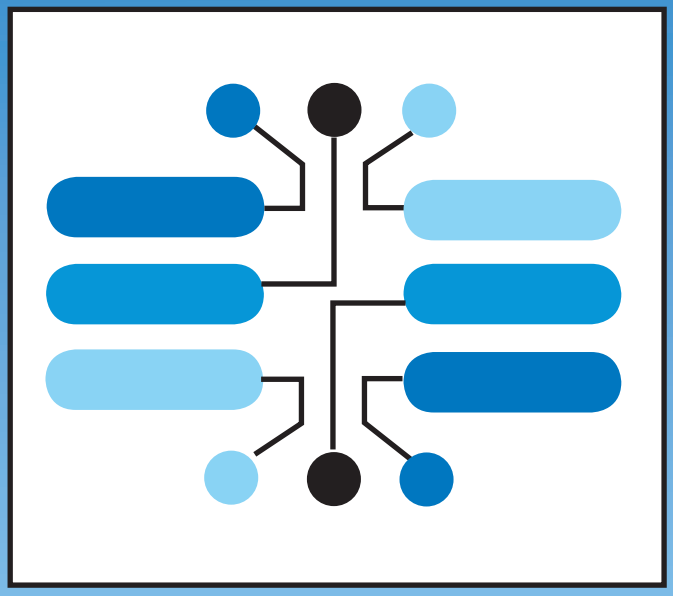
MEMORIA DEL VIII FORO NACIONAL DE ESTADISTICA

INSTITUTO NACIONAL DE ESTADISTICA GEOGRAFIA E INFORMATICA

DEL 27 DE SEPTIEMBRE AL 1° DE OCTUBRE DE 1993



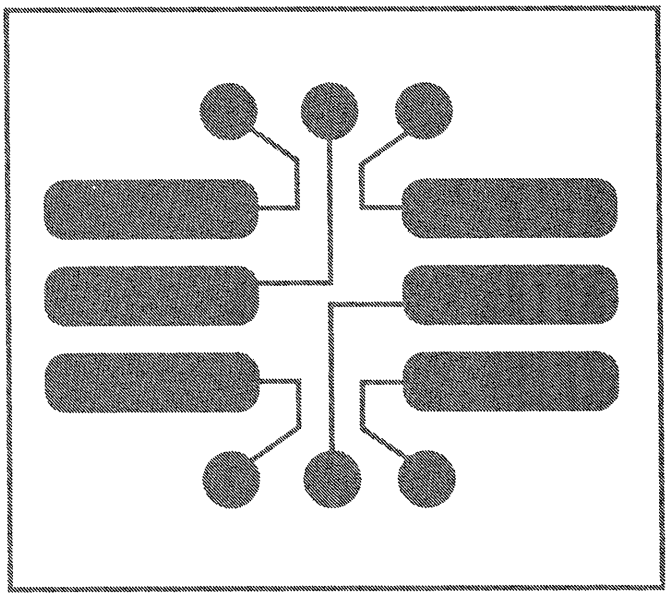
INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

AME

MEMORIA DEL VIII FORO NACIONAL DE ESTADISTICA

INSTITUTO NACIONAL DE ESTADISTICA
GEOGRAFIA E INFORMATICA

DEL 27 DE SEPTIEMBRE AL 1º DE OCTUBRE
DE 1993



INEGI
INSTITUTO NACIONAL DE ESTADISTICA
GEOGRAFIA E INFORMATICA

AME

Memoria del VIII Foro Nacional de Estadística

Impreso en México
ISBN 970-13-0461-6

CONTENIDO

PRESENTACION

En este volumen se presentan los resúmenes de cuatro Conferencias Magnas y de once contribuciones libres aceptadas por el Comité Editorial de la Memorias del VIII Foro Nacional de Estadística. Dicho evento se llevó a cabo del 27 de septiembre al 1 de octubre de 1993 en la sede del INEGI de la Cd. de Aguascalientes, Ags.

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

El Comité Editorial
Agosto 1994

Memoria del VIII Foro Nacional de Estadística

Resúmenes "in extenso"

Editado por:

Belem Trejo Valdivia - IIMAS, UNAM
Rubén Hernández Cid - ITAM

A. M E.
Asociación Mexicana de Estadística

I.N.E.G.I.
Instituto Nacional de Estadística,
Geografía e Informática

CONFERENCIAS

MAGNAS

Pronósticos con restricciones en series de tiempo univariadas

Víctor M. Guerrero
Departamento de Estadística y Actuaría
Instituto Tecnológico Autónomo de México

Introducción.

Los pronósticos de series de tiempo univariadas, comúnmente hacen uso eficiente de la información histórica disponible sobre la variable involucrada. En la práctica, sin embargo, muchas veces se cuenta con información adicional, que no es considerada por el modelo de la serie. No hacer uso de ella constituye un desperdicio que debe ser evitado. Si dicha información adicional se refiere al comportamiento futuro de la variable y está dada en forma de restricciones lineales, su incorporación en los pronósticos puede realizarse de manera óptima mediante argumentos formales de la Estadística.

Aquí se muestran las soluciones que corresponden a diferentes situaciones respecto a la compatibilidad entre las restricciones y la información histórica, en el contexto de modelos ARIMA (autorregresivos, integrados y de promedios móviles). Estas son: a) restricciones ciertas, compatibles con la historia de la serie, b) restricciones inciertas, compatibles o no con la información histórica, y restricciones ciertas, pero incompatibles con la historia debido a un cambio previsible del modelo, ya sea en c) su estructura determinista, d) su parte estocástica, o e) en los valores de sus parámetros.

Notación básica.

Sea $\{Z_t\}$ la serie de tiempo para la cual se desea obtener pronósticos y que admite una representación de modelo ARIMA. El conjunto de datos observados (históricos) será denotado por el vector $Z_0 = (Z_1, \ldots, Z_N)$, mientras que los pronósticos se requieren para el siguiente vector de valores futuros $Z_f = (Z_{N+1}, \ldots, Z_{N+H})$, con H el horizonte de pronóstico. Es bien sabido que el pronóstico óptimo, dado Z_0, está dado por la esperanza condicional $E(Z_f | Z_0)$, cuando el criterio de optimalidad es el del Error Cuadrático Medio Mínimo del error de pronóstico. Por otro lado, se supone que existe información adicional en el vector $\lambda = (\lambda_1, \ldots, \lambda_m)'$, dada en forma de restricciones lineales, es decir, $\lambda = CZ_f$ con C una matriz de constantes conocida.

De esta forma se tienen las siguientes dos fuentes de información

acerca del futuro: $E(Z_t | Z_0)$ y λ. Es necesario entonces verificar la compatibilidad entre fuentes de información para combinarlas apropiadamente, para ello se debe desarrollar un prueba estadística de compatibilidad. Si las fuentes de información resultan ser compatibles (empíricamente), se requiere derivar fórmulas para obtener pronósticos con restricciones y sus correspondientes varianzas.

Como resultados posibles de la verificación de compatibilidad, se tienen los siguientes:

1) Ambas fuentes válidas, lo que da origen a la técnica de "pronósticos con restricciones ciertas". Caso que fue considerado por Cholette (1982) con modelos puramente autorregresivos y por Guerrero (1989) con modelos ARIMA en general.

2) Información histórica válida para el futuro e información adicional incierta, es decir, $E(Z_t | Z_0)$ válida y λ inválida. Lo cual conduce a obtener "pronósticos con restricciones inciertas". Como ejemplos de esta situación, se tiene por un lado, la combinación de pronósticos de modelos alternativos al modelo de series de tiempo, y por el otro, la incorporación de conjeturas o juicios de expertos a los pronósticos ARIMA. Como parte fundamental de este caso, se requiere de un procedimiento para asignar la incertidumbre a λ, la cual, si se trata de utilizar un modelo alternativo al de series de tiempo, surge de la estimación de varianza del modelo. Nuevamente, este caso aparece considerado en Guerrero (1989), pero también es atacado por Pankratz (1989) y por Trabelsi y Hillmer (1989), con diferentes enfoques de solución, pero con resultados equivalentes.

3) Información adicional λ válida, mientras que los pronosticos del modelo ARIMA, $E(Z_t | Z_0)$, son invalidados por cambios previsibles en: i) la estructura determinista del modelo, que conduciría a realizar un análisis de intervención ex-ante, o en ii) la estructura estocástica del modelo. En cualquiera de estos dos casos, se requiere postular la nueva estructura y estimar los parámetros correspondientes. La solución de estos dos problemas se muestra en Guerrero (1991). La tercera posibilidad es que se presente un cambio en iii) los valores de los parámetros del modelo ARIMA original. Para este caso se requiere determinar aquellos parámetros que tengan cambio significativo, en términos estadísticos. La correspondiente solución aparece en Guerrero (1990).

Resumen de resultados.

La derivación de las fórmulas de pronosticos restringidos, surge de la siguiente expresión para el error del pronóstico ARIMA, para $h = 1, \ldots, H$

$$Z_{N+h} - E(Z_{N+h} \mid \mathbf{Z}_0) = \sum_{j=0}^{\infty} \psi_j a_{N+h-j}$$

en donde las ψ_j, $j = 0, 1, \ldots$ son obtenidas de los parámetros del modelo original y de las a's, que provienen de un proceso de ruido blanco Gaussiano, con varianza de σ_a^2. Dicha expresión, en notación matricial, se convierte en

$$\mathbf{Z}_F - E(\mathbf{Z}_F \mid \mathbf{Z}_0) = \Psi \mathbf{a}_F$$

donde Ψ es la matriz triangular interior cor elementos $1, \psi_1, \ldots, \psi_{H-1}$ en su primera columna, $0, 1, \psi_1, \ldots, \psi_{H-2}$ en la segunda columna, y así sucesivamente. Por su lado, $\mathbf{a}_F = (a_{N+1}, \ldots, a_{N+H})'$ es tal que $E(\mathbf{a}_F \mid \mathbf{Z}_0) = \mathbf{0}$ y $E(\mathbf{a}_F \mathbf{a}_F' \mid \mathbf{Z}_0) = \sigma_a^2 I$. Asimismo, la información adicional tiene las siguientes características, $\mathbf{Y} = C\mathbf{Z}_F + \mathbf{U}$ con el error $\mathbf{U} \sim N(\mathbf{0}, U)$, además de que se supone que dicho error no está correlacionado ni con la historia, ni con el futuro de la serie, es decir $I(\mathbf{Z}_0 \mathbf{U}') = 0$ y $E(\mathbf{a}_F \mathbf{U}') = 0$.

A partir de lo anterior, al minimizar el error cuadrático medio del pronóstico, se obtienen las siguientes fórmulas generales, en donde el asterisco denota el caso particular que se considera.

Pronostico restringido óptimo

$$\hat{\mathbf{Z}}_{F^*} = E(\mathbf{Z}_F \mid \mathbf{Z}_0) + A^* [\mathbf{Y} - CE(\mathbf{Z}_F \mid \mathbf{Z}_0)]$$

Covarianza del error de pronóstico, dadas $\mathbf{Z}_0$ y $\mathbf{Y}$

$$Cov(\mathbf{Z}_F - \hat{\mathbf{Z}}_{F^*}) = Cov[\mathbf{Z}_F - E(\mathbf{Z}_F \mid \mathbf{Z}_0)] - \sigma_a^2 \Psi \Psi' C' A^{*\prime} - M^*$$

Prueba estadística de compatibilidad entre fuentes de información

$$K^* = [\mathbf{Y} - CE(\mathbf{Z}_F \mid \mathbf{Z}_0)]' (\sigma_a^2 C \Psi \Psi' C' + U + M^*)^{-1} [\mathbf{Y} - CE(\mathbf{Z}_F \mid \mathbf{Z}_0)] \sim \chi^2_{g.l.^*}$$

La especificación de las fórmulas previas, se muestra a continuación para cada uno de los casos en consideración.

Caso 1) Restricción cierta. Para este caso se tiene que no hay incertidumbre en la información adicional, así que $U = 0$, con $A^* = \Psi \Psi' C' (C \Psi \Psi' C')^{-1}$, $M^* = 0$ y $g.l.^* = m$ (el número de restricciones linealmente independientes), con $m \leq H$.

Caso 2) Restricción incierta. Ahora se tiene $U \neq 0$ y se requiere especificar esta matriz, lo cual se logra al nivel de significación α cuando $K^* < \chi^2_m(\alpha)$. Además $A^* = \sigma_a^2 \Psi \Psi' C' (\sigma_a^2 C \Psi \Psi' C' + U)^{-1}$, $M^* = 0$ y nuevamente $g.l.^* = m$.

Caso 3) Restricción cierta con cambio en la estructura determinista del modelo ARIMA. Para este caso, $U = 0$ y se debe especificar una función dinámica de intervención, por ejemplo de primer orden, dada por $D_t^{N+1} = \omega(1-\delta^{t-N})/(1-\delta)$ si $t > N$. Por otro lado, la matriz A_* resulta ser una inversa generalizada de C, mientras que $M_* = 0$ y $g.l.* = m$.

Caso 4) Restricción cierta con cambio en la estructura estocástica del modelo. Nuevamente se tiene $U = 0$ y se supone que el cambio es debido a contaminación con otro modelo. La consideración más simple de contaminación es por ruido blanco $N(0, \sigma_v^2)$, lo cual conduce a obtener $A_* = (\sigma_a^2 \Psi \Psi' C' + \sigma_v^2 C')(\sigma_a^2 C \Psi \Psi' C' + \sigma_v^2 CC')^{-1}$, $M_* = \sigma_v^2 I$ y $g.l.* = m$.

Caso 5) Restricción cierta con cambio en los parámetros del modelo ARIMA, que se supone inalterado. Así pues, $U = 0$ en esta situación, con parámetros estimados a partir de la historia, que producen la matriz Ψ^0 y los pronósticos $\hat{E}^0(Z_F \mid Z_0)$, mientras que los parámetros estimados haciendo uso de Z_0 y de Y, proporcionan la correspondiente matriz y pronósticos $\hat{\Psi}$ y $\hat{E}(Z_F \mid Z_0, Y)$. Ahora se obtiene entonces $I_* = \hat{\Psi}\hat{\Psi}' C'(C\hat{\Psi}\hat{\Psi}' C')^{-1}$, $M_* = 0$ y $g.l.* = k \leq m$, con k el número de parámetros del modelo. Conviene indicar que se deben realizar pruebas de significación estadística individuales, para determinar qué parámetros cambian significativamente su valor. Además, conviene dar una interpretación de los cambios en los parámetros, de preferencia en términos de componentes no observables de la serie (tendencia y estacionalidad), haciendo uso pra ello de equivalencias entre modelos estructurales y ARIMA's.

Comentarios finales.

