**4.7**(0.0217)
2.00	**3.1**(0.0155)	**3.3**(0.0124)
2.50	**2.3**(0.0104)	**2.5**(0.0086)
3.00	**1.8**(0.0078)	**2.1**(0.0067)
4.00	**1.3**(0.0045)	**1.7**(0.0051)

Tabla 4. Estimaciones del PLC bajo control para algunos valores de h.

Carta RI	
h	PLC
14.25	**356.8**(2.4533)
15.00	**444.6**(2.4493)
15.25	**479.1**(2.4201)

5 Conclusiones

En este trabajo se presentó una carta de control de calidad denominada carta RI, la cual está basada en la regresión isotónica. Esta carta de control está diseñada para monitorear y detectar cambios monotónicos en la media de un proceso. Se presentó un estudio comparativo de desempeño de la carta RI y la carta CUSUM, utilizando el PLC como medida de desempeño. Se observa que la carta RI se desempeña mejor que la carta CUSUM para cambios tipo tendencia lineal en el rango de pendientes $|m| \leq 0.01$ y prácticamente tiene un desempeño similar a la carta CUSUM para $|m| \geq 0.055$. Para el cambio tipo escalón se observa que la carta RI tiene mejor desempeño que la carta CUSUM para el rango de $|\delta| < 0.5$, y un desempeño similar a la carta CUSUM para $|\delta| \geq 0.75$, donde m y δ están en unidades de desviación estándar. La carta de control RI es una buena alternativa para monitorear la media de un proceso, sobre todo si se desea detectar cambios tipo tendencia lineal o tipo escalón muy pequeñas. Para cambios relativamente grandes, el desempeño de la carta RI es prácticamente el mismo que ofrece la carta CUSUM. Además, la carta RI está diseñada para detectar cambios tipo monotónicos, de los cuales los cambios tipo tendencia lineal y tipo escalón con casos particulares.

Referencias

Chan, L. K. y Li, G-Y. (1994). A Multivariate Control Chart for Detecting Linear Trends. *Communications in Statistics and Simulation*, 23, 997-1012.

Coleman, D. E. (1989). Generalized Control Charting. Statistical Process Control in Automated Manufacturing. editores: Keats J. B. y Hubele N. F. *Marcel Dekker Inc. New York N. Y.* 155-191.

Crosier, R. B. (1986). A New Two-Sided Cumulative Sum Quality Control Scheme. *Technometrics*, 28, 187-194.

Flaig, J. J. (1991). Adaptive Control Charts. Statistical Process Control in Manufacturing. editores: Keats J. B. y Montgomery D. C. *Marcel Dekker Inc. New York N. Y.* 111-122.

Grant, E. L. y Leavenworth, R. S. (1980). Statistical Quality Control, Fifth Edition. *McGraw-Hill Book Company.*

Hackl, P. y Ledolter, J. (1992). A New Nonparametric Quality Control Technique. *Communications in Statistics and Simulation*, 21, 423-443.

Hawkins, D. M. y Olwell, D. H. (1998). Cumulative Sum Charts and Charting for Quality Improvement. *Springer-Verlag New York Inc.*

Ríos Alejandro J. G. (1997). Monitoring a Process Mean Under Trend Shift. Tesis Doctoral, *ITESM Campus Monterrey.*

Robertson. T., Wrigth, F T, y Dykstra R. L. (1988). Order Restricted Statistical Inference. *John Wiley and Sons Inc. New York.*

Sweet, A. L. (1988). Using Coupled EWMA Control Charts for Monitor Process with Linear Trends. *I. I. E. Transactions.* 20, 404-408.

Wasserman, G. S. y Sudjianto A. (1993). Short Run SPC Based upon the Second Order Dynamic Linear Model for Trend Detection. *Communications in Statistics and Simulation*, 22, 1011-1036.

Western Electric Company (1956). Statistical Quality Control Handbook. Western Electric Company, Indianapolis, IN.

An Explicit Estimator for the Shape Parameter of the Generalized Gaussian Distribution

Ramón M. Rodríguez-Dagnino y Alberto León-García
ITESM, Monterrey, N.L. *University of Toronto, Canada*

1 Introduction

The generalized Gaussian probability distribution function has been found to be a good model for the distribution of the sample values of each of the bands of a wavelet (subband) video encoder (Westerink et al., (1988), Sharifi & Léon-García, (1995)). Many methods for estimating the parameters for the generalized Gaussian distribution have been proposed so far. A survey of most of them can be found in (Varanasi & Aazhang, (1989)), where the asymptotic statistical properties are discussed for most of them. In this paper, we propose an estimator which is an approximation of the shape parameter estimator obtained by the moment method. Then, our main contribution is in finding an explicit solution for such a shape parameter estimator as a function only of the sample values. We consider that our approximation is good enough in order to be considered as an alternative for look-up tables used in most real-time applications.

As we know, the sample size N is desirable to be large enough to yield reliable estimators but small enough in order to provide an efficient use of resources. Typically, the behavior of the variance of the estimator in the range $10 < N < 100$ is critical to determine the refinaments of the statistical procedures.

In this paper, we also provide an approximate expression for the variance of the extimator which shows the relative importance of the sample size and the specific γ value in order to get a reliable estimate of the shape parameter.

2 Derivation of the Estimator

The generalized Gaussian pdf is given by

$$f_X(x;\mu_X,\sigma_X^2,\gamma) = ae^{-[b|x-\mu_X|]^\gamma}; \qquad x \in \Re \tag{1}$$

where μ_X, σ_X^2, γ are mean, variance and shape parameter of the distribution respectively. The parameters a and b are given by

$$a = \frac{b\gamma}{2\Gamma(1/\gamma)} \tag{2}$$

$$b^2 = \frac{\Gamma(3/\gamma)}{\sigma_X^2 \Gamma(1/\gamma)} \tag{3}$$

where $\Gamma(u)$ is the Gamma function given by

$$\Gamma(u) = \int_0^\infty t^{u-1} e^{-t} dt \qquad ; \quad u > 0 \tag{4}$$

Let x_i, $i = 1, \cdots, N$ be the sample value, where N is the total number of samples in a particular subband in the multiresolution decomposition. Then, we can estimate the mean and variance by the sample moments

$$\hat{\mu}_X = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{5}$$

$$\hat{\sigma}_X^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu}_X)^2 \tag{6}$$

which may give good estimates for large values of N.

A common method of estimating the shape parameter γ is by using the trascendental equation so called the "generalized Gaussian ratio function" which can be obtained by assuming that the random variable X is distributed as a zero-mean generalized Gaussian distribution

$$\frac{\sigma_X^2}{\mathbf{E}^2[|X|]} = \frac{\Gamma(1/\gamma)\Gamma(3/\gamma)}{\Gamma^2(2/\gamma)} \tag{7}$$

which does not give an explicit solution for γ and look-up tables need to be used Mallat (1989), Sharifi y Léon-García (1995).

We can find an explicit solution of equation 7 by using the following Gurland's inequality (Gurland, 1956).

$$G(\alpha, \delta) = \frac{\Gamma^2(\alpha + \delta)}{\Gamma(\alpha)\Gamma(\alpha + 2\delta)} \leq \frac{\alpha}{\alpha + \delta^2} \tag{8}$$

where $\alpha + \delta > 0$; $\alpha \neq 0, 1$; $\delta > 0$; and for convenience, $G(\alpha, \delta)$ will be called the Gurland's ratio function. Now, let $\alpha = \delta$, hence the inequality reduces to

$$G(\alpha, \alpha) = \frac{\Gamma^2(2\alpha)}{\Gamma(\alpha)\Gamma(3\alpha)} \leq \frac{1}{1+\alpha} \tag{9}$$

where by letting $\alpha = 1/\gamma$ we readily obtain the following inequality for the generalized Gaussian ratio function

$$r(\gamma) = G^{-1}(1/\gamma, 1/\gamma) = \frac{\Gamma(1/\gamma)\Gamma(3/\gamma)}{\Gamma^2(2/\gamma)} \geq 1 + \frac{1}{\gamma} \tag{10}$$

The approximation that we have found for the generalized Gaussian ratio, $r(\gamma)$, is very close to the exact values in the range $0.3 < \gamma < 3$ as it is shown in the Figure 1. This is just the important range of values in subband video encoders as it has been shown by experimental results in Sharifi & Léon-García, (1995).

Then, we can obtain an estimator for the shape parameter as

$$\hat{\gamma} = \frac{\hat{\mathbf{E}}^2[|X|]}{\hat{\sigma}_X^2 - \hat{\mathbf{E}}^2[|X|]} \tag{11}$$

where $\hat{\sigma_X^2}$ is given by the equation 6 and the estimator of $E[|X|]$ is given by

$$\hat{\mathbf{E}}[|X|] = \frac{1}{N}\sum_{i=1}^{N} |x_i - \hat{\mu}_X| \tag{12}$$

3 The Variance of the Estimator

We will use the Delta method to calculate the variance of our estimator. Let define

$$M = \frac{\mathbf{E}^2[|X|]}{\mathbf{E}[X^2]} = \frac{\mathbf{E}^2[|X|]}{\sigma_X^2} \approx \frac{\gamma}{1+\gamma} \tag{13}$$

where we have used the Gurland's inequality in the last expression and we have assumed $\mu_X = 0$.

Now, let define

$$u_r = \frac{1}{N}\sum_{j=1}^{N} |X_j|^r \tag{14}$$

as the estimator of $E[|X|^r]$, and

$$m_r = \frac{1}{N}\sum_{j=1}^{N} X_j^r \tag{15}$$

as the estimator of $E[X^r]$. Thus, we can estimate M as

$$\hat{M} = \frac{u_1^2}{m_2} = \frac{\hat{\gamma}}{1+\hat{\gamma}} \tag{16}$$

or

$$\hat{M}(1+\hat{\gamma}) - \hat{\gamma} = 0 \tag{17}$$

The variance of this implicit equation can be approximated as follows

$$Var(\hat{\gamma}) \approx Var(\hat{M}) \left[\left(\frac{d\hat{\gamma}}{d\hat{M}}\right)^2\right]\Bigg|_{\hat{\gamma}=\gamma} \tag{18}$$

Hence, we can readily obtain

$$Var(\hat{\gamma}) \approx Var(\hat{M})(1+\gamma)^4 \tag{19}$$

Now, we need to evaluate $Var(\hat{M})$ which can be approximated by

$$\begin{aligned} Var(\hat{M}) &= Var\left(\frac{u_1^2}{m_2}\right) \\ &\approx \left(\frac{\mathbf{E}(u_1^2)}{\mathbf{E}(m_2)}\right)^2 \left[\frac{Var(u_1^2)}{\mathbf{E}^2(u_1^2)} + \frac{Var(m_2)}{\mathbf{E}^2(m_2)} - \frac{2Cov(u_1^2, m_2)}{\mathbf{E}(u_1^2)\mathbf{E}(m_2)}\right] \end{aligned} \tag{20}$$

The moments of the generalized Gaussian random variable are given by

$$\mathbf{E}(|X|^r) = \frac{2a}{\gamma b^{r+1}}\Gamma\left(\frac{r+1}{\gamma}\right) \tag{21}$$

$$\mathbf{E}(X^r) = \begin{cases} 0 & ; \quad r = 1,3,5,\ldots \\ \frac{2a}{\gamma b^{r+1}}\Gamma\left(\frac{r+1}{\gamma}\right) & ; \quad r = 2,4,6,\ldots \end{cases} \tag{22}$$

The corresponding moments of the sampled moments are given by

$$\mathbf{E}(m_2) = \mathbf{E}(X^2) \tag{23}$$

$$Var(m_2) = \frac{1}{N}\mathbf{E}(X^2) \tag{24}$$

$$\mathbf{E}(u_1^2) = \frac{1}{N}\left[\mathbf{E}(|X|^2) + (N-1)\mathbf{E}^2(|X|)\right] \tag{25}$$

$$\begin{aligned} Var(u_1^2) &= \frac{1}{N^3}\left[\mathbf{E}(|X|^4) + 4(N-1)\mathbf{E}(|X|^3)\mathbf{E}(|X|)\right. \\ &\quad +(2N-3)\mathbf{E}^2(|X|^2) \; + 4(N-1)(N-3)\mathbf{E}(|X|^2)\mathbf{E}^2(|X|) \\ &\quad \left.-2(N-1)(2N-3)\mathbf{E}^4(|X|)\right] \end{aligned} \tag{26}$$

$$\begin{aligned} Cov(u_1^2, m_2) &= \frac{1}{N^2}\left[\mathbf{E}(X^4) + 2(N-1)\mathbf{E}(|X|^3)\mathbf{E}(|X|) - \mathbf{E}^2(X^2)\right. \\ &\quad \left.-2(N-1)\mathbf{E}(X^2)\mathbf{E}^2(|X|)\right] \end{aligned} \tag{27}$$

By substituting the equations 23 to 27 into the equation 20 we obtain

$$\begin{aligned} Var(\hat{M}) &= \frac{1}{N^3}\left[(2N-1) + \frac{1}{\mathbf{E}(X^2)} - \frac{\mathbf{E}(X^4)}{\mathbf{E}^2(X^2)} + (N-1)(8N-10)\frac{\mathbf{E}^2(|X|)}{\mathbf{E}(X^2)}\right. \\ &\quad +6(N-1)(N-1)(8N-10)\frac{\mathbf{E}^4(|X|)}{\mathbf{E}^2(X^2)} + 2(N-1)\frac{\mathbf{E}^2(|X|)}{\mathbf{E}^2(X^2)} \\ &\quad +(N-1)^2\frac{\mathbf{E}^4(|X|)}{\mathbf{E}^3(X^2)} - 2(N-1)\frac{\mathbf{E}^4(X)\mathbf{E}^2(|X|)}{\mathbf{E}^3(X^2)} \\ &\quad \left.-4(N-1)^2\frac{\mathbf{E}^3(|X|)\mathbf{E}(|X|^3)}{\mathbf{E}^3(X^2)}\right] \end{aligned} \tag{28}$$

By using the equations 21 and 22 together with equation 28 and the Gurland's ratio function we can obtain the following expression for the $Var(\hat{\gamma})$

$$\begin{aligned} Var(\hat{\gamma}) &\approx \frac{(1+\gamma)^4}{N^3}\left[2(N-1) + \frac{1}{\sigma_X^2} - \frac{1}{G(1/\gamma, 2/\gamma)}\right. \\ &\quad +(N-1)G(1/\gamma, 1/\gamma)\left(8N - 10 + \frac{2}{\sigma_X^2} - \frac{2}{G(1/\gamma, 2/\gamma)} - \frac{4(N-1)}{G(2/\gamma, 1/\gamma)}\right) \\ &\quad \left.+(N-1)G^2(1/\gamma, 1/\gamma)\left(6 + \frac{(N-1)}{\sigma_X^2}\right)\right] \end{aligned} \tag{29}$$

The variance of this estimator $\hat{\gamma}$ is decreasing as the sample size increases, see Figure 2. We should note that a reliable estimator is ensured for a sample size greater than 100. It is interesting that the variance is strongly affected by the specific value of γ and in some cases it is possible to obtain reliable estimates for moderate sample sizes.

4 Conclusions and Comments

A simple expression, however, accurate enough, has been found for estimating the shape parameter of the generalized Gaussian distribution. The estimator is specially well-behaved in the range of values $0.3 < \gamma < 3$ which appears to be important in subband video encoding applications.

The variance of this estimator has been conveniently expressed as a function of the Gurland's ratio function and it seems to be strongly dependant of the specific γ value, however, the value of this variance is reduced around $\gamma = 1$ independently of the sample size N. We should note that our approximate formula for the variance does not give feasible solutions for some combinations of small values of γ and very small sample sizes ($N \leq 5$).

We believe that this estimator may be useful for real time applications of any data which can be modeled as a generalized Gaussian distribution.

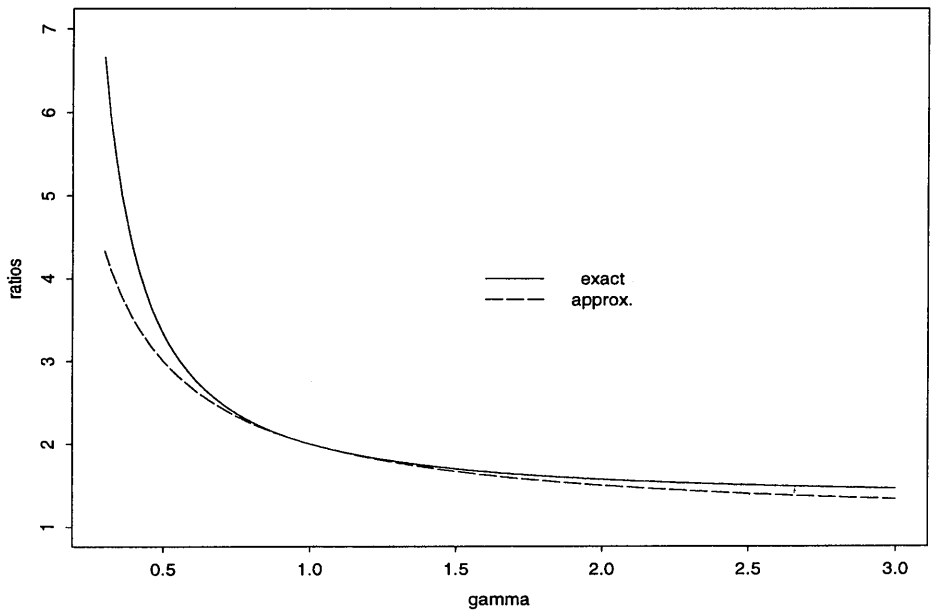


Figure 1. Comparison of the exact generalized Gaussian ratio $r(\gamma)$, and the approximation $1 + 1/\gamma$.

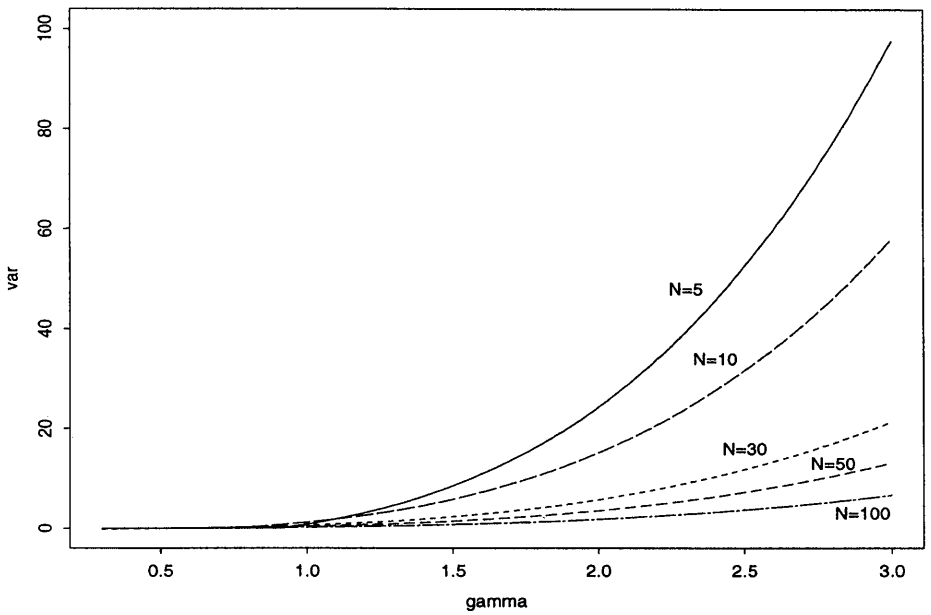


Figure 2. Variance of $\hat{\gamma}$ as a function of the sample size N. We assume $\sigma_X^2 = 1$.

References

Westerink P.H., Biemond J., and Boekee D.E. (1988) Evaluation of image sub-band coding schemes, *Signal Processing IV: Theories and Applications*, Ed. J.L. Lacoume, A. Chehikian, N. Martin, and J. Malbos, Elsevier Science Publishers B.V., North-Holland, EURASIP, pp. 1149-1152.

Varanasi M.K. and Aazhang B. (1989) Parametric generalized Gaussian Density Estimation. *J. Acoust. Soc. Am.*, Vol. 86, No. 4, pp. 1404-1415.

Sharifi K. and A. Leon-Garcia A. (1995) Estimation of shape parameter for generalized Gaussian distributions in subband decomposition of video. *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 5, No. 1, pp. 52-56.

Mallat S.G. (1989), A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, pp. 674-693.

Gurland J. (1956), An Inequality satisfied by the gamma function, *Skandinavisk Aktuarietidskrift*, vol. 39, pp. 171-172.

Una Medida de Exposición Individual a Ozono

Silvia Ruiz-Velasco y Patricia Romero
IIMAS - UNAM

1 Introducción

Durante el año escolar de 1993-1994 se llevó a cabo en tres escuelas de la Ciudad de México, un estudio epidemiológico en niños en edad preescolar para medir el efecto de ozono en la inasistencia a la escuela por causas respiratorias.