El área de pronósticos con restricciones se ha visto enriquecida con la aparición de otros enfoques de solución, en donde sobresalen el bayesiano (véase de Alba, 1993) y el de mínimos cuadrados generalizados (Alvarez, del Rieu y Jareño, 1993). Asimismo, en lo que toca a aplicaciones de este enfoque para solucionar el problema de combinar información histórica y preliminar (Guerrero, 1993) o el de completar series con datos faltantes (Guerrero, 1993; Nieto, 1994; Nieto y Martínez, 1994). Por otro lado, la aplicación de esta técnica a otra clase de modelos de series de tiempo, es una posibilidad que ya ha sido explorada por Pankratz (1989), al considerar series múltiples, y por Rosas y Guerrero (1994) en métodos de suavizamiento exponencial.

Referencias.

Alvarez, L.J., Delrieu, J.C. y Jareño, J. (1993) "Tratamiento de predicciones conflictivas: empleo eficiente de información extramuestral". **Estadística Española 35**, 439-461.

Cholette, P.A. (1982) "Prior information and ARIMA forecasting". **Journal of Forecasting 1**, 375-383.

de Alba, E. (1993) "Constrained forecasting in autoregressive time series models: A Bayesian analysis". **International Journal of Forecasting 9**, 95-108.

Guerrero, V.M. (1989) "Optimal conditional ARIMA forecasts". **Journal of Forecasting 8**, 215-229.

Guerrero, V.M. (1990) "Restricted ARIMA forecasts which account for parameter changes". **ESTADISTICA 42**, 17-31.

Guerrero, V.M. (1991) "ARIMA forecasts with restrictions derived from a structural change". **International Journal of Forecasting 7**, 339-347.

Guerrero, V.M. (1993) "Combining historical and preliminary information to obtain timely time series data". **International Journal of Forecasting 9**,

Guerrero, V.M. (1993) "Restricted forecasts of missing observations in univariate time series". Documento de trabajo DEA-C93.3, Instituto Tecnológico Autónomo de México.

Nieto, F.H. (1994) "Una nota sobre la estimación de datos faltantes en una serie temporal, usando la función de autocorrelación dual". Reporte Interno del Departamento de Matemáticas y Estadística, Facultad de Ciencias, Universidad Nacional de Colombia.

Nieto, F.H. y Martínez, J. (1994) "A recursive approach for estimating missing observations in a univariate time series". Reporte Interno del Departamento de Matemáticas y Estadística, Facultad de Ciencias, Universidad Nacional de Colombia.

Pankratz, A. (1989) "Time series forecasts and extra-model information". **Journal of Forecasting 8**, 75-83.

Rosas, L. y Guerrero, V.M. (1994) "Restricted forecasts using exponential smoothing techniques". **International Journal of Forecasting 10.**

Tabelsi, A. y Hillmer, S.C. (1989) "A benchmarking approach to forecast combination". **Journal of Business and Economic Statistics 7**, 353-362.

Sobre el desarrollo de los métodos estadísticos bayesianos.

MANUEL MENDOZA
Departamento de Estadistica. ITAM
Rio Hondo 1. San Angel México D.F. 01000 MEXICO.

RESUMEN

1. Introducción

Cada vez en un mayor grado, el enfoque bayesiano de la estadistica va incorporándose al acervo de conocimientos básicos de la disciplina y por tanto se encuentra, también con mayor frecuencia, al alcance tanto de los estadísticos profesionales como de los usuarios de las técnicas estadisticas. Los conceptos fundamentales de análisis bayesiano, en ocasiones con una interpretación ciertamente limitada, empiezan a formar parte de la cultura mínima de la comunidad. De una primera fase en la que el centro de las discusiones, entre bayesianos y no-bayesianos, se localizaba en el nivel conceptual e incluso filosófico, se ha trascendido a otro en donde lo importante son los resultados prácticos que, en la aplicación, se pueden obtener de los métodos.

Precisamente ahí, en la implementación de los métodos es donde el enfoque bayesiano ha encontrado algunos de sus más grandes retos. La necesidad de cálculos numericos relativamente elaborados, que por su especificidad no estaban contemplados por la herramienta estadística habitual, ha limitado la aplicabilidad de las técnicas bayesianas. Más aún, la idea de que un análisis estadístico bayesiano (un análisis estadistico, en general) requiere de un estudio de sensibilidad que determine la validez de los supuestos y el peso de la componente subjetiva, ha conducido a la situación en la cual los procedimientos numericos no sólo son necesarios sino que deben satisfacer criterios de eficiencia, reproductibilidad y transportabilidad para permitir un empleo repetido, intensivo, en cada aplicación.

Esta limitación ciertamente ha tenido un impacto en el desarrollo de la alternativa bayesiana. Durante años se ha manifestado la necesidad de contar con paquetes cómputo bayesianos y ciertamente, existen importantes avances en esa dirección. Es previsible que en el futuro cercano aparezca una variedad de productos que seguramente aminorará esta deficiencia. Sin embargo, probablemente la consecuencia más importante, y en alguna forma inesperada, de este desarrollo sea una reconsideración completa de la manera en que puede concebirse y realizarse un análisis estadistico cualquiera.

2. La estructura del análisis estadístico en general.

Para ningún estadistico constituye una novedad que el análisis (estadistico) tiene como propósito la descripción de fenómenos aleatorios comúnmente a partir de la información contenida en muestras propabilísticas. Con distintas variantes, las técnicas estadisticas se agrupan en diferentes clases de acuerdo con el fin especifico que persiguen asi como los supuestos que involucran. La variedad de procedimientos ha causado que la disciplina sea percibida, y también transmitida, como una vasta colección de técnicas particulares y heterogéneas con un sustrato común relativamente vago. La manera como, en un problema

concreto, se integran distintas técnicas estadísticas parece algo que sólo en la práctica profesional puede aprenderse.

Es claro que en todo problema estadístico juegan, o deberían de jugar, un papel importante las técnicas de muestreo, de análisis exploratorio de datos, de inferencia -paramétrica y no paramétrica y de diagnóstico. Específicamente, el ciclo formado por el análisis exploratorio de datos, la inferencia y el diagnóstico, es bien conocido pero sólo en ámbitos muy particulares como el análisis de regresión ocurre que, en forma rutinaria estos tres elementos se integran para dar lugar a una aplicación real completa del análisis estadístico. En cualquier caso, no debe perderse de vista que el objetivo último de las técnicas estadísticas es obtener una descripción razonable y eficiente del fenómeno bajo estudio. En términos muy generales, (Box 1976, por ejemplo) una descripción de este tipo debe contar con las características de relevancia, flexibilidad y parsimonia. Estos es, debe incluir todos los aspectos relevantes del fenómeno, debe también ser general como para poder transportarse a fenómenos similares y finalmente, debe ser lo más simple y concreta posible.

3. Las características básicas del análisis estadístico bayesiano.

La naturaleza del enfoque bayesiano ha sido descrita por distintos autores (DeGroot 1970, Lindley 1971, Berger 1985 y Bernardo y Smith 1994 entre otros). En resumen, una de las más importantes ventajas del enfoque bayesiano, desde un punto de vista estructural, radica precisamente en que concibe el proceso de análisis estadístico como un problema de decisión en ambiente de incertidumbre. De esta manera, la teoría de la decisión proporciona un sustrato común para identificar, formular y resolver cualquier problema estadístico. No solamente se clarifica el papel y la relevancia de cada uno de los componentes del problema y se abre la posibilidad de incorporar toda la información disponible, subjetiva o muestral, sino que además, si se adopta un enfoque axiomático, la solución general es queda totalemente determinada. Se debe *maximizar la utilidad esperada.*

Aquí es conveniente insistir en que el análisis estadístico bayesiano *no* consiste, como en ocasiones se confunde en el contexto de la inferencia paramétrica, en considerar a los parámetros como variables aleatorias, asignar una distribución a esas nuevas *variables* y combinar esa información con la información muestral: a través del teorema de Bayes, para obtener la distribución final.

Como se ha indicado, la esencia del análisis estadístico bayesiano consiste en abordar los procesos de análisis estadístico como problemas de decisión. Para tal fin es necesario identificar las opciones factibles, los sucesos inciertos que pueden modificar las consecuencias asociadas a la elección de cada una de las opciones y las consecuencias mismas. Asimismo, es indispensable cuantificar las preferencias del tomador de decisones respecto a las posibles consecuencias y el grado de conocimiento que posea sobre la posible ocurrencia de los sucesos inciertos. Finalmente, como consecuencia de los axiomas de coherencia, el proceso conduce a seleccionar como la mejor opción a la que minimiza la pérdida esperada.

Esta solución general que es óptima y coherente, sin embargo debe apreciarse en su justa dimensión. En primer lugar, este criterio no invalida o descalifica automáticamente otros procedimientos no bayesianos. Además, tampoco es una garantía de infalibilidad en una aplicación concreta. El criterio general es óptimo pero requiere de la formulación adecuada del problema y de la especificación de una función de pérdida y un modelo de probabilidad para describir el *conocimiento* del tomador de decisiones sobre los sucesos inciertos. Si alguno de estos elementos está mal planteado o simplemente no corresponde con la situación real que se esta abordando puede obtenerse la solución óptima para un problema equivocado y los resultados pueden ser, en el mejor de los casos, inútiles.

4. El análisis bayesiano en la práctica.

Una de las críticas más frecuentes al análisis bayesiano ha sido tradicionalmente la que se dirige a cuestionar la incorportación explícita de información subjetiva en el proceso de análisis. Si bien la subjetividad de las preferencias de un tomador de decisiones no aparece como un elemento polémico, la situación respecto a la distribución de probabilidades que describe su conocimento sobre los sucesos inciertos ha resultado polémica.

Claramente, esa información influye en los resultados finales y desde la perspectiva bayesiana esa influencia no sólo es legítima sino indispensable. Sin embargo, el riesgo de que una mala especificación de esa información, o peor aún una especificación tendenciosa , puede contrarestar la evidencia contenida en la información muestral dando lugar a conclusiones preconcebidas y posiblemente falsas. Es por esa razón que para el análisis bayesiano es aún más importante contemnplar como parte básica de cualquier aplicación, un análisis de sensiblidad que determine el peso que las diferentes fuentes de imformación tienen en los resultados finales.

Ahora bien, para poder llevar a cabo un análisis de sensibilidad como el que se indica, es necesario contar con los medios para efectuar de manera repetida los cálculos que conducen a las distribuciones *a posteriori*

$$p(\theta|Z) = \frac{p(Z|\theta)p(\theta)}{\int p(Z|\theta)p(\theta)d\theta}$$

predictiva

$$p(x|Z) = \int p(x|0)p(\theta|Z)d\theta$$

y *marginales*

$$p(\phi|\varphi) = \int p(\phi,\varphi|Z)d\varphi$$

en donde, $\theta = (\phi,\varphi)$ con φ un parámetro de estorbo Precisamente es aquí en donde han aparecido algunas de las más notables dificultades para implementar las técnicas bayesianas. Concretamente, ocurre con frecuencia que las integrales que involucradas no tienen una expresión cerrada excepto para casos muy específicos de las distribuciones iniciales y en general, es necesario recurrir a métodos numéricos para obtener resultados. Algunas de las primeras soluciones que se propusieron para este tipo de problemas fué restringir la selección de distribuciones iniciales a familias que facilitaran los cálculos y más aún, que permitieran la obtención de expresiones cerradas (las llamadas familias *conjugadas*).

Esta restricción, si bien funciona aceptablemente en una variedad de situaciones y existen resultados que establecen un grado razonable de generalidad (Diaconis y Ylvisaker 1985, por ejemplo), limita de cualquier manera la posibilidad de un análisis de sensibilidad completo.

Más recientemente, el trabajo de investigación se he dirigido hacia el empleo y el desarrollo de métodos de aproximación numérica para el cálculo de estas y otras integrales que aparecen en el análisis bayesiano con el propósito de proveer de buenas aproximaciones que además sean eficientes, de bajo costo y fácil implementación. Esencialmente se pueden mencionar dos tipos de aproximaciones. Las aproximaciones determinísticas como por ejemplo, la aproximación de Laplace o los métodos de cuadratura en sus distintas variantes. Una buena discusión de estos métodos puede encontrarse en Kass, Tierney y Kadane, 1988.

Por otra parte y es aquí en donde el progreso ha tenido un muy importante impacto no sólo en los aspectos computacionales sino también en el futuro completo del análisis bayesiano, son los métodos estadísticos de aproximación. Todos ellos están basados en algoritmos de simulación y en una primera etapa estos métodos utilizaron la idea general de aproximar las integrales de interés con promedios muestrales que convergen a los respectivos valores esperados.

Existen distintas estrategias para generar observaciones de distribuciones fáciles de simular que produzcan los resultados preyistos. Un recuento de las diferentes técnicas de este tipo se encuentra en Naylor y Smith 1988. Aún así, las dificultades con problemas que involucran parámetros de alta dimensionalidad son considerables y sólo hasta muy recientemente han aparecido otros algoritmos más potentes basados en la idea de generar observaciones de las distribuciones de interés para, a partir de ellas, aproximar cualquier tipo de integral, ya sea basada en la distribución final completa o de cualquier distribución final marginal.

Más aún, la posibilidad de contar con muestras de la distribución final de las cantidades de interés, sin importar la dimensión del parámetro, permite la aplicación de técnicas del análisis exploratorio para establecer sus características más relevantes aún cuando no se conozca explícitamente su función de probabilidad o densidad. De entre los métodos de este tipo destaca muy especialmente el muestreador de Gibbs (Geman y Geman 1984). Esta técnica y algunas otras del mismo tipo, como las descritas en Tanner 1991, se basan en algoritmos iterativos y la investigación se encuentra actualmente activa con el propósito de establecer la rapidez de convergencia así como los volúmenes de iteración adecuados.

Conceptualmente, sin embargo, estos procedimientos, los que producen muestras de la distribución final, han tenido un impacto mayor. Problemas que tanto desde una perspectiva bayesiana como frecuentista han sido tradicionalmente considerados intratables o, al menos, complejos, pueden ser ahora abordados con este enfoque vía simulación y análisis exploratorio.

Un caso muy interesante es el de observaciones faltantes. Es bien conocido que la pérdida de observaciones en muchos casos implica un desbalance en la estructura de modelado una de cuyas consecuencias es un incremento en la complejidad del análisis y con fercuencia la imposibilidad de obtener soluciones explícitas de los procedimentos de inferencia. Con la idea de análisis vía el muetreador de Gibbs y sus similares, las observaciones faltantes se pueden considerar magnitudes desconocidas, exactamente igual que los parámetros, y entonces, sin importar su número, pueden obtenerse muestras de la distribución *conjunta* final de los parámetros y esas observaciones.

Más importante aún, es posible obtener muestras de la distribución final *marginal* de los parámetros que naturalmente incorporan la incertidumbre debida a la ausencia de las observaciones perdidas. Esta idea se puede aplicar a una variedad de otros problemas complejos como por ejemplo, datos con censura y mezclas de distribuciones. Estas ideas y otras relacionadas han sido presentadas recientemente por Roberts y Smith 1993.

5. Comentarios finales.

El empleo de los algoritmos numéricos basados en simulación de muestras de la distribución de interés, como el muestreador de Gibbs, para la implementación de las técnicas bayesianas no solamente permite la consideración de una gama de variantes para las distribuciones iniciales. Además de auxiliar en la incorporación del análisis de sensibilidad como una parte rutinaria de la inferencia bayesiana, establece un nuevo papel para las técnicas de análisis exploratorio, ahora aplicadas a la caracterización de las correspondientes distribuciones finales.

Más aún, abriendo todo un nuevo horizonte para la investigación y el desarrollo de nuevas técnicas, permite el tratamiento de problemas que hasta ahora eran considerados inaccesibles por su grado de complejidad. Seguramente en los próximos años aparecerán un buen número de contribuciones en la literatura abordando estos problemas en forma específica y demandando un soprte número y de simulación aún más potente, Sin embargo, la idea del análisis estadístico como un proceso exploratorio, aún en los casos de inferencia paramétrica completamente modelada queda abierta como una opción en el futuro.

Referencias.

Berger, J.O. (1985). *Statistical Decision Theory an Bayesian Analysis.* New York : Springer-Verlag.

Bernardo, J..M. y Smith A.F.M. (1994). *Bayesian Theory.*Chichester : Wiley

Box, G.E.P. (1976). Science and Statsitics. *Journal of the American Statistical Association*, **71**, 356.

DeGroot, M.H. (1970). *Optimal Statistical Decisions.* New York : McGraw-Hill

Diaconis, P. y Ylvisaker, D. (1985). Quantifying prior opinion. *Bayesian Statistics 2.* (Bernardo, J.M., DeGroot. M.H., Lindley, D.V. and Smith, A.F.M. eds.). Amsterdam : North-Holland.

Geman, S, y Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intelligence.* **6**, 721-740.

Kass, R.E., Tierney, L. y Kadane, J.B. (1988). Asymptotics in bayesian computation. *Bayesian Statistics 3.* (Bernardo, J.M., DeGroot. M.H., Lindley, D.V. and Smith, A.F.M. eds.). Amsterdam : North-Holland.

Lindley, D.V. (1971). *Making Decisions.* Chichester : Wiley.

Naylor, J.C. y Smith, A.F.M. (1988). Economic illustrations of novel numerical integration methodology for bayesian inference. *J. Econometrics*, **38**, 103-125.

Dos ejemplos de funciones generatrices utilizadas para realizar inferencia estadística

Miguel Nakamura Savoy

Centro de Investigación en Matemáticas Apdo. Postal 402, Guanajuato, Gto. 36000

1 Introducción.

Es usual que las funciones generatrices de momentos ($m_X(t) = E\exp(tX)$) y de probabilidad ($\varphi_n(t) = E(t^X)$), así como la función característica ($C_X(t) = E\exp(itX)$) de una variable aleatoria X, se introduzcan en cursos de probabilidad y estadística como poderosos instrumentos teóricos que son útiles para demostrar varios resultados (tales como el teorema central del límite o la ley de los grandes números). Es menos común que en estos cursos se preste atención al empleo que pueden darse a las versiones empíricas de dichas funciones, basadas en observaciones $X_1, \ldots, X_n$, y definidas por

$$m_n(t) = (1/n)\sum_{j=1}^{n} \exp(tX_i),$$

$$\varphi_n(t) = (1/n)\sum_{j=1}^{n} t^{X_j},$$

y

$$C_n(t) = (1/n)\sum_{j=1}^{n} \exp(itX_j),$$

respectivamente. Dichos conceptos *empíricos* son casos particulares de los llamados *métodos basados en transformadas estadísticas*. Su utilidad como

herramientas de inferencia estadística se fundamenta en el hecho de que las versiones empíricas pueden aproximar a las funciones generatrices teóricas, cuando $n \to \infty$ (ver [3]).

Mediante una amplia colección de ejemplos se presentarán aquí algunas aplicaciones de la función generatriz de probabilidad empírica y la función característica empírica en problemas de estimación, bondad de ajuste, identificación de modelos, y estimación de transformaciones. El propósito es señalar el potencial de uso que tienen las funciones empíricas en diversos contextos, coincidiendo con el desarrollo que se ha registrado recientemente en esta materia.

2 Función Generatriz de probabilidad empírica.

Para una reseña general acerca de $\varphi_n(t)$, puede consultarse [4]. Aquí hacemos solo breves ilustraciones acerca de su utilización en diversos problemas de inferencia estadística, comenzando con una sencilla técnica exploratoria gráfica. En efecto, la gráfica de la función $Y_n(t) = \ln(\varphi_n(t))$ para $0 \leq t \leq 1$ es un instrumento útil en la identificación de la distribución de una variable aleatoria de conteo, debido a que la forma de esta función posee cualidades específicas que dependen de la variable. Por ejemplo, si las observaciones en la muestra poseen distribución de Poisson, entonces la gráfica de $Y_n(t)$ se aproximará a una línea recta. Esta particularidad se debe a que si X es una variable aleatoria con distribución de Poisson con media λ, entonces $\ln(\varphi_X(t)) = \lambda(t-1)$. Otras peculiaridades de la gráfica de $Y_n(t)$ contra t—relacionadas con la convexidad o concavidad de la misma—permitirán identificar distribuciones binomiales o binomiales negativas, así como discretas de colas pesadas. Por otra parte, la gráfica de dos o más de estas funciones construidas con muestras provenientes de distintas poblaciones, permite llevar a cabo comparaciones de tipo exploratorio. Un ejemplo de ésto último es como sigue: supongamos que dos laboratorios distintos realizan conteos de colonias de bacterias sobre una muestra de medios de cultivo. Es de interés explorar acerca de la distribución de tienen los conteos, así como determinar si ambos laboratorios dan lugar a la misma distribución. Es interesante notar que a pesar de que los datos de conteo poseen distribuciones que no son

continuas, la función $Y_n(t)$ es siempre continua, lo cual la hace más agradable. En [1], puede encontrarse otro ejemplo de la utilización de $\varphi_n(t)$ para identificar distribuciones discretas, en un contexto de contaminación del aire a lo largo de las estaciones del año.

La función $\varphi_n(t)$ ha motivado, por otra parte, varias pruebas analíticas de bondad de ajuste diseñadas para datos de conteo, particularmente para la distribución de Poisson. El problema de bondad de ajuste, en esencia consiste en determinar si a un juego de datos puede atribuírsele una distribución de probabilidad determinada de antemano. Todas las pruebas tienen en común el que aprovechan simultáneamente las propiedades que tiene $\varphi_n(t)$ como aproximación de $\varphi_X(t)$, y las propiedades que tiene la familia de distribuciones que es de interés (ver lista de referencias contenidas en [4]). Por ejemplo, como se mencionó anteriormente, cualquier distribución de Poisson posee una función generatriz de probabilidad que es de la forma $\exp(\lambda(t-1))$, de modo que si la muestra es auténticamente Poisson, entonces $Y_n(t)$ debería ser una función próxima a una línea recta. Esto permite basar un criterio de bondad de ajuste en una cuantificación de la medida en que $Y_n(t)$ dista de ser una recta como función de t.

La función generatriz de probabilidades empírica también puede emplearse en la estimación de parámetros en familias de distribuciones discretas. Supongamos que se ha especificado una familia de distribuciones a través de funciones generatrices de probabilidades, $\{\varphi_n(t;\theta)\}$, donde θ varía en cierto espacio paramétrico, y que el objetivo es estimar θ. Existen contextos en los cuales $\varphi_n(t;\theta)$ es directamente especificable, mientras que la función de densidad es inaccesible; más aún, es posible que algún método clásico de estimación tal como máxima verosimilitud, no sea aplicable. Supongamos que la dimensión de θ es p. Un método para estimar θ basado en la función de probabilidades empírica consiste en seleccionar p valores $t_1 < t_2 < \ldots < t_p$, establecer el sistema de p ecuaciones

$$\varphi_n(t_i) = \varphi(t_i;\theta), 1 \leq i \leq p,$$

y resolverlas como función de θ. Esta solución proporciona un estimador de θ similar en concepto al método de momentos, el cual propone ecuaciones con la igualación de momentos teóricos y empíricos (estimados). El método basado en $\varphi_n(t)$ posee una teoría asintótica en forma cerrada ver [2]), la cual es auxiliar en la importante pregunta acerca de la selección de los valores

de t en los que debe basarse el método con el fin de obtener estimadores eficientes. Con estos métodos, se demuestra también que en algunos casos, el método puede hacerse tan eficiente como el método de máxima verosimilitud, invirtiendo menor esfuerzo de cómputo.

Finalmente, puede mencionarse aquí otro problema de inferencia estadística que se presta a una solución basada en $\varphi_n(t)$: el llamado problema de *punto de cambio*. Este problema consiste en obtener una estimación del valor de k, basada en observaciones $X_1, \ldots, X_k$ que provienen de una distribución F, y $X_{k+1}, \ldots, X_n$ que provienen de una distribución G (consultar referencias contenidas en [4]). Esta situación se presenta si se tiene conocimiento de que en algún punto del tiempo de muestreo, haya sucedido algún fenómeno causante de un cambio en la distribución de las respuestas observadas; en fenómenos económicos y biológicos son comunes estas situaciones.

3 Función característica empírica.

El siguiente es un ejemplo específico de cómo puede utilizarse la función $C_n(t)$ para conseguir una solución no convencional a un problema de estimación. Una propiedad que caracteriza a una variable con distribución simétrica, es que su función característica es real. Este hecho puede explotarse para resolver el siguiente problema: Dadas observaciones $X_1, \ldots, X_n$, estimar el valor de α tal que la distribución de X^α sea de la forma $\mu + \epsilon$, donde ϵ posee una distribución *simétrica*—no necesariamente normal.

La motivación para el planteamiento de este problema radica en que este sencillo modelo describe adecuadamente distribuciones para X que son sesgadas, como las que ocurren con mucha frecuencia en fenómenos químicos o económicos; de hecho, los ejemplos presentados involucran mediciones de concentraciones de sustancias contaminantes en aguas de desecho industrial. Con estos datos, se observa que aunque la distribución de las mediciones originales no siguen una distribución simétrica, una potencia es capaz de inducir una distribución simétrica. Con ello, el sesgo en la distribución de las mediciones y los datos aparentemente atípicos presentes—se explican con un mecanismo relativamente sencillo. En caso de que la transformación no sea exitosa en el sentido de lograr una simetría aparente, entonces sería necesario recurrir a alguna distribución más compleja, posiblemente requiriendo un mayor número de parámetros.

El procedimiento (ver [5]) consiste en evaluar la función característica empírica en residuos construidos en distintos valores de α, y examinando la parte imaginaria de la misma para determinar si puede considerarse que ésta es cero en todas partes. Más precisamente: para cada valor candidato del parámetro α, consideremos los n valores de los *residuos* $r_j(\alpha) = X_j^\alpha - \hat{\mu}(\alpha)$, donde $\hat{\mu}(\alpha) = (1/n)\sum_{j=1}^n X_j^\alpha$. Considerando la función característica empírica basada en los residuos,

$$C_n(t;\alpha) = (1/n)\sum_{j=1}^{n} \exp(itr_j(\alpha)),$$

se elige α tal que $\int Im^2(C_n(t;\alpha))d\iota$ sea mínimo, produciendo así un estimador de la transformación a simetría. Dicho estimador posee propiedades atractivas que lo hacen competir favorablemente contra otras alternativas.

4 Bibliografía.

[1] Castro, I. (1993) "Algunos Aspectos Estadísticos de los Episodios de Niveles Altos de Ozono en la Ciudad de México", Tesis Profesional, Escuela de Ciencias, Universidad de las Américas–Puebla. (ejemplos adicionales de empleo de funciones generatrices empíricas)

[2] Dowling, M. and Nakamura, M. (1994) "Estimating Parameters in Discrete Distributions via the Empirical Probability Generating Function", *Comunicaciones Técnicas del CIMAT, No.* I–94-08. (descripción de la teoría asintótica asociada al procedimiento para estimar parámetros utilizando la función generatriz de probabilidades empírica).

[3] Feuerverger, A. and McDunnough, P. (1981) "On the efficiency of Empirical Characteristic Function Procedures", *Journal of the Royal Statistical Society*, B, **43**, no. 1, 20-27. (algunas propiedades teóricas sobre funciones características empíricas)

[4] Nakamura, M. and Pérez-Abreu, V. (1993) "Empirical probability generating function: An overview", *Insurance: Mathematics and Economics* **12**, 287-295. (contiene lista extensa de referencias adicionales)

[5] Pérez, F. (1993) "Estimación de transformaciones mediante minimización de criterios de simetría", Tesis de Maestría, Facultad de Matemáticas, Universidad de Guanajuato. (descripción detallada de un procedimiento para estimar una transformación)

CONTRASTE BAYESIANO DE HIPOTESIS PARAMETRICAS

Raúl Rueda
Departamento de Probabilidad y Estadística
IIMAS. UNAM

INTRODUCCION

Supóngase que existe

$$\mathcal{F} = \{p(X|\theta,\omega) : \theta \in \Theta, \omega \in \Omega\}$$

una familia paramétrica de distribuciones.

Sea $Z = \{X_1, X_2, \ldots, X_n\}$ una muestra aleatoria de $\mathcal{F}$ y sean

$$\mathcal{F}_0 = \{p(X|\theta,\omega) : \theta \in \Theta_0, \omega \in \Omega\}$$

$$\mathcal{F}_1 = \{p(X|\theta,\omega) : \theta \in \Theta_1, \omega \in \Omega\}$$

con Θ_0 y Θ_1 ajenos.

Se desea contrastar las siguientes hipótesis

$$H_0 : p(X|\theta,\omega) \in \mathcal{F}_0 \text{ vs. } H_1 : p(X|\theta,\omega) \in \mathcal{F}_1$$

o en forma equivalente

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

Existen diferentes soluciones a este problema. Estas soluciones dependen fundamentalmente de dos consideraciones: el enfoque estadístico y las dimensiones de Θ_0 y Θ_1. La solución que aquí se propone está dentro del marco formal de la Teoría de la Decisión.

SOLUCION GENERAL

Sea $D = \{d_0, d_1\}$ donde d_0 es elegir H_0 y d_1 elegir H_1. Si $d_0 < d_1$ significa que $p(X|\theta,\omega)$ puede aproximarse más adecuadamente por un elemento de $\mathcal{F}_0$ que de $\mathcal{F}_1$.