2 Estudio

En las tres escuelas se instaló una estación de monitoreo, es decir monitores continuos de ozono equivalentes a los de la Red de Monitoreo Ambiental de la Ciudad. Estos monitores tiene la característica de que se puede obtener la medida acumulada de ozono por cierta unidad de tiempo, usualmente una hora.

De los niños de estas escuelas se seleccionó una muestra aleatoria de los que vivían a menos de 10 Km. de la escuela, para llevar a cabo mediciones utilizando monitores pasivos. La muestra de niños quedó distribuida en 36 niños de la escuela 1, 37 de la escuela 2 y 45 de la escuela 3; con edades entre 2 y 7 años.

En las casas de los niños seleccionados se colocaron monitores pasivos de ozono dentro y fuera de las casas, así como dentro y fuera de sus salones de clase, por un mínimo de siete días. Una característica de los monitores pasivos es que se obtiene una medición acumulada de ozono durante todo el tiempo en que el monitor estuvo prendido.

En las casas de los participantes se colocaron monitores pasivos por un mínimo de siete días, comenzando a las ocho de la mañana del día uno. Dentro de los salones de clase de las escuelas, los monitores se colocaron por un mínimo de cinco días, entre las ocho de la mañana y la una de la tarde. Por otra parte durante este tiempo se pidió a los padres de los niños en la muestra, que llevaran un calendario de las actividades de sus hijos.

Para construir la medida de exposición, lo primero que se hizo fue obtener la tasa de penetración de ozono como el cociente de ozono interior y ozono exterior. En las escuelas esto se hizo considerando la medida acumulada del monitor continuo en los mismos

intervalos de tiempo que estuvo prendido el monitor pasivo, es decir se supone que la tasa de penetración es constante. Con esta tasa se calcula la exposición horaria interior, utilizando la medida horaria exterior.

En las casas se contaba con un monitor pasivo en el exterior de las casas y un monitor pasivo en el interior. Asumiendo que la tasa de desplazamiento del ozono del sitio donde se encuentra el monitor continuo hasta la casa es constante, se puede calcular cual sería el valor horario del monitor pasivo exterior y por lo tanto cual sería el valor horario del monitor pasivo interior, utilizando la tasa de penetración.

El siguiente paso es calcular la exposición diaria por niño, utilizando el calendario de actividades. En este calendario se detallan todas las actividades del niño desde que se levantó hasta que se durmió, especificando las horas en que empezaba y terminaba cada actividad, incluyendo el transporte de un lugar a otro y el sitio en que se encontraba.

Con estos datos se calculó la exposición individual, sumando la exposición en cada actividad. Si se encontraba en el exterior se usó la del monitor continuo y si era en el interior la de su casa, con excepción de cuando se encontraba en la escuela.

3 Resultados

En la figura se observa el comportamiento de la medida individual de ozono contra el promedio de ozono de 24 hrs. que es la medida comúnmente utilizada como exposición. El coeficiente de correlación entre las dos medidas es de 0.914. De la gráfica se observa que a medida que aumenta la concentración de ozono la variabilidad en la medida individual aumenta. La tabla 1 muestra las correlaciones entre el monitor continuo y el pasivo exterior, así como entre el monitor pasivo exterior y pasivo interior.

Tabla 1

		pasivo exterior	
	escuela 1	escuela 2	escuela 3
continuo	0.9938	0.9982	0.9997
pasivo int.	0.3791	0.6163	0.2745

Se ajustó un modelo de regresión lineal con la medida individual de ozono como variable dependiente y la edad, sexo, temperatura mínima y el promedio de ozono de 24 hrs. como variables independientes. Esto se hizo únicamente con fines descriptivos, para investigar si la exposición individual se relacionaba con alguna de estas variables. Los resultados del ajuste son:

$$indiv = 0.005 + 0.0004\ sexo + 0.13\ ozono - 0.0003\ tempmin + 0.0004\ edad$$

El ajuste es significativo con una $R^2 = 0.254$, en donde la variable sexo no es significativa. Hay una relación positiva de la medida individual con el promedio de ozono de 24 hrs., en donde por cada unidad de ozono que aumente la media de 24 hrs., la media de la exposición individual aumenta 0.13. En cuanto a la variable edad, la medida de exposición individual aumenta por cada año de edad y hay una relación negativa con temperatura mínima.

Debido a que las condiciones son diferentes en las tres escuelas, se consideró ajustar un modelo de regresión para cada escuela, los resultados son:

En la escuela 1
Regresión significativa con $R^2 = 0.155$, siendo el modelo resultante:

indiv = 0.009 + 0.13 ozono - 0.0004 tempmin - 0.0001 edad

Resultando la variable edad no significativa.

En la escuela 2
Regresión significativa con $R^2 = 0.282$, siendo el modelo ajustado:

indiv = 0.002 + 0.14 ozono - 0.00007 tempmin +0.00008 edad

Resultando la variable tempmin no significativa.

En la escuela 3
Regresión significativa con $R^2 = 0.33$, resultando el modelo:
indiv = - 0.001 + 0.15 ozono - 0.0001 tempmin - 0.0009 edad

En este caso la variable tempmin es no significativa.

De esto se concluye que hay una relación lineal positiva entre la medida individual de ozono y la medida del promedio de ozono 24 hrs., resultando similar en las tres escuelas. Habiendo también relación con la edad de los niños, siendo ésta diferente en las tres escuelas y con la temperatura mínima en una de ellas. Además es claro que no es sufiente conocer estas variables para predecir la exposición individual.

En el estudio completo los datos se analizaron utilizando regresiones Poisson, considerando como variable respuesta el número de niños ausentes por causa respiratoria por primera vez, cada día y como "offset" el número de niños a riesgo el día anterior. Para los niños en la muestra se ajustaron algunos de los modelos, para ver si mostraban el mismo comportamiento que en la población total. La tabla 2 muestra los resultados para el aumento de ausentismo considerando un incremento en 50 ppm de ozono, tanto para el estudio completo como para la muestra de niños.

Tabla 2

		Estudio Completo	Muestra	
	exp (β)	90%IC	exp(β)	90%IC
Escuela 1	1.05	0.80,1.39	1.17	0.71,195
Escuela 2	1.20	1.05,1.38	1.31	0.98,1.87
Escuela 3	0.92	0.68,1.25	0.87	0.54,1.39

Modelo ln(aus)=ln(total$_{-1}$) + α + β ozono$_{-1}$ + $\beta' x$
La variable de exposición es el promedio 24 hrs de ozono del día anterior, corregido por fecha tempmin.

Con esta medida individual y diaria se pretendía analizar el ausentismo de manera individual, sin embargo, y posiblemente por estar conscientes del monitoreo, ningún niño faltó a la escuela en el período en que fue monitoreado.

4 Conclusiones

De los resultados arriba expuestos, se concluye que la medida individual de ozono, considerando las actividades de los individuos, aunque siempre es menor o igual que la medida usual de ozono proveniente del monitor más cercano, está muy correlacionada con ella.

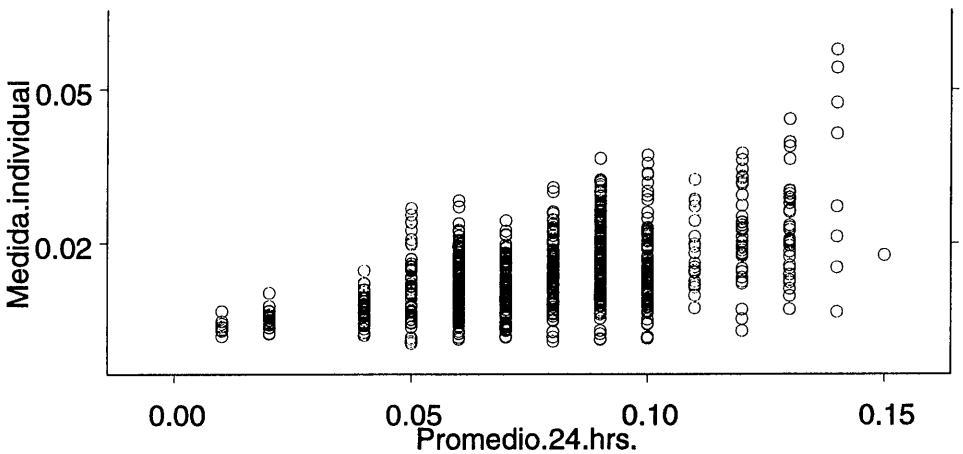


Figura 1. Medida individual de ozono vs promedio 24 hrs.

Es importante continuar con el estudio de esta nueva medición de ozono, ya que nos brinda medidas más cercanas a la realidad en la exposición al ozono.

Como vimos para calcular la medida de exposición individual se utilizó la tasa de penetración. En esta población en un estudio previo se vió que el cociente de penetración se puede explicar por variables como el tener alfombra en casa, el tiempo en que se abren

las ventanas. Entonces es posible si conocemos estas variables para otros niños obtener un estimador de la tasa de penetración en su domicilio y con su calendario de actividades calcular una medida de exposición individual.

Optimización Mediante Recocido Simulado en Regresión No-Lineal

Javier Trejos y Mario Villalobos
Universidad de Costa Rica, Costa Rica

1 Introducción

Es conocido que los métodos de regresión no lineal encuentran un óptimo local del criterio de mínimos cuadrados. En general, a partir de una estimación incial de los parámetros, se hace un descenso del gradiente, encontrándose una solución suboptimal. En muchas de las implementaciones computacionales, como por ejemplo en SAS©, se hace primero un mallado del espacio de los parámetros para tomar entonces como configuración inicial de las iteraciones, el mejor conjunto de parámetros de ese mallado. La aplicación de técnicas de optimización global, como el recocido simulado, parece bien adaptada, en particular por las propiedades de convergencia asintótica al óptimo global que posee.