Dados los axiomas de coherencia, habrá que especificar:
p una distribución inicial sobre $\Theta \times \Omega$ y u una función de utilidad sobre $D \times \Theta \times \Omega$
y elegir d^* que maximice la utilidad esperada final

$$\int\int u(d;\theta,\omega)p(\theta,\omega|Z)d\theta d\omega$$

con $p(\theta,\omega|Z) \propto p(\theta,\omega)p(Z|\theta,\omega)$.

Así, $d_1 > d_0$ si

$$\int\int u(d_1,\theta,\omega)p(\theta,\omega|Z)d\theta d\omega > \int\int u(d_0,\theta,\omega)p(\theta,\omega|Z)d\theta d\omega$$

SOLUCION PARTICULAR

Puesto que elegir la acción d_i significa que la verdadera distribución de X es aproximada por un elemento de $\mathcal{F}_i$, es natural proponer como parte de la función de utilidad a una medida de la discrepancia entre los dos modelos: el propuesto por H_i y el verdadero.

Sea $\delta(\theta,\omega;\theta_i)$ una medida de la discrepancia entre $p(X|\theta,\omega)$ y $p(X|\theta_i,\omega)$ con $\theta \in \Theta$. Intuitivamente, si $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$ y $\delta(\theta,\omega;\theta_1) < \delta(\theta,\omega;\theta_0)$ sería razonable elegir a d_1.

Defínase a $u(d_i;\theta,\omega)$ como

$$u(d_i;\theta,\omega) = -A\delta(\theta,\omega;\theta_i) + B_i \text{ con } A \in \mathbf{R}^+ \text{ y } B_i \in \mathbf{R} \ (i = 0,1)$$

Si Θ_i $(i = 0,1)$ contienen sólo un elemento, la maximización de la utilidad esperada lleva a rechazar H_0 si y sólo si

$$\mathrm{E}_{\theta,\omega|Z}\Big\{\delta(\theta,\omega;\theta_1) - \delta(\theta,\omega;\theta_0)\Big\} < \frac{B_1 - B_0}{A}.$$

Si Θ_i $(i = 0,1)$ contienen a más de un elemento, el procedimiento usual es estimar θ_i en el conjunto Θ_i definido por la hipótesis H_i, y usar este estimador, por ejemplo θ_i, en lugar de θ_i.

Para este nuevo problema de decisión, se propone como función de utilidad a

$$u(\theta,\omega;\tilde{\theta}) = -\delta(\theta,\omega;\tilde{\theta})$$

por lo que el estimador de θ en Θ_i, $\tilde{\theta}_i^*$, es tal que

$$u(\tilde{\theta}_i^*) = \inf\Big\{\mathrm{E}_{\theta,\omega|Z}(\delta(\theta,\omega;\tilde{\theta}_i)) : \theta_i \in \Theta_i\Big\}$$

Así, una solución general para

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

es rechazar H_0 si y sólo si

$$\mathrm{E}_{\theta,\omega|Z}\Big\{\delta(\theta,\omega;\tilde{\theta}_1^*) - \delta(\theta,\omega;\tilde{\theta}_0^*)\Big\} < \frac{B_1 - B_0}{A}$$

con $\tilde{\theta}_i^*$ las correspondientes soluciones de

$$u(\tilde{\theta}_i^*) = \inf\Big\{\mathrm{E}_{\theta,\omega|Z}(\delta(\theta,\omega;\tilde{\theta}_i)) : \tilde{\theta}_i \in \Theta_i\Big\}$$

LA DIVERGENCIA LOGARITMICA COMO EJEMPLO DE UNA FUNCION DE DISCREPANCIA

Sea

$$\delta(\theta,\omega;\theta^*) = \int p(X|\theta,\omega)\log\frac{p(X|\theta,\omega)}{p(X|\theta^*,\omega)}dX, \quad \theta^* \in \Theta$$

Usando esta función de discrepancia, la solución queda como: rechazar H_0 si y sólo si

$$\mathrm{E}_{\theta,\omega|Z}\Big[\int p(X|\theta,\omega)\log\frac{p(X|\tilde{\theta}_0^*,\omega)}{p(X|\tilde{\theta}_1^*,\omega)}dX\Big] < \frac{B_1-B_0}{A}$$

con $\tilde{\theta}_i^*$ las correspondientes soluciones de

$$u(\tilde{\theta}_i^*) = \inf_{\theta_i^*\in\Theta_i}\int p(\theta,\omega|Z)\Big\{\int p(X|\theta,\omega)\log\frac{p(X|\theta,\omega)}{p(X|\tilde{\theta}^*,\omega)}dX\Big\}d\theta d\omega$$

FAMILIA EXPONENCIAL REGULAR

Sea X una variable aleatoria cuya densidad pertenece a

$$\mathcal{F} = \Big\{p(X|\theta) = a(\theta)b(X)\exp\theta t(X) : \theta\in\Theta\Big\}$$

con $\Theta = \{\theta\in\mathbf{R} : a(\theta)\in\mathbf{R}^+\}$ no vacío y abierto.

Supóngase que $p(\theta)\propto a^{n_0}(\theta)\exp(\theta t)$ y sea $Z = \{X_1, X_2, \ldots, X_n\}$ una muestra aleatoria de la distribución de X. Finalmente, supóngase las siguientes hipótesis

$$H_0 : \theta\in\Theta_0 \text{ vs. } H_1 : \theta\in\Theta_1$$

con Θ_0, Θ_1 subconjuntos ajenos de Θ.

En este caso

$$\mathrm{E}_{\theta|Z}(\delta(\theta;\tilde{\theta})) = -\frac{\partial}{\partial n_1}\log H(n_1,t_1) - \log a(\tilde{\theta}) + \frac{1}{n_1}\Big\{1+\frac{\partial}{\partial t_1}H(n_1,t_1) - \tilde{\theta}\Big\}$$

con $H(n,t)$ la constante de proporcionalidad en $p(\theta)$ $(p(\theta|Z)), n_1 = n_0 + n$ y $t_1 = t_0 + \sum_{i=1}^{n} t(X_i)$, y entonces se rechaza H_0 si y sólo si

$$\log\frac{a(\theta_1^*)}{a(\tilde{\theta}_0^*)} + (\tilde{\theta}_1^* - \tilde{\theta}_0^*)\frac{t_1}{n_1} > \frac{B_0-B_1}{A}$$

donde $\tilde{\theta}_i^*$ son las correspondientes soluciones de

$$-\frac{\partial}{\partial\tilde{\theta}_i}\log a(\tilde{\theta}_i) = \frac{t_1}{n_1},$$

de aquí es fácil demostrar que estas soluciones coinciden con las modas de las distribuciones finales restringidas a $\Theta_i, (i = 0, 1)$.

Si las mismas hipótesis fuerán consideradas, la solución por cociente de verosimilitudes generalizadas es

$$\log\frac{a(\hat{\theta}_1)}{a(\hat{\theta}_0)} + (\theta_1 - \hat{\theta}_0)\frac{1}{n}\sum_{i=1}^{n} t(X_i) > k$$

El estimador $\hat{\theta}_i$ es la solución de la ecuación

$$-\frac{\partial}{\partial\theta_i}\log a(\theta)i) = \frac{1}{n}\sum_{i=0}^{n} t(X_i)$$

así, si $n_0 = t_0 = 0$, *i.e.* usando una distribución límite, posiblemente no informativa, se tiene que

$$\tilde{\theta}_i^* = \hat{\theta}_i \quad (\imath = 0, 1)$$

y las soluciones "coinciden".

Un tipo de hipótesis ha merecido especial atención en la literatura,

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

en donde además se supone que

$$p(\theta = \theta_0) = p \text{ y } p(\theta \neq \theta_0) = (1-p)g(\theta) \text{ con } p \in (0,1) \text{ y } \int g(\theta)d\theta = 1$$

Dentro de la familia exponencial y usando la divergencia logarítmica, se tiene que

$$\delta(\theta,\tilde{\theta}) = \log\frac{a(\theta)}{a(\tilde{\theta})} + (\theta - \tilde{\theta})\mu(\theta) \text{ con } \mu(\theta) = \mathrm{E}_{X|\theta}(t(X))$$

lo que implica que $H_0 : \theta = \theta_0$ será rechazada si

$$\log\frac{a(\tilde{\theta}_0^*)}{a(\tilde{\theta}_1^*)} + (\tilde{\theta}_0^* - \tilde{\theta}_1^*)\mathrm{E}_{\theta|Z}(\mu(\theta)) < \frac{B_1 - B_0}{A}$$

con $\tilde{\theta}_i^* \in \Theta_i$ tal que $\mu(\tilde{\theta}_i^*) = \mathrm{E}_{\theta|Z}(\mu(\theta))$ para cualquier distribución inicial $p(\theta)$.

Supóngase que

$$p(\theta) = \sum_{i=1}^{k} \Pi_i g(\theta) \text{ con } \sum_{i=1}^{k} \Pi_i = 1, \Pi_i \geq 0 \text{ y } \int g_i(\theta)d\theta = 1 \ \forall\, i \in J_k$$

entonces

$$p(\theta|Z) = \sum_{i=0}^{\tilde{n}} \Pi_i' g_i(\theta|Z)$$

con

$$\Pi_i' = \frac{\Pi_i g_i(Z)}{\sum_{i=1}^{k} \Pi_i g_i(Z)}, \; g_i(\theta|Z) \propto g_i(\theta)p(Z|\theta) \text{ y } g_i(Z) = \int g_i(\theta)p(Z|\theta)d\theta \ \forall \imath \in J_k$$

y ademas

$$\mathrm{E}_{\theta|Z}(\mu(\theta)) = \sum_{i=0}^{k} \Pi_i' \int \mu(\theta) g_i(\theta|Z) d\theta$$

En resumen, usando la divergencia logarítmica y dentro de la familia exponencial, el procedimiento es

a. Encontrar $\mathrm{E}(\mu(\theta))$
b. Resolver la ecuación $\mu(\theta^*) = \mathrm{E}_{\theta|Z}(\mu(\theta))$
c. Decidir rechazar H_0 si $\log \frac{a(\tilde{\theta}_0^*)}{a(\tilde{\theta}_1^*)} + (\tilde{\theta}_0^* - \tilde{\theta}_1^*)\mathrm{E}_{\theta|Z}(\mu(\theta)) < \frac{B_1 - B_0}{A}$.

CONCLUSIONES

El procedimiento presentado tiene la característica principal de uniformizar el contraste bayesiano de hipótesis paramétricas y bajo ciertas condiciones, reproduce la solución clásica. Posiblemente la única dificultad que presenta es la elección de las constantes A, B_0 y B_1 que aparecen en la función de utilidad.

AGRADECIMIENTOS

El autor agradece a los organizadores del Foro, la amable invitación a participar. También agradece la "ayuda involuntaria" de Eduardo Gutiérrez y Efraín Santos.

CONTRIBUCIONES

LIBRES

CARACTERIZACIÓN DE LAS PRUEBAS PARA CAUSAS ESPECIALES EN LAS CARTAS DE CONTROL p Y np

Osvaldo Camacho Castillo[1]

Humberto Gutiérrez Pulido[1]

Las ocho pruebas más usuales para detectar cambios especiales en las cartas de Shewhart se han derivado a partir del supuesto de normalidad e independencia de los datos generados por el proceso (Western Electric, 1958). Existen varios estudios sobre el desempeño de las ocho pruebas para cartas $\bar{X}$-R, (Schilling and Nelson,1976; Champ & Woodall, 1987, por ejemplo). Sin embargo, este no es el caso para las cartas de atributos. Las más de las fuentes bibliográficas se limitan a dar recomendaciones de índole general, por ejemplo:

Western Electric(1958) dice "En las más de las cartas donde los límites de control son razonablemente simétricos, es suficientemente seguro aplicar las pruebas estándar".

Nelson(1987) recomienda "Las pruebas 1, 5 y 6, pueden ser usadas en las cartas p, np, c y u. También la prueba 2, si las distribuciones son suficientemente simétricas. Use las tablas de la distribución Binomial o Poisson para verificar situaciones específicas".

Montgomery(1991) presenta las pruebas de manera general para las cartas de control de Shewhart, y no asume una posición explícita sobre cuáles se deben usar en la cartas p.

Besterfield(1990) señala "Un estado de control para una carta p es tratado de manera similar a lo descrito para cartas de control para variables", en donde se presentaron las pruebas estándar.

Como se puede apreciar, en las recomendaciones hay ambigüedades, omisiones y contradicciones, y lo que es peor, algunas están equivoca-

[1]Facultad de Ingeniería, Universidad de Guadalajara. Campus Tecnológico, Guadalajara.

das, como lo veremos más adelante. Lo anterior puede provocar que el usuarios de las cartas de control aplique indiscriminadamente las pruebas y eso lleve a declarar con frecuencia que el proceso estuvo fuera de control estadístico, cuando en realidad no ocurrió así.

En este trabajo se presenta los resultados de un estudio de la significancia de las pruebas estándar aplicadas a las cartas p y np.

METODOLOGIA

Para las pruebas se calculó la probabilidad de que mediante éstas se detecte una señal de falta de control cuando en realidad el proceso no ha cambiado, es decir, se calculó el valor de la α para distintos valores de los parámetros (n,p) de una distribución binomial. Para las pruebas 1 a 4, 7 y 8, el calculo se hizo de manera exacta, mientras que para las pruebas 6 y 7, se estimó la probabilidad mediante el método Monte Carlo. Para evaluar la magnitud de las α encontradas se tomó como punto de referencia las de cada prueba bajo el supuesto de normalidad.

RESULTADOS

A continuación se presentan los resultados encontrados en función de 2000 valores de np, distribuidos entre 0.1 y 50, $p<=0.5$.

PRUEBA 1. UN PUNTO FUERA DE LOS LÍMITES DE CONTROL. Bajo el supuesto de normalidad esta prueba tiene una significancia (sig.) de 0.00135 por cada lado de la carta.

Esta prueba aplicada al LADO INFERIOR de la carta p no tiene problemas, ya que en general se logran $\alpha<0.00135$. Se detectaron casos aislados que superan tal sig., pero son menores que 0.002.

En el LADO SUPERIOR de la carta p, la aplicación de la prueba 1 presenta α altas, aún con valores grandes de np se dan casos donde $\alpha>0.00135$, ver figura 1. Significancias menores que 0.005 se logran a

partir de np>10. La $\bar{\alpha}$=0.002839 y S=0.002508; el 95.4% tiene valores de α>0.00135; el 77.5% de α es menor que 0.003.

PRUEBA 2. DOS DE TRES PUNTOS EN LA ZONA A. Bajo el supuesto de normalidad, la significancia de esta prueba para un solo lado de la carta es de 0.001075.

En el LADO INFERIOR la prueba 2 trabaja con α ligeramente mayores que el caso normal; $\bar{\alpha}$=0.000911 y S=0.000414, y sólo por excepción se tiene α>0.0025. Con lo que esta prueba aplicada a la parte inferior no generará demasiadas falsas alarmas.

En el LADO SUPERIOR la prueba 2 es muy poco segura ya que tiene en general significancias altas, sobre todo en valores de np menores que 6. Se obtuvo una $\bar{\alpha}$=0.002005 y S=0.001613; sólo el 8% de las combinaciones de np tuvo una α<0.00107, el 63.6% son menores que 0.002.

PRUEBA 3. CUATRO DE CINCO PUNTOS EN LA ZONA B Ó MÁS ALLA, SIN SALIRSE DEL LÍMITE DE CONTROL. Bajo el supuesto de normalidad, la significancia de esta prueba para un solo lado de la carta es de 0.0027.

En el LADO INFERIOR la prueba 3 genera α>0.0027. La $\bar{\alpha}$ fue de 0.003076, S=0.003184 y max=0.053. El 47% de los casos tiene una α mayor que 0.0027, el 8% rebalsa la sig. de 0.005.

En el LADO SUPERIOR la prueba 3 genera α moderadamente más altas que el caso normal, sobre todo en valores de np>3. Se obtuvo una $\bar{\alpha}$=0.002740 y S=0.001706. El 57% de los casos se tiene que α<0.0027.

PRUEBA 4. OCHO PUNTOS CONSECUTIVOS DE UN SOLO LADO DE LA LÍNEA CENTRAL, SIN SALIRSE DE LOS LÍMITES DE CONTROL. Bajo el supuesto de normalidad, la α de esta prueba para un solo lado de la carta es de 0.003822.

En el LADO INFERIOR la prueba 4 tiene problemas serios con np<1, en este caso se tiene $\bar{\alpha}$=0.4635. Para np>=1 las significancias siguen

siendo grandes, sobre todo para valores de np<10, ver figura 2 . La $\bar{\alpha}$ de 1978 casos de np>=1 fue de 0.005145 con S=0.004241.

En el LADO SUPERIOR la prueba 4 trabaja con significancias más altas que la normal, sobre todo con np<10. Se tiene $\bar{\alpha}$=0.003104, S=0.00168 y max=0.01808; el 33% de los casos tiene α>0.003822.

PRUEBA 5. SEIS PUNTOS CONSECUTIVOS EN AUMENTO O DISMINUCIÓN. El estudio Monte Carlo muestra que esta prueba no tiene problemas de exceso de falsas alarmas, ya que $\bar{\alpha}$=0.000059, S=0.000112 y max=0.0007.

PRUEBA 6. CATORCE PUNTOS CONSECUTIVOS ALTERNANDO ENTRE ALTOS Y BAJOS. Las 10,000 simulaciones para cada uno de los 2000 valores de np, muestran que esta prueba no tiene problemas de exceso de falsas alarmas, ya que $\bar{\alpha}$=0.001085, S=0.001001 y max=0.0052.

PRUEBA 7. OCHO PUNTOS CONSECUTIVOS A AMBOS LADOS DE LA LÍNEA CENTRAL CON NINGUNO EN LA ZONA C. Bajo el supuesto de normalidad, la significancia de esta prueba es de 0.000103.

Se tienen problemas serios en np=1, ya que $\bar{\alpha}$=0.01. En el resto de valores de np no hay problemas. Se obtuvo $\bar{\alpha}$=0.000162, S=0.000749.

PRUEBA 8. QUINCE PUNTOS CONSECUTIVOS EN LA ZONA C, ARRIBA O ABAJO DE LA LÍNEA CENTRAL. Bajo el supuesto de normalidad, la significancia de esta prueba es de 0.003254.

Con mucha frecuencia se tienen significancias más altas que la normal. Presenta problemas muy serios de falsas alarmas con np<1. En general, se obtuvo $\bar{\alpha}$=0.00677, S=0.024098 y max=0.355; el 20% tiene α menores que 0.002; el 50.8% tienen sig. mayores que 0.0032. Así, en general esta prueba no es confiable salvo para algunos valores de np.

CONCLUSIONES

Es riesgoso aplicar las ocho pruebas a las cartas p, ya que para ciertas combinaciones de np, se tendrán falsas alarmas frecuentemente. De la descripción del desempeño de cada prueba que se presentó anteriormente, se puede ver que los mayores riesgos se presentan cuando se aplica la prueba 1 y 2 al lado superior, con ciertas combinaciones de np<10, y cuando se aplica la prueba 3, 4 y 8 para todo valor de np (en combinaciones especificas). En los casos descritos antes a medida que np es más pequeños el riesgo se incrementa.

De lo anterior se concluye que la aplicación indiscriminada de las pruebas estandar a la carta p (o np), producira una mayor cantidad de falsas alarmas que en las cartas para procesos con distribución normal. Esto está en contradicción con las recomendaciones de Besterfield(1990) y Western Electric(1958). Si bien la afirmación de Nelson(1987) es correcta para las pruebas 5 y 6, no lo es para la prueba 1.

REFERENCIAS

Besterfield, D.H. (1990). Quality Control, 3e. Prentice-Hall, Englewood, New Jersey.

Champ, C.W. & W.H. Woodall (1987). "Exact results for shewhart control charts with supplementary runs rules", Technometrics, 26,4.

Gutiérrez Pulido, H. (1992). Control Total de Calidad. Edug, Guad.

Nelson, L. S. (1984). "The Shewhart control chart-tests for special causes", Journal of Quality Technology, 16,4, 237-39.

Schilling, E. G. and P. R. Nelson (1976). "The efect of nonnormality on the control limits of *X* charts", J. of Quality Technology, 8.

Western Electric (1958). Statistical Quality Control Handbook. AT&T, Chicago.

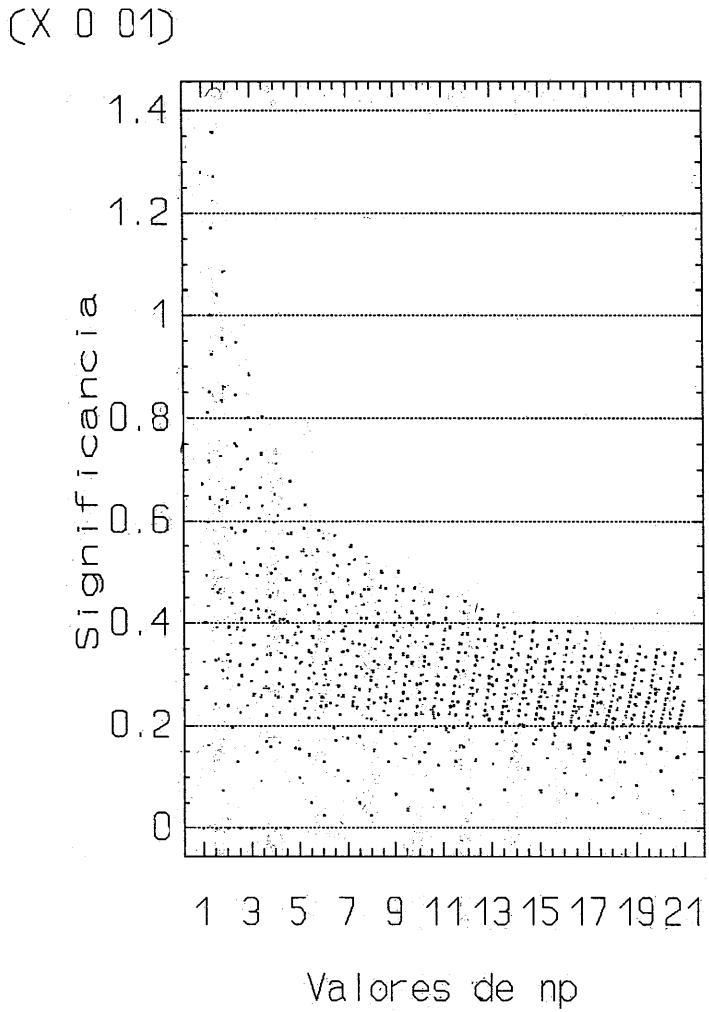


Fig. 1. Prueba 1, lado superior, p<0.5.

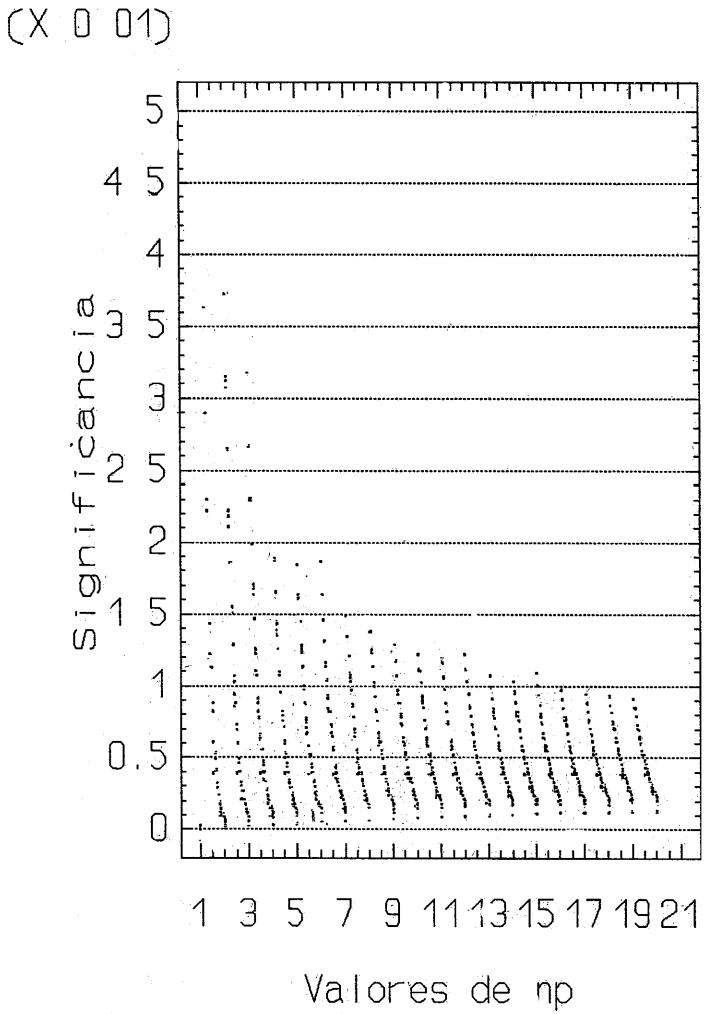


Fig. 2. Prueba 4, lado inferior, p<0.5.

Pruebas no paramétricas de homogeneidad para K muestras multivariadas[1]

Mario Cortina Borja
Departamento de Estadística y Actuaría
Instituto Tecnológico Autónomo de México
Río Hondo 1, San Angel, México 01000 D.F.
cortina@gauss.rhon.itam.mx

1 Introducción

Las dificultades encontradas al extender el concepto de orden para espacios con dimensión mayor que uno son la causa de la falta de generalizaciones de pruebas no paramétricas a problemas multivariados.
Una alternativa a una lista ordenada de valores muestrales para el caso multivariado consiste en asociar cada observación con otros individuos que se consideren cercanos a ella. Una forma natural de establecer cercanía entre observaciones multivariadas es el uso de gráficas basadas en matrices de disimilaridads.
La idea central es análoga al condicionamiento en las estadísticas de orden en el que se basan las pruebas no paramétricas univariadas más conocidas. En el caso multivariado se condiciona en las aristas que definen a una gráfica generada a partir de una matriz de disimilaridads. Las estadísticas de prueba estarán basadas en el número de aristas definidas por observaciones provenientes de muestras diferentes.
En este trabajo presentamos algunas generalizaciones de pruebas no paramétricas para probar la hipótesis de homogeneidad en K muestras multivariadas contra alternativas generales. Supondremos que hay $n = n_1 + n_2 + \cdots + n_K$ observaciones.
Este planteamiento general del problema puede utilizarse para establecer si varias muestras son diferentes con énfasis especialmente en los extremos de cada muestra. En el caso multivariado resulta díficil dar caracterizaciones precisas de alternativas específicas. Esto es lo contrario de lo que sucede con datos univariados donde es natural proponer alternativas referentes a, por ejemplo, el orden de las localizaciones de las K poblaciones analizadas.
En la siguiente sección presentamos algunas gráficas que pueden utilizarse para representar relaciones de cercanía en datos multivariados. Posteriormente explicamos cómo se obtienen las estadísticas de prueba y sus distribuciones. Finalmente mencionamos muy brevemente algunos resultados referentes a la potencia de pruebas de esta clase.

2 Gráficas basadas en matrices de disimilaridads

Todas las gráficas mencionadas en esta sección tienen por nodos a las n observaciones; sus aristas están ponderadas por alguna medida de disimilaridad d. Por facilidad supondremos que $d(x_i, x_j) \neq d(x_k, x_l)$ para toda i, j, k, l. Este supuesto es equivalente a suponer que no hay empates en datos univariados. Cabe mencionar que es posible modificar las estadísticas de prueba para considerar el caso en que posiblemente haya empates en las disimilaridades.

[1] Resumen de la ponencia presentada en el VIII Foro Nacional de Estadística, Aguascalientes, 1993

Para una discusión más completa de los algoritmos necesarios para calcular estas gráficas véase el trabajo de Cortina Borja (1992).

2.1 Gráficas de vecinos más cercanos

El n-*VMC* del punto x_i es el punto x_j tal que $d(x_i, x_k) < d(x_i, x_j)$ para exactamente $n-1$ valores de k, con $1 \leq k \leq n$ y $k \neq i, j$. La *n-gráfica de VMC* se obtiene ligando los puntos que son *VMC* de orden m, con . $1 \leq m \leq n$.

2.2 Arboles ortogonales de longitud mínima

Un *árbol* es una gráfica conexa y sin ciclos. Un *árbol de longitud mínima (ALM)* es tal que la suma de las ponderaciones de las aristas es mínima entre todos los árboles que es posible definir para los n nodos. Si hay n observaciones entonces un *ALM* tiene $n-1$ aristas que representan parejas de observaciones que pueden considerarse cercanas entre ellas.
Dos gráficas son *ortogonales* si tienen el mismo conjunto de nodos pero la intersección de sus conjuntos de aristas es vacía.
Entonces el 2-*ALM* es la unión del 1-*ALM* y el *ALM* obtenido minimizando su longitud total si se eliminan las aristas incluídas en el 1-*ALM*. En general, un n-*ALM* es la unión de los primeros $(n-1)$-*ALM* y el *ALM* obtenido sin incluir ninguna arista perteneciente a *ALM*s obtenidos previamente.