En efecto, en otros problemas de estadística multidimensional hemos aplicado esas técnicas de optimización global, obteniendo excelentes resultados. Este es el caso de la rotaciones varimax oblicuas, el particionamiento unimodal y bimodal, el "multidimensional scaling" y los llamados "rough sets". Así en el problema de realizar rotaciones oblicuas para maximizar el criterio varimax (Trejos (1992)) obtuvimos resultados comparables a los de otros métodos, pero superiores en algunos casos. En particionamiento, dentro del conjunto de técnicas de análisis de conglomerados o clasificación automática, aplicamos el recocido simulado, la búsqueda tabú y un algoritmo genético, obteniendo resultados notablemente superiores a los de los métodos tradicionales, como el de k-medias o el de Ward en clasificación jerárquica (ver Trejos el al., (1998)). En cambio, para el particionamiento bimodal, también conocido como clasificación cruzada de líneas y columnas de una tabla de contingencia (cf. Gaul and Schader (1996)), obtuvimos los mismos resultados aplicando el recocido simulado (Trejos and Castillo (1999)). Por otra parte, en el llamado "multidimensional scaling", conocido en castellano como escalamiento multidimensional o análisis de proximidades, obtuvimos de nuevo resultados superiores a los de los métodos que usan otras técnicas de optimización, como el método de mayorización o el de excavación de túneles (cf. Groenen (1993)). Los resultados se pueden consultar en Trejos y Villalobos (1999), mientras que el algoritmo es descrito en Trejos y Villalobos (1998). Finalmente, también hemos emprendido investigaciones en la optimización de las soluciones de los

llamados conjuntos aproximados o "rough sets", mediante la aplicación de un algoritmo genético. Resultados preliminares se pueden consultar en Espinoza (1998). Una síntesis de todo lo anterior se encuentra en Trejos (1999).

En el presente trabajo, se plantea una aplicación a la regresión no lineal del recocido simulado mediante un mallado del espacio de parámetros, y se aplica la regla de aceptación de Metropolis para nuevos conjuntos de parámetros generados. El mecanismo de generación de soluciones se basa en el *movimiento* definido por un paso dado por el usuario, de tal forma que los parámetros se ajusten optimizando el criterio de mínimos cuadrados. Se presentarán las principales características del método propuesto, así como algunas comparaciones someras con técnicas clásicas de regresión no lineal sobre datos reales.

2 El recocido simulado

El recocido simulado, también conocido como *sobrecalentamiento simulado*, es una técnica de optimización global estocástica. Fue introducida por Kirkpatrick el al., (1983) y ha dado excelentes resultados en problemas de optimización combinatoria (ver por ejemplo Aarts and Korst (1989)). Se basa en una analogía con el procedimiento físico del sobrecalentamiento de materiales ("annenaling", en inglés) con el fin de obtener cristales muy puros. Este procedimiento establece que primero se calienta el material a muy altas temperaturas (de allí el nombre de sobrecalentamiento), y luego se enfría muy lentamente, de forma que las partículas encuentren una posición de equilibrio térmico. Un hecho curioso que tiene el procedimiento, es que a pesar de que la temperatura se desciende sistemáticamente, la energía de las partículas puede aumentar, y es esta propiedad la que precisamente hace que el método sea exitoso, ya que de esta forma permite que la función de energía se "escape" de los "valles profundos" que atraen hacia los mínimos locales, y entonces permite al material encontrar su posición de equilibrio térmico en el estado óptimo global.

En un problema de optimización combinatoria, se hace una implementación iterativa de las ideas anteriores introduciendo un parámetro externo que juegue el papel de la temperatura, llamado *parámetro de control* y denotado c_k, que será el que controlará la aceptación de nuevos estados susceptibles de ser soluciones del problema. Se requiere de la definición de los estados admisibles, y para cada estado se definirá un *vecindario* que incluye a los estados admisibles a los que se puede acceder en un paso a partir del estado actual.

La principal ventaja que tiene el recocido simulado sobre cualquier otra técnica de optimización combinatoria, es que se demuestra que *converge asintóticamente* al óptimo global del problema tratado. Para una prueba completa de la convergencia, se puede consultar Aarts and Korst (1989). Las condiciones de convergencia asintótica son relativamente simples y naturales: los vecinos de un estado deben tener todos la misma probabilidad de acceso, los vecindarios deben ser todos del mismo tamaño, se debe satisfacer una condición de reversibilidad (si se pasa de un estado a a un estado b con una probabilidad positiva,

entonces debe ser posible regresar de b a a con la misma probabilidad, y finalmente debe haber conexidad entre los estados (entre cualesquiera dos estados a y b, hay una cadena finita de estados intermedios tales que la probabilidad de pasar de un estado al siguiente es positiva).

Una adecuada implementación, requiere además la definición de lo que se llama un *programa de enfriamiento*:

1. ¿Qué se entiende por "alta temperatura" en un problema de optimización combinatoria? Es decir, qué valor inicial debe tener c_k para que al inicio de las iteraciones se acepten casi todos los nuevos estados, aún si no mejoran la función a optimizar.

2. ¿Qué significa que el sistema esté frío? Esto es, cuándo se debe terminar (en general, se decide detener las iteraciones cuando $c_k \approx 0$).

3. ¿Qué se entiende por descender la temperatura muy lentamente? Es decir, con qué ritmo se debe bajar el valor de c_k.

4. Finalmente, ¿en cuánto tiempo encuentra el sistema una posición de equilibrio térmico para cada valor de c_k? Esto es, cuántas iteraciones se deben hacer para cada valor de c_k (este valor corresponde a la longitud de las cadenas de Markov que sirven para modelar el método en la prueba de convergencia).

Para los tres primeros puntos de un programa de enfriamiento, hay escogencias más o menos estándar que se hacen y que nosotros hemos adoptado (en particular, las mencionadas en Aarts and Korst (1989)). Ahora bien, la escogencia del número de iteraciones asociado a cada valor del parámetro de control, es un problema difícil de resolver y requiere de una cierta experimentación, pero en general se aconseja que este número sea bastante grande, es decir, del orden de varios miles (con más razón si el tamaño de las tablas de datos es grande). Se puede consultar De los Cobos et al. (1997) para una introducción al recocido simulado y a las otras técnicas de optimización global.

3 Regresión no-lineal

Se supone que se tienen dos variables numéricas observadas sobre n individuos, denotando x la variable explicativa y y la variable a explicar. Se quiere encontrar $\hat{y} = f(x, \vec{\beta})$, con $\vec{\beta} = (\beta_1, \ldots, \beta_r)$, tal que:

$$\text{Min } S(\vec{\beta}) = \|y - f(x, \vec{\beta})\| \tag{1}$$

siendo $\|\cdot\|$ una norma cualquiera. Es sabido que en el caso de la norma euclídea clásica y si f es lineal, entonces la solución se obtiene por simples operaciones algebraicas. También es

sabido que en muchos casos en que f es no lineal, la estimación de los parámetros se puede hacer mediante una linealización del problema. Sobre estos asuntos se puede consultar una extensa bibliografía, pero solo citamos Draper and Smith (1968) y Tomassone et al. (1992).

En el caso en que f es una función no lineal ni el problema es linealizable, que es el caso que nos interesa, no existe un método exacto que pueda dar la solución general al problema (1). El método más usado es el de Gauss-Newton, que mediante un desarrollo de Taylor de orden 1, hace una aproximación lineal de f para mejorar iterativamente una solución inicial (se puede consultar Bates and Watts para una amplia descripción del método de Gauss-Newton). Ahora bien, la definición de las iteraciones se hace de tal forma que los nuevos conjuntos de parámetros son aceptados únicamente si mejoran el criterio de mínimos cuadrados, por lo que irremediablemente el método converge a soluciones que pueden ser suboptimales. Es decir, si la estimación inicial de $\vec{\beta}$ está en el "valle" de un mínimo relativo, entonces no será posible que el método se escape de este mínimo.

4 Características del método propuesto

Definiremos un *estado* del problema (1) como $I = (\beta_1^I, \ldots, \beta_r^I) \in \mathbb{R}^r$. Dado $h \in \mathbb{R}$, se dirá que un estado J es un *vecino* de I si existe $l \in \{1, \ldots, r\}$ y $\alpha \in \{-1, 1\}$ tales que

$$\beta_i^J = \begin{cases} \beta_i^I + h\alpha & \text{si } i = l \\ \beta_i^I & \text{si } i \neq l \end{cases}$$

Así, la generación de un nuevo estado J a partir de I se hará según el siguiente mecanismo aleatorio: (a) sea $J := I$, (b) para cada l en $\{1, \ldots, r\}$, hacer: (i) decidir al azar si l se incluye, (ii) en caso de que l se incluye, escoger al azar α en $\{-1, 1\}$, y (iii) sea $\beta_l^J := \beta_l^I + h\alpha$. Se supone que las escogencias al azar se hacen con una ley uniforme. Este procedimiento corresponde a construir un mallado de ancho h del espacio de los parámetros $\vec{\beta}$, y realizar una caminata aleatoria en este mallado, escogiendo al azar el vecino del estado actual de $\vec{\beta}$. Por ejemplo, en un caso de dos parámetros, uno podría imaginarse que los movimientos se hacen al "norte", noreste", "este", etc. Por otra parte, el ancho de la malla puede ser diferente para cada parámetro, en cuyo caso habría que sustituir h por h_l en la descripción anterior. Así mismo, este ancho de malla puede disminuir con el algoritmo de recocido simulado, en cuyo caso también debería tener un índice k que corresponda a la iteración del parámetro de control c_k. Sin embargo, no incluimos estos casos en la descripción que sigue con el fin de no hacer pesada la presentación.

Puede verse que entonces se cumplen las condiciones de convergencia asintótica:

- **Reversibilidad:** se escogen (al azar, con la misma probabilidad) los mismos índices $l \in \{1, \ldots, r\}$ y se escoge $\alpha := -\alpha$, regresando entonces de J a I.

- Conexidad: cualquier punto de la malla del espacio de los $\vec{\beta}$ puede ser alcanzado mediante un número finito de pasos.
- Tamaño de los vecindarios: cada vecindario tiene tamaño $3^r - 1$, por lo que la probabilidad de generar cualquier vecino de un estado I es $\frac{1}{3^r-1}$.

5 Algoritmo

A continuación presentamos el algoritmo RNL_SS, que implementa el recocido simulado para obtener los parámetros en un problema de regresión no lineal. El procedimento Generar(J, I) es el mecanismo de generación de nuevos estados descrito en la sección anterior. Se denota con L_k la longitud de las cadenas de Markov asociadas a cada valor de c_k (que dependen de un factor F_L dado por el usuario) y por S la suma de los cuadrados de las diferencias entre y y $f(x, \vec{\beta}) = \hat{y}$, es decir, la función a minimizar.

```
procedure RNL_SS;
begin
  Inicializar (I_o, c_o, γ);
  k := 0; I := I_o;
  c_k := c_o; L_k := F_L * r;
  S_i := S(I);
  opt := I; S_opt := S_I; {estado optimal}
  repeat
    for L := 1 to L_k do
      Generar(J, I);
      S_J := S(J));
      if S_J ≤ S_I then
        I := J; S_I := S_J;
        if S_I < S_opt then
          opt = I; S_opt = S_I;
        end;
      endthen
      else
        if exp(S_I − S_J)/c_k > random [0, 1) then
          I := J; S_I := S_J;
        endif
      endelseif
    endfor;
    k := k + 1; Calcular L_k; c_k := c_k γ;
  until c_k ≈ 0 u otro criterio;
end;
```

6 Algunos resultados

En nuestras implementaciones, usamos la norma euclídea clásica al cuadrado en el problema (1). Los resultados que siguen corresponden a un estudio del crecimiento de los tejidos del ñame alado (ver Rodríguez y Trejos (1995)), en el cual se usó un programa comercial; las limitaciones de la metodología allí implementada, nos indujeron a hacer este estudio de la optimalidad de la regresión no lineal. A manera de ilustración, presentamos los resultados sobre los tubérculos, usando un modelo logístico:

$$f(x; M, \kappa, g) = \frac{M}{1 + \exp[-\kappa(x - g)]} \tag{2}$$

y sobre los tallos, usando un modelo de campana truncada:

$$f(x; M, \kappa, g) = Me^{[-\kappa(x-g)^2]}. \tag{3}$$

En la tabla 1 se presentan los resultados obtenidos para los tubérculos usando el modelo logístico, comparándolos con los obtenidos con el método de Gauss-Newton y para dos mallados diferentes. En la tabla 2 se presentan los resultados obtenidos para los tallos usando el modelo de campana truncada. Se puede apreciar que los resultados obtenidos son prácticamente los mismos, y que la precisión mejora un poco si el mallado es más fino. A manera de ilustración diremos que en el primer caso el coeficiente de determinación es de 96.37% mientras que en el segundo caso es de 86.25% (aunque prevenimos al lector que este coeficiente, por ser lineal, tiene poco sentido en regresión no lineal).

7 Comentarios y perspectivas

Si bien es cierto que para el ejemplo presentado se obtuvieron los mismos resultados que con el método de Gauss-Newton, creemos que se pueden encontrar conjuntos de datos donde se pueda apreciar la utilidad de usar una técnica de optimización global, como es el caso en otros problemas de análisis de datos. De esta forma, estamos emprendiendo comparaciones exhaustivas sobre la capacidad que tiene el método propuesto para encontrar los óptimos, en diversas situaciones. Finalmente, también estamos implementando otras técnicas de optimización global, como los algoritmos genéticos y la búsqueda tabú.

Referencias

Aarts, E.M. and Korst, J. (1989). *Simulated Annealing and Boltzmann Machines.* John Wiley & Sons, Chichester.

	G-N (SAS©)	RNL_SS (100)	RNL_SS (1000)
M	9726.58654	9809.8803	9735.37488
κ	0.069572	0.06845	0.069464
g	283.295855	283.630	283.379
$S/\|\|y\|\|^2$	3.6273%	3.6293%	3.6274%

Tabla 1. Resultados comparativos para los tubérculos con el modelo logístico entre la regresión no lineal usando el método de Gauss-Newton (G-N) implementado en SAS© y usando recocido simulado (RNL_SS), para mallas de 100 y 1000 puntos en cada dimensión.

	G-N (SAS©)	RNL_SS (100)	RNL_SS (1000)
M	3079.70691	3085.8357	3080.60547
κ	0.000262	0.00026627	0.00026194
g	280.861285	281.12	280.869
$S/\|\|y\|\|^2$	13.73878%	13.74685%	13.73879

Tabla 2. Resultados comparativos para los tallos con el modelo de campana truncada entre la regresión no lineal usando el método de Gauss-Newton (G-N) implementado en SAS© y usando recocido simulado (RNL_SS), para mallas de 100 y 1000 puntos en cada dimensión.

Bates, D.M. and Watts D.G. (1988). *Nonlinear Regression Analysis and Its Applications.* John Wiley & Sons, New York.

De los Cobos, S., Pérez, B.R. y Gutiérrez, M.A. (1997). Programación estocástica en optimización. In: *X Simposio Internacional de Métodos Matemáticos Aplicados a las Ciencias*, W. Castillo & J. Trejos (Eds.), Liberia, UCR: 31–45.

Draper, N. R. and Smith, H. (1968). *Applied Regression Analysis.* John Wiley & Sons, New York.

Espinoza, J.L. (1998) Conjuntos aproximados y algoritmos genéticos. In: *Estudios en Análisis de Datos y Estadística*, W. Castillo & J. Trejos (Eds.), Santa Clara, UCR–ITCR: 215–223.

Gaul, W. and Schader, M. (1996) . A new algorithm for two-mode clustering. In: *Data Analysis and Information Systems*, H.-H. Bock & W. Polasek (Eds.), Springer, Heidelberg.

Groenen, P.J.M. (1993). A comparison of two methods for global optimization in multidimensional scaling. In: *Information and Classification*, O. Opitz et al. (Eds.), Springer, Heidelberg.

Kirkpatrick, S., Gellat, C.D. and Vecchi, M.P. (1983). Optimization by simulated annealing, *Science*, 220, 671–680.

Rodríguez, W. y Trejos, J. (1995) *Análisis y modelado del crecimiento del ñame alado (Dioscorea alata).* Informe Final de Investigación, Universidad de Costa Rica.

Tomassone, R., Audrain, S., Lesquoy, E., Millier, C. (1992). *La Régression. Des Nouveaux Regards sur une Ancienne Méthode Statistique.* Masson, Paris.

Trejos, J. (1992). A simulated annealing implementation for oblique varimax rotations. In: *Applied Stochastic Models and Data Analysis*, Vol. II, J. Janssen & C.H. Skiadas (Eds.), World Scientific, Singapur, 981–989.

Trejos, J., Murillo, A. and Piza, E. (1998). Global stochastic optimization for partitioning. In: *Advances in Data Science and Classification*, A. Rizzi et al. (Eds.), Springer, Heidelberg, 185–190.

Trejos, J. y Villalobos, M. (1998) Análisis de proximidades usando sobrecalentamiento simulado. In: *Estudios en Análisis de Datos y Estadística*, W. Castillo & J. Trejos (Eds.), Santa Clara, UCR–ITCR: 41–52.

Trejos, J. (1999). Stochastic optimization applied in data analysis. In: *Applied Stochastic Models and Data Analysis*, Proceedings of the IX International Symposium, Lisbon, June 1999.

Trejos, J. y Villalobos, M. (1999). Une implémentation du recuit simulé en analyse des proximités. In: Actes des *XXXI Journées de Statistique*, Grenoble, Mai 1999.

Trejos, J. and Castillo, W. (1999). Simulated annealing optimization for two-mode partitioning. In: *Classification and Information Processing at the Turn of the Millenium*, W. Gaul & R. Decker (Eds.), Springer, Heidelberg.

Estudio Exploratorio de Indices Ecológicos en una Muestra de Arbolado Urbano de la Ciudad de México

Héctor Javier Vázquez
UAM-Azcapotzalco, México

Jaime Grabinsky
UAM-Azcapotzalco, México

Alicia Chacalo
UAM-Azcapotzalco, México

Alejandro Aldama
UAM-Azcapotzalco, México

1 Introducción y objetivos

Los árboles en las calles de las ciudades han mostrado múltiples beneficios desde que el desarrollo urbano internacional generó ciudades densamente pobladas. Sus efectos fisiológicos, psicológicos, en la química ambiental, financieros, estéticos han sido motivo de estudios gubernamentales, académicos y de los industriales de viveros, entre otros.

En la investigación de arbolado urbano se vive la etapa fenomenológica y los inicios de la cuantitativa. El nivel de experimentación y teorización no ha alcanzado la madurez de otras disciplinas; por ejemplo, el de las forestales y aún menos que el de las ciencias "biológicas duras". Para esto se ha trabajado con muestreos representativos integrando algunas variables como especie, localización y estado del lugar, altura, ancho de troncos, problemas sanitarios, etc.

En la Universidad Autónoma Metropolitana Azcapotzalco, se ha trabajado en los problemas de crecimiento y mantenimiento del arbolado de las calles con énfasis en el D.F. Se han evidenciado entre otros fenómenos que:

el mantenimiento es insatisfactorio

- hay gran cantidad de interferencias para el sano desarrollo de los árboles
- las delegaciones muestran fuertes diferencias en la cantidad promedio de árboles
- hay muy poca diversidad de especies en algunas delegaciones
- se dedican muchos recursos y trabajo al arbolado aunque con una inadecuada planeación
- el bosque urbano sufre de vandalismo y descuidos de la población

Sin embargo dado el interés para establecer evaluaciones y comparaciones más generales en bosques urbanos (Richards, 1993), se procedió a la evaluación de una muestra represen-

tativa en términos de características ecológicas como: distribución espacial de individuos de una o más especies, diversidad en términos de abundancia y equitatividad, asociación inter-especies. El uso de índices ecológicos para evaluar estas propiedades en una población o muestra biótica es tradición en sistemas naturales (Ludwig y Reynolds, 1988). En sistemas urbanos, particularmente en bosques urbanos, no se conocen estudios del uso de este tipo de índices. Inicialmente en un primer estudio se integraron estos índices como nuevas variables y se evaluó mediante un estudio multivariado la relación de estos índices con variables como la calidad de los sitios, la condición general de los árboles, la densidad (Grabinsky et al., 1998).

El primer objetivo de éste trabajo es el de estudiar una muestra representativa del bosque urbano de la Ciudad de México mediante el uso de estos índices y describir así la estructura en términos de las características ecológicas ya mencionadas. Sin embargo a pesar de las ventajas de estos índices para resumir varias propiedades en un sólo número, se observaron algunas desventajas. El análisis multivariado resultó un complemento valioso para la exploración (Lebart et al., 1984). Presentar los resultados de este análisis y compararlos con los índices ecológicos constituye el segundo objetivo de este trabajo.

2 Evaluación de la estructura de una muestra árboles de las calles mediante el uso de índices ecológicos

Los datos provienen de una muestra representativa de 1260 árboles (Chacalo et al., 1994, 1996), con más de 50 especies diferentes, localizados en 229 manzanas repartidas en las 16 delegaciones políticas del Distrito Federal (**A Obr**: Alvaro Obregón, **MiHi**: Miguel Hidalgo, **Cuauh**: Cuauhtemoc, **B Ju**: Benito Juárez, **Azc** : Azcapotzalco, **Izpa**: Iztapalapa, **MiAl**: Milpa Alta, **GAM**: Gustavo A. Madero, **Tlalp** : Tlalpan, **Tlah**: Tláhuac, **Izco**: Iztacalco, **Xoch**: Xochimilco, **Cua**: Cuajimalpa, **MCon**: Magdalena Contreras, **Coy**: Coyoacán, **V Ca**: Venustiano Carranza). La base de datos con 58 variables contiene mas de 75,000 entradas.