2.3 Gráficas de vecindad relativa

Una posibilidad para considerar una gráfica de vecindad relativa *(GVR)* es la siguiente, debida a Toussaint (1980):

> x_i y x_j definen una arista de la *GVR* sí y sólo sí
>
> $$d(x_i, x_j) \leq \max_{k \neq i,j} \max \left[d(x_i, x_k), d(x_j, x_k) \right]$$

Intuitivamente esto significa que dos observaciones están ligadas en la *GVR* si están al menos tan cerca entre ellas como lo están con respecto a cualquier otra observación.
En espacios euclideanos, la siguiente definición es equivalente:

> x_i, x_j forman una arista en la *GVR* sí y sólo sí la intersección de las hiperesferas abiertas con radios $d(x_i, x_j)$ centradas en x_i y x_j no contiene a ninguna otra observación.

Lefkovitch (1985) propuso una generalización para construir *GVR* ortogonales de órdenes superiores que puede expresarse como

> Las aristas de n-*GVR* con $n > 1$ ligan puntos que no estaban ligados en *GVR* previas y que tienen al menos un vecino relativo tienen al menos un vecino relativo de orden menor común.

2.4 Gráficas de Gabriel

Esta gráfica fue propuesta primeramente por Gabriel and Sokal (1969) para definir conexidad en un conjunto de regiones geográficas. Su definición es como sigue:

> x_i and x_j definen una arista en GG sí y sólo sí la hiperesfera abierta con diámetro $d(x_i, x_j)$ y centrada en el punto medio del segmento que une a x_i con x_j no contiene a ninguna observación.

Esto es equivalente a decir que x_i y x_j definen una arista en GG sí y sólo sí

$$d^2(x_i, x_j) \leq d^2(x_i, x_k) + d^2(x_j, x_k)$$

para toda $k \neq i, j$.
Es posible usar la generalización de Lefkovitch para obtener una sucesión de GGs ortogonales.

3 Generalizaciones de la prueba de rachas multivariadas para K muestras

3.1 Coeficientes de correlación generalizados

Los métodos estadísticos basados en rangos constituyen procedimientos alternativos al enfoque paramétrico clásico para probar la hipótesis nula de homogeneidad para K poblaciones:

$$H_0 : F_{X_1} = F_{X_2} = \cdots = F_{X_K}.$$

En este trabajo consideraremos hipótesis alternativas generales. A continuación presentamos la generalización de la estadística de rachas multivariadas de Friedman–Rafsky (1979) para K muestras. La teoría que utilizamos es la de los coeficientes de correlación generalizados.
Considere una muestra $(X_i, Y_i), i = 1, \cdots, n$ de pares ordenados y sean a_{ij}, b_{ij} funciones para cada par (i, j) de observaciones X y Y respectivamente. Entonces, sin incluir una forma de estandarización específica, un coeficiente de correlación generalizado *(CCG)* es de la forma

$$\Gamma = \sum_{i}^{n} \sum_{j}^{n} a_{ij}\, b_{ij} \tag{1}$$

Si condicionamos sobre los valores observados de X y Y es posible probar la hipótesis nula de no correlación ordenando el valor observado de T respecto a la distribución de

$$\mathrm{T}(\pi) = \sum_{i}^{n} \sum_{j}^{n} a_{ij}\, b_{\pi(i)\,\pi(j)} \tag{2}$$

donde π es una permutación de los enteros $\{1,\cdots,n\}$. Este procedimiento es adecuado puesto que bajo la hipótesis nula de no correlación entre X y Y todas las permutaciones (las parejas (X,Y)) son igualmente probables. Esta distribución permutacional determina si el valor observado de Γ es significativo.
La ecuación 2 proporciona la forma de calcular la distribución permutacional nula de Γ. Daniels (1944) da condiciones generales para establecer la normalidad asintótica de una amplia clase de CCGs. La traducción de las condiciones de Daniels al contexto de teoría de gráficas impone algunas restricciones sobre la topología de las gráficas. Sin entrar en detalles técnicos, es posible afirmar que estas condiciones se refieren a que las gráficas deben ser asintóticamente densas, i.e. deben contener una proporción alta de las aristas de la gráfica completa. Para asegurar esto es suficiente ver que el grado de cada nodo crece linealmente con n. Es posible mostrar que aún si las gráficas no son asintóticamente densas se logra la normalidad asintótica bajo ciertas condiciones bastante generales. Los detalles técnicos están en Cortina (1992).

3.2 Pruebas de homogeneidad

En el contexto de pruebas de homogenidad, la falta de correlación se refiere a las relaciones entre cercanía en el espacio multivariado X y cercanía entre las identidades muestrales Y. Si ambas variables están positivamente correlacionadas entonces habrá evidencia de falta de homogeneidad entre las K poblaciones.
Sea $\mathcal{G}_X$ una gráfica construída a partir de una matriz de disimilaridad definida sobre las observaciones X en la muestra conjunta y sea $\mathcal{G}_Y = \cup_{j=1}^{K} \mathcal{K}_{n_j}$, donde $\mathcal{K}_{n_j}$ es la gráfica completa formada ligando todas las observaciones en la j-ésima muestra; i.e. sus aristas están definidas por nodos de la misma muestra. Esto nos permite escribir la estadística de prueba como un *CCG* entre puntos que definen aristas en $\mathcal{G}_X$ (que son puntos que deben estar cerca en el espacio de las observaciones) e identidad muestral. Para esto definimos a a_{ij} como 1 si los nodos i y j forman una arista en $\mathcal{G}_X$ y como 0 en otro caso; las funciones b_{ij} se definen del mismo modo para las aristas de $\mathcal{G}_Y$.
Supongamos que tenemos observaciones de K poblaciones con tamaños muestrales $n_1, n_2, \cdots, n_K$ con distribuciones $F_{X_1}, F_{X_2}, \cdots, F_{X_K}$. El tamaño de la muestra conjunta es $n = \sum_{j=1}^{K} n_j$.
Sea la v.a. Z_i definida como

$$Z_i = \begin{cases} 1 & \text{si } i \in \mathcal{G}_Y \\ 0 & \text{e.o.c.} \end{cases} \qquad 1 \leq i \leq e_X$$

siendo e_X el número de aristas de $\mathcal{G}_X$, de manera que el número de aristas en $\mathcal{G}_X \cap \mathcal{G}_Y$ es

$$\mathrm{T}_R = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}\, b_{ij} = \sum_{i=1}^{e_X} Z_i \tag{3}$$

Bajo H_0 no deberíamos observar una correlación alta entre los nodos que definen aristas en $\mathcal{G}_X$ y aquellos que lo hacen en $\mathcal{G}_Y$. Entonces, los valores de T_R que llevarían a rechazar H_0 serán relativamente grandes.
Cualquiera de las gráficas mencionadas en la sección 2 es adecuada para probar la hipótesis

de homogeneidad. Sin embargo la potencia de la prueba dependerá de la clase de gráfica seleccionada.

Para obtener el momento de orden r para T_R es necesario condicionar sobre todas las configuraciones que pueden formarse con r aristas diferentes; en general, el r-ésimo momento puede expresarse en términos de

$$\sum_{i_1<\cdots<i_r}^{e_X} \mathrm{E}\,(Z_{i_1} Z_{i_2} \cdots Z_{i_r}) = \sum_{m=1}^{g_r} q_m \Pr\,[\,\mathrm{E}\,(Z_{i_1} Z_{i_2} \cdots Z_{i_r}) = 1 \,|\, \mathcal{D}_m]$$

donde g_r es el número de subgráficas que es posible formar con r aristas diferentes y q_m es el número observado de la subgráfica de tipo m ($\mathcal{D}_m$) que consiste de r aristas diferentes en $\mathcal{G}_X$. Para el primer momento no hay más que una subgráfica con 1 arista. El segundo momento involucra dos subgráficas con dos aristas: una en la que estas comparten un nodo y otra en la que no lo hacen. Para los momentos de orden 3 y 4 los números de subgráficas necesarias son 5 y 11, respectivamente.

Las probabilidades

$$\Pr\,[Z_{i_1} Z_{i_2} \cdots Z_{i_r} = 1 \,|\, \mathcal{D}_m]$$

pueden expresarse como funciones de los tamaños de muestra $n_1, \cdots, n_K$. Los detalles se desarrollan en Cortina y Robinson (1993). Desgraciadamente, el cálculo de los coeficientes q_i necesarios para obtener el tercero y el cuarto momento de Γ_R puede ser computacionalmente muy costoso. Al contrario con lo que ocurre para los primeros dos momentos no hay forma alguna de obtener estos coeficientes con una función simple del conjunto de grados de los nodos en $\mathcal{G}_X$.

La única posibilidad consiste en utilizar enumeración directa para las distintas configuraciones observadas en $\mathcal{G}_X$. Esta labor puede efectuarse para gráficas poco densas, es decir, con relativamente pocas aristas o bien en gráficas que tienen un grádo máximo D_p^* que es una función que depende únicamente de la dimensión p del espacio de las observaciones y que está acotada independientemente de n, de tal suerte que $D_p^* \ll n$. En cualquier caso, el número de operaciones necesarias en el procedimiento de enumeración para obtener el r-ésimo momento es a lo más proporcional a la r-ésima potencia del grado máximo observado en $\mathcal{G}_X$. Como esta cantidad puede ser en muchos casos de casi el mismo orden de magnitud de, por ejemplo, $n/2$, la carga computacional puede ser demasiado pesada. Sin embargo, para tamaños de muestra moderados estos cálculos son factibles. Los detalles computacionales y los algoritmos necesarios pueden verse en Cortina (1992).

Una vez obtenidos los primeros cuatro momentos de Γ_R es posible aproximar su distribución nula utilizando curvas de Pearson. El algoritmo propuesto por Davis y Stephens (1983) es muy útil para obtener los percentils más comunes de la densidad de Pearson correspondiente.

Como se mencionó en la subsección anterior, es posible obtener una muestra de la distribución permutacional nula o establecer normalidad asintótica para Γ_R con el fin de construir pruebas de significancia. Ambos enfoques son computacionalmente simples. Sin embargo, no es fácil determinar un número de permutaciones muestreadas para que la distribución permutacional nula esté bien representada, particularmente en sus extremos. Lo mismo sucede con los tamaños de muestra n que garanticen la normalidad asintótica. El uso de curvas de Pearson es un camino adecuado para aproximar la distribución nula de Γ_R.

Finalmente, mencionaremos que las pruebas basadas en Γ_R tienen buena potencia aún para tamaños de muestra pequeños y datos con alta dimensionalidad. Varios resultados que apoyan lo anterior aparecen en Cortina y Robinson (s/f).
Ciertamente estas pruebas no paramétricas son una excelente alternativa para las pruebas clásicas multivariadas basadas en cocientes de verosimilitud en las que los supuestos pueden ser tan difícil de probar (o aún más) que la misma hipótesis de interés.

REFERENCIAS

CORTINA BORJA, M (1992) *Graph-Theoretic Multivariate Nonparametric Procedures.* PhD thesis, University of Bath.

CORTINA BORJA, M Y ROBINSON, T (1993) *Some generalizations of the Friedman–Rafsky tests* Cuadernos de trabajo. Departamento de Estadística y Actuaría, ITAM.

CORTINA BORJA, M Y ROBINSON, T (s/f) The power of some graph-theoretic multivariate nonparametric tests Enviado a publicación.

DANIELS, HE (1944) The relation between measures of correlation in the universe of sample permutations. *Biometrika* **33**, 120-135.

DAVIS, CS AND STEPHENS, MA (1983) Approximate percentage points using Pearson curves. *Applied Statistics* **32**, *322-324.*

FRIEDMAN, JH AND RAFSKY, LC *(1979) Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests.* Annals of Statistics **7** 697-717.

GABRIEL, KR AND SOKAL, RR (1969). A new statistical approach to geographic variation analysis. *Systematic Zoology*, **18**, 259-278.

LEFKOVITCH, LP (1985) Further nonparametric tests for comparing dissimilarity matrices based on the relative neighborhood graph. *Mathematical Biosciences* **73**, 71-88.

TOUSSAINT, GT (1980) The relative neighbourhood graph of a finite planar set. *Pattern Recognition* **12**, 261-268.

Instituto Nacional de Estadística, Geografía e Informática

Tema:

Sistema para la Consulta de Información Censal (SCINCE) Versión 2.0

Expositor :

Lic. Mario Chavarria Espinosa
Lic. Víctor Esparza de Lira
José Luis Olarte Quiroz
Coordinación de Cartografía Censal.
Subdirección de Cartografía Automatizada.
Depto. de Desarrollo de Sistemas.
Héroe de Nacozari 2301 Sur.
Puerta 6 Mezzanine.
Tel. 16-67-92
Aguascalientes, Ags., México.

RESUMEN

En décadas anteriores, la información que resultaba de los censos se difundía por medio de tabulados impresos que sólo contenían identificadores y cifras. Este tipo de productos cubrio cierto espectro de necesidades; sin embargo, en ellos no resultaba sencillo reflejar una clara relación entre la información generada y el espacio geográfico que le dió origen. En este orden de ideas, es importante recordar que todo dato estadístico únicamente cobra sentido al referenciarse a un punto en el tiempo y en el espacio, a un cuándo y un dónde; el cuándo lo dan los calendarios, el dónde la cartografía. Para salvar parcialmente dicha problemática, los tabulados impresos eran complementados con cartas temáticas, las cuales apenas contemplaban los indicadores más importantes, sobre todo porque éstos eran trabajos totalmente manuales.

En el Instituto Nacional de Estadística, Geografía e Informática (INEGI), la implementación de nuevas tecnologías ha devenido en el desarrollo de nuevos productos, entre los cuales se encuentra un sistema para microcomputadoras que permite al usuario relacionar la Información Censal con su correspondiente Espacio Geográfico: el Sistema para la Consulta de Información Censal (SCINCE).

El Sistema para la Consulta de Información Censal es un producto cuyo objetivo esencial es ofrecer a los usuarios de la información censal un software que, de manera ágil e interactiva, permita obtener el mayor provecho posible de los resultados que arrojan los diversos censos, valiéndose para ello, de Información tanto Geográfica como Estadística, y de la mutua relación que existe entre ambas. Así, el SCINCE permite analizar la distribución y el comportamiento de la Información Estadística en el espacio al cual pertenece, ubicándolo como un instrumento generador de nueva información.

Concebido como un sistema interactivo, el SCINCE puede ser manejado por cualquier usuario de microcomputadoras, ya que para su operación no es necesario ser un especialista en informática, sino tan solo desear conocer la relación Gráfico-Estadística de la información.

A fin de alcanzar su objetivo, el SCINCE se diseñó para:

- Permitir la definición de la Unidad Geográfica que será sujeta a estudio así como del Indicador Censal que será analizado en la unidad elegida.

- Permitir visualizar la Unidad Geográfica definida y analizarla a través de acercamientos sucesivos, de recuperación de acercamientos previos y de desplazamientos; así como la localización gráfica de alguna Subunidad Geográfica por medio de su clave.