Distribución Espacial.- Entre los índices desarrollados para evaluar la distribución espacial se encuentran aquellos basados en la razón **varianza/media**, razón que evalúa el grado de agrupación o agregación de individuos. En el caso de una distribución al azar, la varianza es igual a la media. Esto supone que la posición de cada individuo está determinada por factores independientes de aquellos que determinan la localización de otros individuos. El modelo estadístico correspondiente es el modelo de Poisson. Si la razón **varianza/media** aumenta se observa una mayor agrupación de individuos; por el contrario si esta razón es inferior a 1, los individuos se encuentran espaciados de manera uniforme. La Figura 1 muestra los modelos asociados a las tres distribuciones y los valores de los índices de Dispersión (**ID**) y de Green (**GI**).

A partir de la base de datos se obtienen las tablas de distribuciones del número de

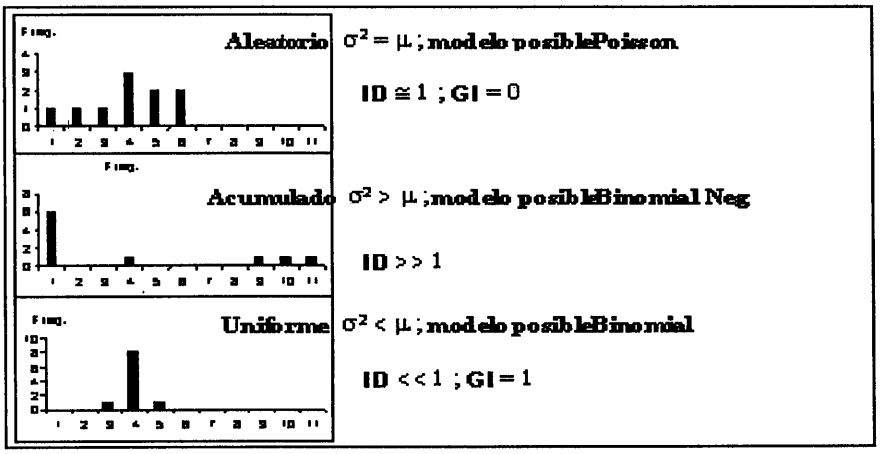


Figura 1: Modelos estadísticos de distribuciones espaciales

individuos de especies dominantes (**CDR**: Cedro, **OLM**: Olmo, **CSR**: Casuarina, **FRS**: Fresno, **ECL**: Eucalipto, **TRN**: Trueno, **JCR**: Jacaranda, **CLR**: Colorín). en las 16 delegaciones. Se observa, por ejemplo

que 34 fresnos se encuentran sólo en una delegación, mientras que hay cuatro delegaciones que no tienen olmos. Esta información con los índices de dispersión y el índice de Green se representa en diagramas de barras (Figura 2). En todas las distribuciones se observan valores del **ID** mayores que uno e índices de Green entre 0 y 1, lo que sugiere que los individuos se agrupan en ciertas delegaciones.

Los patrones que se observan en los diagramas son diferentes y sin embargo los índices dejan pensar que las distribuciones son parecidas. Dada esta situación, se procedió a realizar un estudio multivariado partiendo de la tabla de distribución. El Análisis de Componentes Principales permite concentrar 62.7 % de la varianza total en dos ejes. Las proyecciones sobre los dos primeros ejes permiten detectar tres niveles de agrupación: **Nivel 1:** los fresnos se encuentran en mayor número de delegaciones, con acumulaciones relativamente más numerosas que las otras especies; **Nivel 2 :** la casuarina, el olmo y la jacaranda se encuentran agrupados a un nivel intermedio; **Nivel 3 :** el eucalipto, el trueno, el colorín y el cedro se encuentran relativamente poco agrupados Figura 3. Los resultados del análisis de conglomerados confirman estos tres niveles de agrupación.

Diversidad.- En este concepto se integran dos nociones: "abundancia" (número de especies) y la equitatividad (abundancia relativa). La diversidad se puede evaluar en términos del tamaño de la muestra (**N**), del número total de especies presentes en la muestra (**N0** = **S**), del número de especies "abundantes" (**N1**) o del número de especies muy abundantes (**N2**). A partir de estos términos se han propuesto las siguientes definiciones (Ludwig et al., 1998):

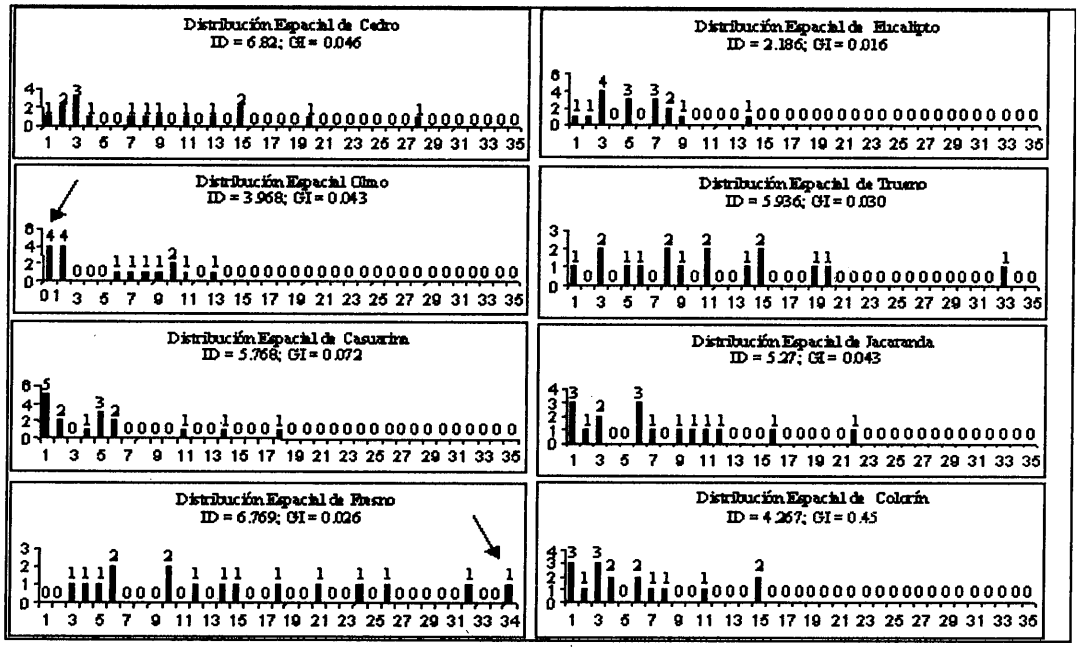


Figura 2: Distribuciones Espaciales de Especies Dominantes.

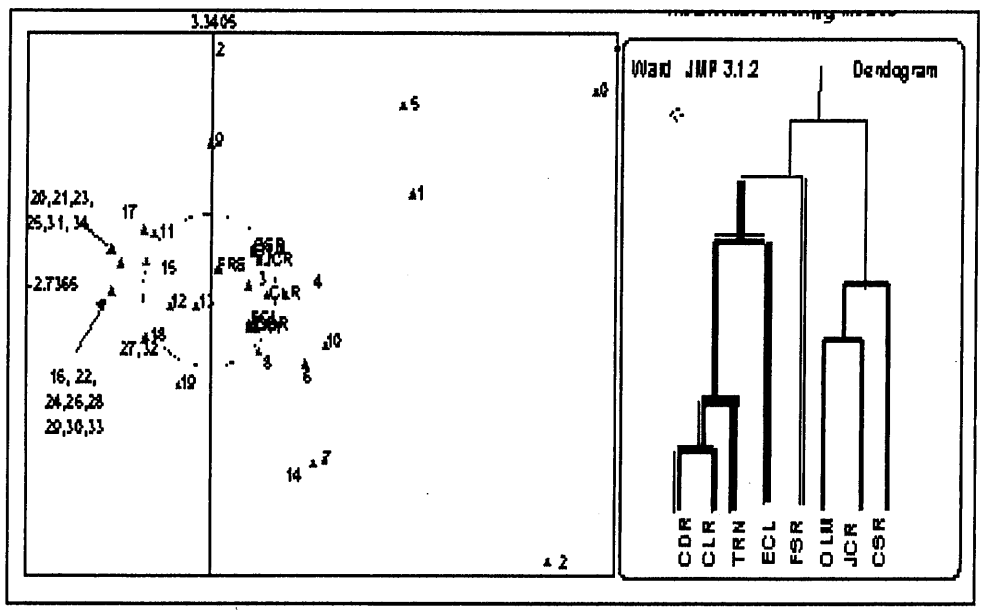
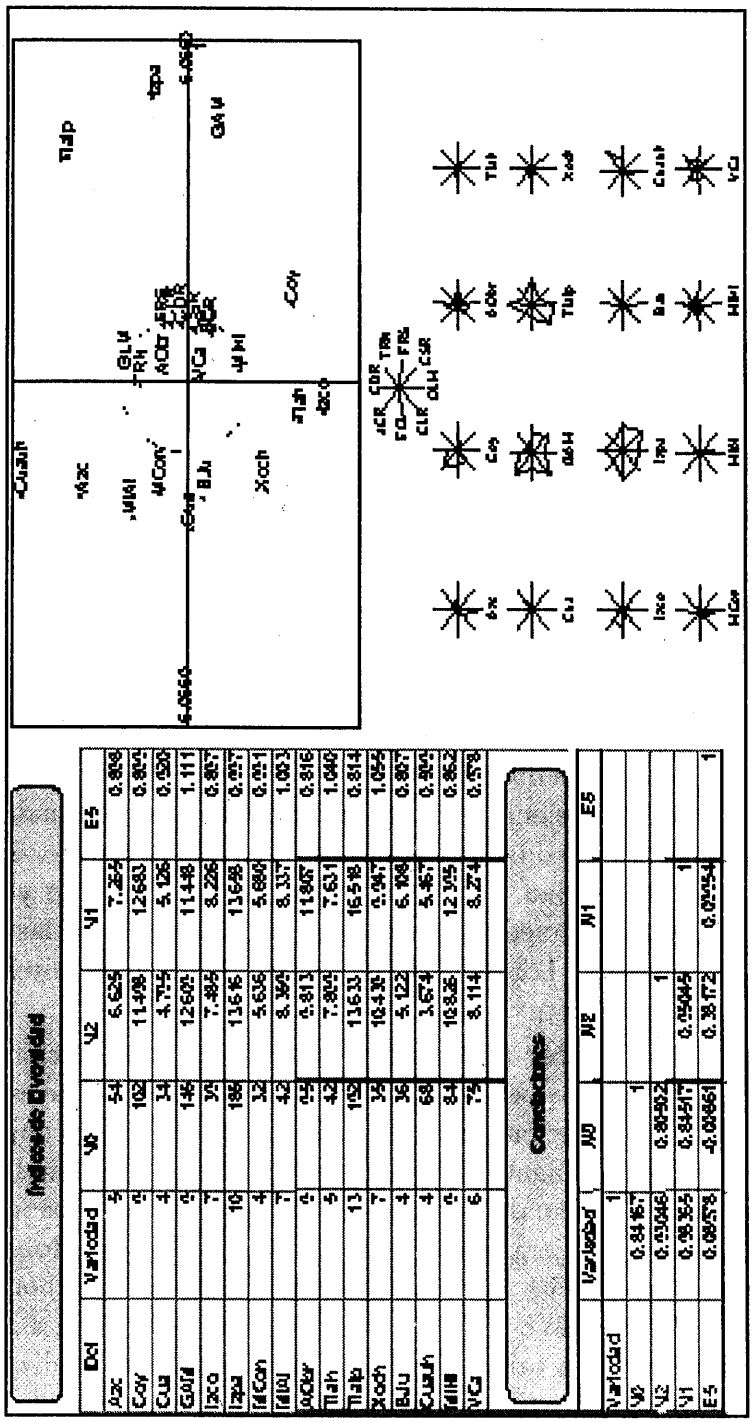


Figura 3: Primer plano resultado del Análisis de Componentes Principales y Dendograma del Análisis de Conglomerados para Especies Dominantes.