- Relacionar la información estadística con el espacio geográfico al cual pertenece y representar tal relación gráficamente, con la generación de un plano temático, que para su construcción es posible determinar de dos a siete intervalos, lo cual asociado a una escala cromática o de achurados, permite mostrar las variaciones en la distribución espacial de la información, siendo posible establecer libremente la dimensión de cada intervalo.

- Presentar, en gráficas de barras, la variación cuantitativa de la información, con la finalidad de analizar el comportamiento del Indicador Censal en estudio, dentro de la Unidad Geográfica seleccionada.

- Definir Indicadores Censales Compuestos mediante operaciones matemáticas entre los Indicadores Censales.

- Producir impresiones de los gráficos y tabulados por medio de:
  - Impresora de Matriz de puntos (configurada en modo EPSON).
  - Impresora Laser.
  - Impresora Paint Jet.
  - Periférico que utilice código HPGL.

- Producir, tanto en pantalla como en papel, reportes de totales de la información relativa a cada Unidad Geográfica, así como reportes del Indicador Censal elegido o del Indicador Compuesto construido.

- Importar Información Estadística propia del usuario.

- Permitir la operación matemática entre Indicadores Censales de archivos de información independientes.

-Exportar la Información Cartográfica original a formato DXF.

-Exportar la Información Estadística original a formato ASCII delimitado.

-Exportar la Información Estadística, resultado de la operación matemática entre los Indicadores Censales a formato ASCII delimitado.

ELEMENTOS UTILIZADOS POR EL SCINCE

Para su funcionamiento, el SCINCE se apoya en dos Elementos Fundamentales; el Cartográfico y el Estadístico.

El Elemento Cartográfico

Con respecto a este elemento, se manejan tres niveles de cobertura:

- Nacional :

  Por estado, donde cada estado es una Subunidad Geográfica.

  Por municipio, donde cada municipio es una Subunidad Geográfica.

- Estatal por municipio, donde cada municipio es una Subunidad Geográfica; y

- Localidad Urbana por Area Geoestadística Básica (AGEB); donde cada AGEB es una Subunidad Geográfica.

Estos niveles de cobertura tienen su expresión física en la formulación de Unidades Geográficas y Sub-Unidades Geográficas. Se entiende por Unidad Geográfica al conjunto de Subunidades Geográficas que tienen por objetivo la representación por zonas de los resultados producidos por los diversos proyectos censales.

El SCINCE representa el territorio nacional en tres niveles de Unidades Geográficas:

1. País
2. Estado
3. Localidad

Los límites de las Unidades Geográficas corresponden a los establecidos por la Cartografía Censal, y no necesariamente a los Político-Administrativos. Cada Subunidad Geográfica tiene identificadores socio-económicos que pueden ser fácilmente cuantificados y representados en planos y gráficas.

El Elemento Estadístico

El elemento estadístico, está conformado por los datos numéricos producidos por la actividad censal para cada nivel de cobertura y para cada Subunidad Geográfica, y son precísamente estos datos los que, por medio del SCINCE, pueden referirse a su espacio geográfico correspondiente.

INDICADORES CENSALES

Los indicadores utilizados en el sistema son 71, derivados del XI Censo General de Población y Vivienda, 1990, y son el contenido de las publicaciones de datos por AGEB Urbana. Su clave se compone de la letra "P" seguida de dos dígitos del 01 al 71, como puede observarse a continuación.

DESCRIPCION DEL INDICADOR	CLAVE
Población Total	P01
Pob. Femenina	P02
Pob. de 5 años y más	P03
Pob. de 6 años y más	P04
Pob. de 12 años y más	P05
Pob. de 15 años y más	P06
Pob. de 16 años y más	P07
Pob. de 18 años y más	P08
Pob. de 35 años y más	P09
Pob. de 65 años y más	P10
Pob. Nacida en la Entidad	P11
Pob. Nacida fuera de la Entidad	P12
Pob. de 5 años y más Residentes en la Entidad en 1985	P13
Pob. de 5 años y más Resid. fuera de la Entidad en 1985	P14
Pob. de 5 años y más Católica	P15
Pob. de 5 años y más No Católica	P16
Pob. de 6-14 años que Saben Leer y Escribir	P17
Pob. de 15 años y más Alfabeta	P18
Pob. de 6-14 años que Asisten a la Escuela	P19
Pob. de 15 años y más sin Instrucción	P20
Pob. de 15 años y más con Primaria Completa	P21
Pob. de 15 años y más con Instrucción Postprimaria	P22
Pob. de 15 años y más sin Instrucción Media Básica	P23
Pob. de 15 años y más con Secundaria Completa	P24
Pob. de 15 años y más con Educación Postmedia Básica	P25
Pob. de 18 años y más sin Educación Media Superior	P26
Pob. de 18 años y más con Instrucción Superior	P27
Pob. de 18 años y más sin Instrucción Superior	P28
Pob. de 12 años y más Soltera	P29
Pob. de 12 años y más Casada	P30
Pob. Femenina de 12 años y más	P31
Promedio de Hijos Nacidos Vivos	P32
Promedio de Hijos Sobrevivientes	P33
Pob. Económicamente Activa Ocupada	P34
Pob. Económicamente Activa Desocupada	P35
Pob. de 12 años y más Estudiante	P36
Pob. de 12 años y más Dedicada a Quehaceres del Hogar	P37
Pob. Ocupada en el Sector Secundario	P38
Pob. Ocupada en el Sector Terciario	P39
Pob. Ocupada como Empleado u Obrero	P40
Pob. Ocupada como Jornalero o Peón	P41
Pob. Trabajadora por Cuenta Propia	P42
Pob. Ocupada que Trabajó hasta 32 hrs. en la Semana	P43
Pob. Ocupada que Trabajó de 32-40 hrs. en la Semana	P44
Pob. Ocupada que Trabajó de 41-48 hrs. en la Semana	P45
Pob. Ocupada con menos de un S. M. Mens. de Ingreso	P46
Pob. Ocupada con más de 1 y hasta 2 S. M. Mens. de Ingr.	P47
Pob. Ocupada con más de 2 y hasta 5 S. M. Mens. de Ingr.	P48
Viviendas Particulares Habitadas	P49
Viv. Part. con Techo de Losa	P50
Viv. Part. con Techo de Lámina de Asbesto, Cartón o Met.	P51
Viv. Part. con Paredes de Tabique	P52
Viv. Part. con Paredes de Madera	P53
Viv. Part. con Paredes de Adobe	P54
Viv. Part. con Piso de Cemento	P55
Viv. Part. con Piso de Mosaico, Madera u otros Recubrim.	P56
Viv. Part. con 1 Cuarto	P57
Viv. Part. con 2-5 Cuartos	P58
Viv. Part. con 1 Dormitorio	P59
Viv. Part. con 2-4 Dormitorios	P60
Viv. Part. con Cocina Exclusiva	P61
Viv. Part. con Cocina No Exclusiva	P62
Viv. Part. que usa Gas para Cocinar	P63
Viv. Part. con Drenaje Conectado a la Calle	P64
Viv. Part. con Drenaje Conectado a Suelo o Fosa	P65
Viv. Part. que Disponen de Energía Eléctrica	P66
Viv. Part. con Agua Entubada a la Vivienda	P67
Viv. Part. con Agua Entubada en el Predio	P68
Viv. Part. con Agua en Llave Pública	P69
Viv. Part. Propias	P70
Viv. Part. Rentadas	P71

Producción de información estadística demográfica y social. Registros administrativos

Antonio Escobedo Aguirre

Dirección General de Estadística
Dirección de Estadísticas Demográficas y Sociales
Instituto Nacional de Estadística, Geografía e Informática

PRODUCCION DE INFORMACION ESTADISTICA DEMOGRAFICA Y SOCIAL

REGISTROS ADMINISTRATIVOS

Los registros administrativos constituyen, junto con los censos y las encuestas, las fuentes de información básicas del Sistema Nacional de Información Estadística; en ellos se asienta de manera continua información demográfica, económica y social.

En el marco de la Ley de Información Estadística y Geográfica (LIEG) se encuentran los principios y las normas jurídico-administrativas mediante las cuales las dependencias y entidades de la administración pública federal deben ejercer las funciones que les corresponde, como partes integrantes de los Servicios Nacionales de Estadística y de Información Geográfica.

El Servicio Nacional de Estadística comprende, entre otros aspectos: la generación de estadísticas que observen hechos económicos, demográficos y sociales de interés nacional; las estadísticas permanentes, básicas o derivadas, las cuentas nacionales; indicadores que elaboren las dependencias, instituciones públicas y privadas, los poderes y servicios estatales; así como, la publicación de los resultados de las actividades que corresponden al Servicio Nacional de Estadística como tal.

Corresponde a la Dirección de Estadísticas Demográficas y Sociales (DEDS) por encargo de la Dirección General de Estadística, la planeación, programación, supervisión y control de las actividades relacionadas con la generación e integración de la información estadística de nueve temas básicos: Nacimientos, Matrimonios, Divorcios, Defunciones generales y fetales, Salud, Educación, Relaciones Laborales, Seguridad y Orden Público y Cultura (bibliotecas, asistencia a cines, museos y otros espectáculos públicos).

En esta oportunidad, se presenta de manera sucinta el proceso seguido para la generación de información demográfica y social proveniente de registros administrativos; cuyo objetivo general, consiste en generar y difundir información proveniente de registros administrativos y civiles sobre fenómenos socio-demográficos.

Dicha información es recolectada, integrada y difundida mediante procesos coordinados por la DEDS, a través de las diez Direcciones Regionales del Instituto Nacional de Estadística Geografía e Informática. Estos procesos están sustentados por una serie de manuales e instructivos distribuidos a las Direcciones Regionales con la finalidad de orientar todas y cada una de las fases de cada proceso. En la DEDS, también se diseñan los criterios de validación de la información y los planes básicos de tabulaciones para consulta y difusión.

Estadísticas Vitales

La información demográfica constituye un insumo básico en la definición de políticas de carácter social y para el conocimiento de la evolución del país, así como, un valioso conjunto de elementos de análisis del entorno social con los cuales la población en general se apoye para formar su propia opinión, opinen y participen decididamente en la toma de decisiones.

En nuestro país, el sistema de estadísticas vitales ha experimentado un cambio radical pues, de haber permanecido prácticamente intacto durante casi un siglo, en la década de 1980 se reestructura permitiendo acciones interinstitucionales necesarias para su consolidación.

Actualmente, el sistema funciona en forma desconcentrada a través de diez oficinas regionales cuyo ámbito de competencia lo componen 3 ó 4 entidades federativas. En ellas se realizan a partir de 1985, las actividades siguientes: recopilan la información en las entidades que les corresponden; mantienen el contacto con las fuentes informantes; efectúan el tratamiento manual y electrónico de la información, integran y divulgan los datos a nivel estatal.

Las Oficialías del Registro Civil proporcionan a las oficinas estatales de estadísticas continuas, actas de nacimiento, de matrimonio y de defunción, certificados de defunción y de muerte fetal y cuadernillos de divorcios administrativos. Las Agencias del Ministerio Público entregan cuadernillos de defunciones accidentales y violentas, en tanto que los Juzgados de lo Familiar, de lo Civil y los Mixtos, proveen los cuadernillos de divorcios judiciales.

En el caso de las defunciones, se lotifican los formatos, previa selección de las actas, certificados y cuadernillos. Entre ellos, el certificado de defunción es el documento fuente en primera instancia, de no contar con él, se toma el acta y, en el caso extremo de no recibir certificado ni acta en una muerte accidental o violenta, se emplean los cuadernillos. También se realiza la crítica codificación de las variables geográficas y de las características sociodemográficas de los involucrados en el evento registrado. Además, se codifican las causas de muerte tanto en las defunciones generales como fetales.

Cabe destacar que, en el caso de las defunciones, se realizan actividades específicas a fin de evitar el doble conteo de casos y para asegurar la calidad de la información sobre la causa básica de la defunción.

Con el propósito de detectar y atacar oportunamente brotes epidémicos, la Secretaría de Salud (SSA) ha definido 20 enfermedades sujetas a vigilancia epidemiológica. Cuando un médico o persona autorizada asienta entre las causas de muerte una de esas enfermedades, se debe realizar una investigación epidemiológica en el lugar donde acaeció la defunción, así como, en el de residencia habitual del fallecido y determinar la veracidad del diagnóstico.

Así, las Direcciones Regionales del INEGI reportan a las delegaciones estatales de la SSA los certificados correspondientes a la situación descrita y reciben a cambio, la ratificación o la rectificación del diagnóstico médico original. Condición sin la cual, el Instituto no publicaría la información sobre mortalidad.

A la fecha, la información sobre los hechos vitales se integra y difunde en un lapso menor a 12 meses de ceral año estadístico, tal es el caso de los Cuadernos de Población números 4 y 5 con datos correspondientes a 1991 y 1992 respectivamente. Estos avances en cuanto a la oportunidad de la información presentada descansa de manera determinante en la desconcentración efectuada hacia las Direcciones Regionales.

Los hechos vitales son caracterizados mediante una serie de variables sociodemográficas como son la edad, el sexo, la escolaridad, el lugar de residencia habitual, la ocupación, etc.; con las cuales se pueda formular hipótesis explicativas de los fenómenos en estudio. Cabe destacar que

de ninguna manera pueden considerarse como opción o alternativa a las mediciones de fenómenos tales como el empleo o la migración, para los cuales se siguen metodologías especificas aplicadas por personal especializado.

Las estadísticas vitales permiten en conjunto, conocer aspectos de la dinámica de la población, como son su crecimiento natural (proporcionado por la diferencia de los nacimientos y las defunciones) y el proceso de formación-disolución de hogares (a través de los matrimonios y divorcios registrado), en los ámbitos nacional, estatal y municipal.

A partir de los nacimientos registrados se obtienen indicadores del nivel de la fecundidad y algunas características de las madres, como son su edad, escolaridad, participación en actividades económicas y su lugar de residencia.

La información de los matrimonios permite conocer la edad de los contrayentes, además de su lugar de residencia, escolaridad y participación en actividades económicas. También con ello se puede saber la edad promedio al matrimonio de los hombres y de las mujeres, la cual es considerada como determinante de la fecundidad.

Los divorcios, por su parte, ofrecen datos sobre las causas de las rupturas conyugales, la duración del matrimonio y sobre características de los divorciados, entre los que se anotan su edad, escolaridad y participación en actividades económicas.

Las estadísticas de defunciones muestran en su vertiente demográfica, la estructura por edad y sexo de la mortalidad y la residencia habitual de los fallecidos. En su otra vertiente, permiten un acercamiento al estado de salud de la población a través del estudio de la causa básica de muerte, el lugar de ocurrencia del deceso y si el fallecido recibió atención médica durante su último padecimiento o accidente.