Figura 4. Indices de densidad, correlaciones, primer plano del análisis en comp. principales



N1 e^H donde H es el índice de Shannon H = -$\Sigma p_i \ln p_i$

$p_i = n_i$ /**N** y **N2** = **1**/λ donde $\lambda = -\Sigma p_i^2$

La equitatividad **es la uniformidad de la distribución del número de individuos** de cada especie. Los índices más comunes son :

E1 = e H' /S = (N1)/(N0) ; E2 = e H' /ln(S) = (ln N1)/(ln(N0)

E5 = (N2 - 1)/(N1 - 1) modificada de Hill. Si E5 0 una especie se vuelve dominante

La Figura 4 presenta la diversidad de las 16 delegaciones evaluada con algunos de los índices de abundancia mencionados, el índice de equitatividad **E5** y la variable **Variedad** (definida como el número de especies para completar el 80% de individuos). De acuerdo a los índices de diversidad **N0**, **N2** y **Variedad**, las delegaciones con mayor diversidad son: Tlalpan, Iztapalapa, Gustavo A. Madero y Miguel Hidalgo. Sin embargo, no es evidente definir cuales son aquellas delegaciones con menor diversidad ya que entre los índices existen diferencias. Por ejemplo con el índice **N0** se obtienen (MCon, Xoch y BJu), con el índice **N2** (Cuauh, Cuaj y BJu). A pesar de esta dificultad los índices están bien correlacionados entre sí. Según el índice de equitatividad **E5** la delegación Gustavo A. Madero presenta mayor equitatividad y la delegación Cuauhtémoc la menor.

El Análisis Multivariado, en particular el Análisis de Componentes Principales Figura 5, permite visualizar la tabla de abundancias, obtener conclusiones similares y detectar rasgos interesantes que no se observan fácilmente del estudio de la tabla de abundancias, ni de la tabla de índices. Por ejemplo las delegaciones Tlalpan, Iztapalapa y Gustavo A. Madero tienen una mayor diversidad, Cuajimalpa, Magdalena Contreras, Milpa Alta y Benito Juárez tienen poca diversidad. Miguel Hidalgo, Alvaro Obregón y Venustiano Carranza están en una situación intermedia. También es posible visualizar las delegaciones con muy poca presencia de especies. Los diagramas polares Figura 4 permiten apreciar abundancia y diversidad simultáneamente.

Asociación Inter-especies.- Para evaluar la asociación entre especies se usa una tabla binaria en donde se señalan las presencias o ausencias de especies por delegación. Es posible evaluar la asociación de especies de dos en dos o en conjuntos mayores. En el caso de ocho especies dominantes se obtienen así 28 combinaciones. Para cada pareja se obtienen los índices de **Jaccard**, **Ochiai** y **Dice**.(Figura 5). Estos índices son consistentes y permiten establecer una jerarquía de asociación de **dos en dos**. La evaluación de la asociación de las ocho especies dominantes en **conjunto** se hace por medio del estadístico Chi2 en el intervalo centrado del 90% y no se observa asociación, a nivel delegación, entre las especies dominantes. La tabla de presencias ausencias se analiza mediante el Análisis de Componentes Principales (Figura 6). En el primer plano, se observan delegaciones en el centro y otro conjunto en la periferia, lo mismo para las especies.

Pres/Aus	Azc	Coy	Cua	GAM	Izco	Izpa	MCon	MAl	AObr	Tlah	Tlalp	Xoch	BJu	Cuauh	MHi	VCa	
ESPECIE																	
FRS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16
TRN	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	15
CDR	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	15
JCR	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	13
CLR	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
ECL	1	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	13
OLM	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	0	12
CSR	1	1	0	1	1	1	0	1	1	1	1	0	0	0	1	1	11
	7	7	5	8	8	8	5	6	8	6	8	6	6	7	8	7	

	Presente	Ausente	
Presente	a	b	m = a+b
Ausente	c	d	n = c+d
	r = a+c	s = b + d	N = a+b+c+d

Indice de Jaccard = a/(a+b+c)

Indice de Ochiai = a/√((a+b)*(a+c))

Indice de Dice = ID = (2a)/(2a+b+c)

Figura 5. Tabla de presencias-ausencias y definición de índices de asociación inter-especie

La interpretación es la siguiente: las delegaciones situadas en el centro tienen la mayoría o todas las especies. Si una delegación está en la periferia y cercana a una especie quiere decir también que esa especie está presente en esa delegación. Por el contrario, si la especie está en la periferia, pero lejos de esta delegación, la especie está ausente de la delegación. Por ejemplo, la delegación Gustavo A. Madero se encuentra en el centro y ninguna especie está ausente. Magdalena Contreras situada en la periferia no tiene cedros, casuarinas, ni jacarandas.

3 Discusión y conclusiones

El estudio de una muestra representativa del bosque urbano de la Ciudad de México mediante el uso de índices ha permitido evaluar algunas características ecológicas (distribución espacial de los árboles, diversidad y asociación inter-especies) y establecer jerarquías entre grupos de especies y delegaciones.

Es importante resaltar que como la mayoría de índices, son medidas que integran varios indicadores o características en un solo número. Esto permite obtener evaluaciones o jerarquías dentro de un conjunto de unidades estadísticas, sin embargo estos índices no siempre integran toda la información disponible y no garantizan una total representación del indicador o característica asociada. Esto puede dar lugar a falsas interpretaciones. En general, se observa lo siguiente:

- diferencias en las interpretaciones asociadas a los tres grupos de índices
- discordancia entre índices que evalúan el mismo concepto

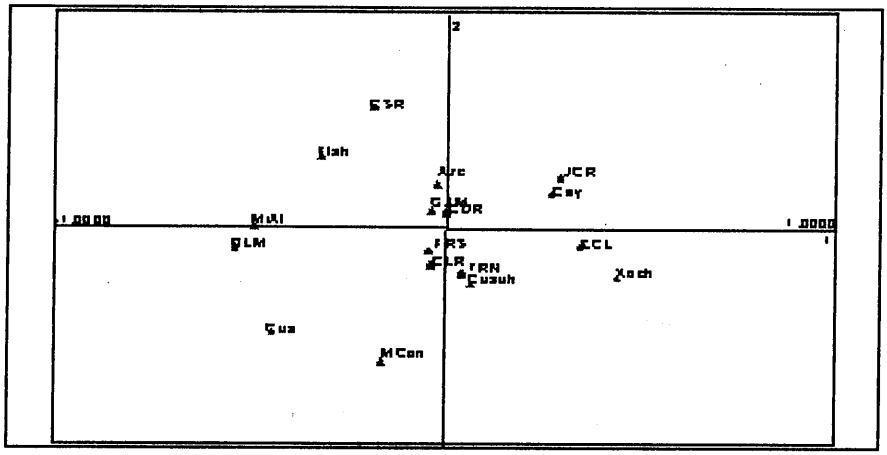


Figura 6. Primer plano resultado del Análisis en Componentes Principales de la Matriz de presencias-ausencias (delegación x especie)

- dificultad para distinguir patrones diferentes

- pérdida de información significativa y valiosa.

Por ejemplo se observa que varios índices propuestos para evaluar la misma característica pueden ser inconsistentes; en otros casos el mismo valor del índice puede obtenerse para dos muestras substancialmente diferentes.

En el caso particular del índice de Distribución Espacial, no integra las posiciones de los individuos en el espacio. Una vía de trabajo a desarrollar sería efectuar estudios con ayuda de Sistemas de Información Geográfica (GIS). Con un conocimiento de las posiciones de los árboles dentro de la mancha urbana del Distrito Federal sería posible construir modelos estocásticos y de simulación.

Para reducir algunas de las limitaciones observadas con estos índices el Análisis Multivariado resultó útil ya que permitió consolidar y definir mejor los resultados obtenidos, al visualizar grupos de individuos con características similares y resumir rápidamente la información, con poco trabajo y tiempo de computadora. Con un estudio más detallado de las matrices es posible también obtener conclusiones semejantes, pero requiere una mayor experiencia con estos métodos.

En lo que respecta a la posibilidad de conocer las causas responsables de un cierto comportamiento, en sistemas naturales es común generar conjuntos de hipótesis explicativas. Por ejemplo una distribución de árboles con muchos agrupamientos puede explicarse a partir del grado de dispersión de los árboles padres, por diferencias del ambiente o por relaciones inter-especies (competencia, simbiosis). En sistemas urbanos un buen número de los árboles de las calles han sido plantados ex profeso, o sea que los mecanismos

naturales son poco importantes. Sin embargo es posible explicar el estado del bosque a partir de causas de origen social. Para esto sería interesante disponer de varias muestras representativas a lo largo del tiempo.

Referencias

Chacalo, A., Aldama, A. and Grabinsky, J. (1994). Street Tree Inventory in Mexico City. *Journal of Arboriculture*, 20(4): 222-226.

Chacalo, A., Grabinsky, J. and Aldama A. (1996). Inventario del Arbolado de Alineación de la Ciudad de México. *Ciencia Forestal.* Vol. 21, Num 79, México D.F., México.

Grabinsky. J., Aldama A., Chacalo, A. and Vázquez H. J. (1998). Integration of Ecological Indices in the Multivariate Evaluation of an Urban Inventory of Street Trees. *Integrated Tools for Natural Resource Inventories in the 21st Century.*, Boise, Idaho. U.S.A. (publicación de memoria en evaluación).

Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis*, John Wiley.

Ludwig J.A. and Reynolds J. F. (1988). *Statistical Ecology Wiley.*

Richards N.A.(1993). Reasonable Guidelines for Street Tree Diversity. 344-*349 Journal of Arboriculture* 19 (6) November.

Filtrado y Selección de Variedades

María del Carmen Ybarra Moncada Guillermo Zárate de Lara

Martha Elva Ramírez

ISEI Programa de Estadística, Colegio de Postgraduados

1 Introducción y objetivos

La planeación de experimentos con un gran número de tratamientos es común en la actualidad, siendo necesario elegir eficientemente los mejores elementos; sin embargo, con frecuencia la disponibilidad del material en estudio no es suficiente, con ello, la planeación de los experimentos se enfrenta a aspectos de diseño que ocasionan problemas en la ejecución.

La experimentación con una gran cantidad de tratamientos nuevos, da lugar a la puesta en marcha de programas de producción, conformados de varias fases como: producción inicial, filtrado, selección, validación y optimización; a medida que avanzan las fases del programa de producción, se va haciendo una selección más fina de los tratamientos; en cada fase es necesario contar con los métodos estadísticos adecuados para obtener información relevante al problema y poder concluir con un adecuado nivel de confianza.

Con la finalidad de contribuir en la difusión y conocimiento de herramientas estadísticas adecuadas para el filtrado y selección de tratamientos, el objetivo del presente trabajo consiste en recomendar un paquete de diseños experimentales que permiten al investigador adoptar el modelo experimental que ajuste mejor a cada caso, con objeto de controlar de modo adecuado la variabilidad, comparar objetivamente nuevos tratamientos y seleccionar los mejores.

Las técnicas estadísticas recomendadas se enfocan a los diseños experimentales aumentados, los cuales son una vía objetiva para el ajuste de tratamientos con una observación, facilitando la identificación de los elementos más eficientes hasta lograr una depuración en la cual el número de tratamientos es razonable desde el punto de vista económico para implementar más de una repetición.

Cuando es factible la repetición de los tratamientos, el objetivo es seleccionar al más pequeño de los subconjuntos que contengan los mejores tratamientos; para ello se sugiere el procedimiento de selección del mejor subconjunto de tratamientos bajo una probabilidad de éxito especificada.

2 Metodología

La sugerencia hecha por Federer (1961) resolvió el problema de estudiar el efecto de tratamientos con una sola repetición,. Federer propone un diseño experimental aumentado como cualquier diseño estándar, aumentado con tratamientos adicionales en bloques completos, bloques incompletos, hileras o columnas cuyo objetivo es estimar el cuadrado medio del error de las observaciones (s^2) y estimar las varianzas para diversas comparaciones entre tratamientos los cuales son ajustados por bloques. En estos diseños se tienen dos tipos de tratamientos, a saber:

a) Tratamientos repetidos r veces y que ocurren una vez en cada bloque denominados tratamientos control o estándar.

b) Tratamientos que ocurren una vez en uno de los r bloques denominados tratamientos nuevos.

En el siguiente cuadro se ilustran las clases de tratamientos mencionados.

Tabla 1. Tratamientos nuevos y tratamientos control en un diseño de bloques completos al azar aumentados.

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

En la fase de selección, el procedimiento de Gupta y Sobel establece con una confianza P*, que para todos los tratamientos eliminados, el valor del parámetro de interés μ_i es menor que el del control μ_0, dándose la selección de un subconjunto que contiene todos los tratamientos tan buenos o mejores que el control. Los supuestos estadísticos del procedimiento indican que:

i) Sean t muestras aleatorias e independientes cada una con n observaciones $Y_{i1},...,Y_{in}$ ($0 \leq i \leq t$) tomadas para $t \geq 1$ tratamientos normales τ_0, ... , τ_t (Bechhofer et al. 1995). Donde t_0 es el tratamiento control y τ_1, ... , τ_t son los tratamientos experimentales.

ii) Los tratamientos τ_i tienen media μ_i desconocida y varianza común desconocida, σ^2 .

Meta: Seleccionar un subconjunto de tratamientos que incluya a todos aquellos elementos con media $\mu_i > \mu_0$. Si la meta es alcanzada, se dice que con una confianza P* ha ocurrido la **selección correcta (SC).**

Los requerimientos probabilísticos del procedimiento se expresan bajo los términos siguientes: Dada una P* constante con $\frac{1}{t} < P^* < 1$ se requiereque la P$\{SC\} \geq P^*$, para toda μ y σ^2, con σ^2 desconocida.

El procedimiento Gupta-Sobel en un experimento balanceado establece lo siguiente: Sean $Y_{i1},...,Y_{in}$ variables aleatorias independientes con distribución común $\eta(\mu, \sigma^2)$.

Procedimiento: Retener en el subconjunto seleccionado aquellos tratamientos τ_i para los cuales se cumpla que:

$$\bar{y}_i > \bar{y}_0 - hs_\nu\sqrt{\frac{2}{n}}$$

Donde n es el número de observaciones, $\bar{y}_i$es la media muestral del tratamiento i-ésimo, $\bar{y}_0$es la media muestral del tratamiento control, s^2_ν es el estimador insesgado combinado de σ^2, s_ν es la desviación estándar muestral y $h=T^{(1-P*)}_{t,\nu,\frac{1}{2}}$ es la t multivariada central con α=(1-P*), t tratamientos y correlación ρ=1/2. Los valores para la t multivariada central y no central (h) se calcularon con el algoritmo AS-251 (Dunnett, 1989; 1993; 1995), programa en lenguaje Fortran al cual se accede vía Internet.

En el procedimiento Gupta-Sobel en un experimento desbalanceado, los supuestos son los mismos que en el caso anterior, excepto que ahora el experimento se diseña con n_0 observaciones para el control y n observaciones para los tratamientos experimentales con $n_0 \neq n$

Procedimiento: Incluir los τ_i tratamientos en el subconjunto seleccionado si y sólo si:

$$\bar{y}_i > \bar{y}_0 - hs_\nu\sqrt{\frac{1}{n} + \frac{1}{n_0}}$$

3 Resultados

Para elegir los mejores tratamientos en la fase de filtrado, los diseños experimentales propuestos corresponden a cuatro diseños estándar en los que la propiedad de ser aumentados no altera su correspondiente modelo lineal, esto se confirma al realizar el análisis de varianza, que se ejecuta de igual modo que en el diseño estándar y exclusivamente sobre los tratamientos control.

i) Diseños experimentales en bloques completos al azar.
ii) Diseños experimentales en cuadros de Youden
iii) Diseños experimentales en látice.
iv) Diseños experimentales en bloques enlazados.

En el caso tener repeticiones del material experimental se propone el procedimiento de Gupta y Sobel con un tratamiento control, σ^2 desconocida, en situaciones experimentales balanceadas y desbalanceadas; fue diseñado un programa de cómputo que permite al usuario manejar datos de tratamientos, elaborar e imprimir reportes de ellos.

La Figura 1 muestra la pantalla principal de la aplicación que ejecuta la selección del mejor subconjunto de tratamientos. Como puede observarse todos los procedimientos de selección son invocados desde esta pantalla por algún componente de la misma, el acceso a la rutina deseada se logra oprimiendo con el mouse la letra de la columna derecha correspondiente al procedimiento.

SELECCION DEL MEJOR SUBCONJUNTO DE TRATAMIENTOS	
Procedimiento G-S en un DCA	C
Procedimiento G-S en un DBCA	B
Procedimiento G-S en un D. desbalanceado	D
Ayuda	A
Salir a un nuevo libro	S
Presione el botón correspondiente al procedimiento deseado	

Figura 1. Menú principal para la selección del mejor subconjunto de tratamientos.

Al ejecutar la opción **C** del menú principal, el algoritmo seleccionará el mejor subconjunto de tratamientos cuando el experimento se realiza bajo un diseño completamente al azar, invocándose la pantalla de introducción de datos.

El programa muestra un resumen del procedimiento reportando las medias de tratamientos y la conclusión de la prueba, es decir, muestra la etiqueta "seleccionado" si el tratamiento pertenece al mejor subconjunto o "eliminado" de otro modo.

La segunda opción de la pantalla principal (**B**) corresponde al procedimiento Gupta-Sobel bajo un diseño de bloques completos al azar.. La tercera opción (**D**) de la pantalla principal evalúa el procedimiento Gupta-Sobel a los niveles de confianza de P*=0.99, P*=0.95, P*=0.90 y P*=0.80 cuando las observaciones de los tratamientos provienen de un diseño experimental desbalanceado.

4 Conclusiones

En el presente documento se proporcionan herramientas estadísticas las cuales se adaptan a experimentos que involucran varias repeticiones, con el objetivo común de seleccionar el subconjunto más pequeño que contiene a los mejores tratamientos a un nivel de confianza especificado. El procedimiento de análisis estadístico propuesto fue sistematizado y apoyado con procedimientos de cómputo usando como software de desarrollo Excel y SAS.

Si se desea estimar los efectos de nuevos tratamientos con una sola repetición, los diseños experimentales aumentados son más eficientes y económicos que los diseños estándar porque permiten reducir el tamaño del experimento, evaluar el cuadrado medio del error y realizar diversas comparaciones entre tratamientos.

El procedimiento de análisis estadístico propuesto fue sistematizado y apoyado con procedimientos de cómputo usando los programas Excel y SAS.

Referencias

Bechhofer, R.E., T. J. Santner, and D. M. Goldsman. (1995). *Design and analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons.* John Wiley. New York 325 p.

Dunnett, C. (1995). A multiple comparison precedure for comparing several treatments with a control. *Journal of the American Statistical Association* 50:1096-1121.

Dunnett, C. (1989). Algorithm AS-251: Multivariate Normal Probability Integrals with Product Correlation Structure. *Applied Statistics* 38:564-579.

Dunnett, C. (1993). Correction to Algorithm AS-251: Multivariate Normal probability integrals. *Applied Statistics* 42:709.

Federer, W.T. (1961). Augmented designs with one-way elimination of heterogeneity. *Biometrics.* 17:447-473.

Federer, W.T. and D. Raghavarao. (1975). On Augmented Designs. *Biometrics* 31: 29-35.

Esta publicación consta de 410 ejemplares y se terminó de
imprimir en el mes de abril del 2000 en los talleres gráficos del
Instituto Nacional de Estadística, Geografía e Informática
Av. Héroe de Nacozari Núm. 2301 Sur, Acceso 11, PB
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.
México