El INEGI ha divulgado tradicionalmente las estadísticas vitales en publicaciones de carácter general, como son el Anuario Estadístico de los Estados Unidos Mexicanos y la Agenda Estadística, aunque también las difunde en publicaciones especializadas, como la serie de los Cuadernos de Población, que a la fecha cuenta con cinco números abarcando información definitiva desde 1970 hasta 1992.

Las **ESTADISTICAS DE ASISTENCIA Y SERVICIOS MEDICOS** son, por una parte, el complemento natural de la información sobre mortalidad pues los datos sobre morbilidad hospitalaria redondean el panorama del estado general de salud de la población y por otra, dan cuenta de los recursos materiales y humanos empleados por las instituciones hospitalarias.

Esta información es generada por las instituciones públicas que prestan servicios de atención a la salud de la población; se integran y se difunden mediante los mecanismos establecidos en el seno del Grupo Interinstitucional de Información y Evaluación de Salud conformado por representantes de las instituciones del Sistema Nacional de Salud y por el INEGI.

Asimismo, el INEGI capta información sobre recursos, servicios y morbilidad hospitalaria de los establecimientos particulares de atención a la salud, a través de un sistema estadístico que

incluye: un cuadernillo de captación, un instructivo de llenado, un manual para la critica-codificación de la información y programas de captura, de depuración y de obtención de resultados.

El disponer y proporcionar información de asistencia y servicios médicos tanto a los responsables de la toma de decisiones, como el sector académico y a la población en general, permite optimizar la asignación y aprovechamiento de los recursos, por una parte, así como, el conocimiento necesario de los problemas de salud existentes y del tipo de atención médico-hospitalaria al que se puede tener acceso.

Estadísticas Sociales

La generación e integración de estadísticas sociales también tiene como fuente básica a diversos registros administrativos, donde se asientan en forma continua, datos sobre fenómenos sociales, cuya traducción a información estadística permita un mejor conocimiento de la realidad social. En este apartado nos referimos a las Estadísticas de Educación, sobre Relaciones Laborales, sobre Seguridad y Orden Público y las de Cultura.

Esta labor es encomendada a la Dirección General de Estadística desde su conformación como tal en 1882. Así, se puede mencionar la producción de estadísticas referentes a Bibliotecas, Espectáculos Públicos y Museos. Si bien es 1928 el año en que inicia la publicación en forma sistemática de esta información, ya en 1889-1909 se presentan algunos datos sobre Bibliotecas y Museos.

La estadística sobre el **Sistema Educativo Nacional** es producto de un avanzado esquema de concertación interinstitucional y un funcional esquema de coordinación y desconcentración. En efecto, conforme a las bases de coordinación que celebraran en 1976 la Secretaría de Educación Pública (SEP) y la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES), con la Dirección General de Estadística, son estas dependencias las que se ocupan de captar e integrar la estadística del Sector Educativo, la cual comprende una amplia temática y la totalidad de los ciclos educativos, tanto de los establecimientos públicos como privados.

La recopilación de los datos se hace a través de los cuestionarios que distribuyen la SEP y la ANUIES a cada uno de los centros educativos, al inicio y al fin de cursos, mismos que son recabados y procesados por la SEP en cada estado, realizando esta misma dependencia la integración de cifras nacionales a nivel central en la Ciudad de México, pudiéndose obtener datos con desagregación por entidad federativa y municipio, área urbana o rural, sexo, edad, control administrativo, ciclo educativo y otras variables.

Si bien la generación de información se centra en la producción de estadística básica, es importante mencionar que la disponibilidad de datos para series históricas amplias y la calidad de los datos permite la integración de indicadores educativos que, conjuntamente con la estadística básica, sirven al objetivo con el cual se producen: apoyar la toma de decisiones en los procesos de planeación y evaluación de las actividades propias del sector, así como en la formulación de políticas y en la investigación del fenómeno, además de fomentar la participación informada de todos los sectores de la población que involucra un aspecto social tan trascendente y amplio como es el de la educación.

Otro tema que se cubre aprovechando los registros administrativos, es el referente a las negociaciones obrero-patronales. Bajo el nombre de **Estadísticas sobre Relaciones Laborales** se capta información de los diferentes momentos por los que atraviesa la negociación obrero-patronal ante el surgimiento de un conflicto de intereses, cubriendo desde los convenios de trabajo fuera de juicio hasta la solución de huelgas, haciendo distinción de los conflictos individuales de trabajo respecto a los conflictos colectivos. Además, se capta la solución de dichos conflictos, los emplazamientos a huelga y su solución, así como las huelgas estalladas.

Conviene destacar que la información que genera el INEGI sobre el tema, corresponde únicamente a los eventos de jurisdicción local, siendo la Secretaría del Trabajo y Previsión Social la dependencia encargada de hacer lo propio con los conflictos laborales de jurisdicción federal. De esta manera, son las Juntas Locales de Conciliación y Arbitraje, así como las municipales y regionales de conciliación, las instancias que mensualmente proporcionan información en boletas estadísticas que con tal fin ha diseñado la DEDS.

La estadística que se genera a partir de los datos recabados se desagrega por entidad federativa y municipio, así como por sector de actividad, obteniéndose el total de eventos y los trabajadores involucrados en cada uno de los siguientes conceptos: Convenios de trabajo fuera de juicio, Conflictos individuales, Conflictos colectivos, Solución de conflictos, Emplazamiento a huelga, Solución de emplazamientos a huelga, Huelgas y Solución de huelgas.

También, se tabula información conforme a variables como: mes de registro, motivo del conflicto, condición sindical, modalidad de contratación, clase de contrato, organización obrera y tipo de sindicato emplazante, entre otras.

La información que así se produce, aporta importantes elementos de análisis para investigar, planear y evaluar las características y algunos de los efectos de la política laboral, así como para medir el grado, frecuencia, magnitud y características con que se presentan los conflictos laborales, entre los que destaca la huelga.

Bajo el rubro de **Estadísticas de Seguridad y Orden Público** actualmente, y desde los años treinta, se viene produciendo información sobre dos grandes temas relacionados: las estadísticas judiciales y las estadísticas sobre suicidios e intentos de suicidio. Otras estadísticas que también se producían bajo este rubro y dejaron de captarse a finales de la década pasada fueron las de juicios de amparo, incendios y cuerpo de bomberos.

Las estadísticas judiciales comprenden dos conceptos, a saber: los presuntos delincuentes y los delincuentes sentenciados. Estos se captan tanto para el fuero común como para el fuero federal, por lo que las fuentes de información son los juzgados de primera instancia, de ambos fueros, que conocen de delitos en materia penal.

Las variables que se captan abarcan aspectos propios de los hechos delictivos, como: delito, situación jurídico-penal, sentencia y estado psicofisiológico en que se encontraba la persona al momento de cometer el delito; características sociodemográficas de los sujetos, como: sexo, edad, estado civil, ocupación y condición de alfabetismo; y variables geográficas y temporales: lugar de registro, de ocurrencia y de residencia habitual del sujeto (desagregada por entidad federativa y municipio), así como mes de registro.

La información se recaba mensualmente de cada uno de los juzgados a través de las áreas estatales de estadísticas continuas, utilizando los formatos (boletas unitarias) diseñados y distribuidos por el INEGI para este propósito. Una vez requisitados los formatos son concentrados en la ciudad de Aguascalientes, donde reciben tratamiento manual y electrónico, haciendo uso de manuales y catálogos que permiten la homogeneidad en el tratamiento y la comparabilidad de los resultados.

Sobre este último aspecto, conviene mencionar que en la Dirección de Estadísticas Demográficas y Sociales se ha diseñado un catálogo de delitos que recoge los distintos tipos penales que se encontraban considerados en los códigos penales de los estados hasta el año de 1990. Haciendo uso de este catálogo se codifica y tabula la información, buscando que los resultados por entidad federativa puedan ser comparables.

Ahora bien, la utilidad de la estadística que así se obtiene radica en el uso que de ella hacen las autoridades responsables de la procuración, administración e impartición de justicia a nivel federal y estatal, como sería la Procuraduría General de la República y las Procuradurías de Justicia en los Estados, los cuerpos policíacos y los diferentes órganos del Poder Judicial, tanto federal como estatal. Asimismo, la información es demandada por otras instituciones abocadas al estudio de fenómenos sociales, así como a la docencia e investigación.

El otro tema que se capta actualmente, es: suicidios e intentos de suicidio, fenómeno de sumo interés por las interpretaciones e implicaciones de carácter social que se encuentran asociadas al mismo. Aunque se tienen noticias de que esta información se obtenía desde el siglo pasado, no es sino a partir de los años treinta que se ha venido publicando regularmente.

La fuente de información para la estadística de suicidios e intentos de suicidio son las agencias del ministerio público, las cuales, a partir de sus averiguaciones llenan mensualmente los formatos que el INEGI les hace llegar a través de sus áreas estatales, mismas que se encargan de recabarlos y remitirlos a la ciudad de Aguascalientes, donde la información es procesada.

La cobertura temática abarca variables asociadas a la comisión del acto suicida, como lo es el lugar y medio empleado, el motivo y antecedentes de suicidas en la familia del suicida; por otra parte se recaba información sociodemográfica de los suicidas (sexo, edad, alfabetismo, ocupación, estado civil, posesión de hijos y religión).

La cultura constituye un fenómeno vasto y complejo, cuya medición representa, igualmente, una tarea de amplios alcances y gran dificultad. Sin embargo, el INEGI, haciendo uso de los registros administrativos existentes, viene produciendo información sobre algunos tópicos de interés. En coordinación con la Secretaría de Educación Pública y bajo el mismo esquema que se genera la estadística educativa, el INEGI produce las **Estadísticas de Cultura**, esto es, integra y difunde información sobre bibliotecas (número, acervo, servicios y usuarios); asimismo, en forma directa, el INEGI capta información sobre cines, museos y espectáculos públicos.

En la estadística de cines se recaba mensualmente, de cada uno de los establecimientos existentes en el país, datos sobre capacidad, asistencia, proyecciones y número de funciones, así como información sobre la nacionalidad de las películas. En el caso de la estadística sobre espectáculos públicos se capta información similar, solo que se indaga sobre el tipo de espectáculo presentado, clasificándolos en: deportivos, teatrales, taurinos y recreativos.

La estadística de museos abarca todo tipo de establecimientos que tengan como actividad permanente y sin fines de lucro la exhibición pública de colecciones en que se muestre la obra del hombre o de su medio ambiente, ya sea con propósito de estudio, educación o deleite. Así, se capta información de institutos de conservación y galerías de exposición; lugares y monumentos arqueológicos, etnográficos y naturales, establecimientos que exponen especies vivientes, tales como jardines botánicos, zoológicos, acuarios y viveros, así como planetarios y centros científicos.

Cabe mencionar que la integración de esta información se realiza de dos modos: uno, directo, a través de las áreas estatales de estadísticas continuas y otro, indirecto, por medio del Instituto Nacional de Antropología e Historia, quien proporciona a nivel central la información de los centros que se encuentran bajo su control.

Según puede observarse, el aprovechamiento de los registros administrativos como fuente de información permite cubrir un amplio espectro temático, además de que involucra a un gran número de instituciones y facilita la obtención de información con mayor detalle y continuidad que la obtenida a través de censos o encuestas. Sin embargo, el uso combinado de los datos que el INEGI genera por los distintos métodos estadísticos, ayuda al estudio de la realidad económica y social del país.

De esta manera, con la colaboración de la población, que participa en la generación estadística al declarar y proporcionar datos; con el apoyo de las instituciones, que participan en el llenado y envío de los cuestionarios, y con la cooperación de especialistas e investigadores, que en su calidad de usuarios enriquecen el quehacer estadístico con sugerencias y aportaciones, el Servicio Nacional de Estadística cumple cada vez mejor con su objetivo, contribuyendo así al progreso y modernización del país.

UN ESTIMADOR PARA AJUSTAR MODELOS DE REGRESION LINEAL CON DATOS DE MUESTRAS COMPLEJAS, BASADO EN EL ESTIMADOR DE REGRESION GENERALIZADO: CONSTRUCCION Y CARACTERISTICAS[1]

MARTIN HUMBERTO FELIX MEDINA
ESC. DE CIENCIAS FISICO MATEMATICAS-UAS

1. Introducción

Diversos estudios muestran que el estimador de mínimos cuadrados ordinarios (EMCO), es generalmente inadecuado para el ajuste de modelos de regresión lineal con datos de muestras complejas de poblaciones finitas. En particular, Nathan y Holt (1980) muestran que con muestras no autoponderadas se tienen problemas de sesgo que afectan sus propiedades de cobertura.

Se han propuesto varios estimadores para esta situación, algunos derivados bajo el enfoque inferencial basado en diseño y otros bajo el basado en modelo. Los primeros funcionan aceptablemente en una amplia gama de diseños y poblaciones, sin embargo, son menos eficientes (en error cuadrático medio) que los basados en modelo, aunque estos últimos no son robustos a la especificación errónea del modelo.

Con la idea de obtener estimadores robustos y eficientes, Pfeffermann y Holmes (1985) sugieren construir estimadores que combinen la información sobre las relaciones entre las variables del modelo de regresión y las diseño. Plantean algunas estrategias particulares, una de las cuales se desarrolla en la presente investigación.

En este trabajo se construye un estimador con las características anteriores, el cual se obtiene a partir del estimador de regresión generalizado (ERG), propuesto por Cassel et al (1976) para estimar medias poblacionales. Se presentan algunas propiedades estadísticas asintóticas del estimador obtenido, y mediante un estudio de simulación, se compara el comportamiento de este estimador con el de otros que se han propuesto.

[1] **Tesis de la Maestría en Estadística e Investigación de Operaciones de la UACPyP-CCH UNAM, dirigida por el Dr. Ignacio Méndez.**

2. El coeficiente de regresión de población finita

Considérese una población finita $U=\{1,\ldots,N\}$, cuyo i-ésimo elemento tiene asociado el vector (y_i, x_i, z_i^*), donde $y_i \in \mathbb{R}$ es la variable respuesta, $x_i \in \mathbb{R}^p$ la variable regresora, y $z_i^* \in \mathbb{R}^{q^*}$ la variable diseño. La población finita de vectores es $P=\{(y_i, x_i, z_i^*): i=1,\ldots,N\}$, y al igual que en Kish y Frankel (1974), el parámetro de regresión de población finita de interés se define por:

$$B_U = (X_U' X_U)^{-1} X_U' Y_U$$

donde $Y_U=[y_i]_N$ y $X_U=[x_i]_{N,p}$. Cabe aclarar que la idea subyacente a esta definición es la existencia de un modelo superpoblacional de regresión lineal entre **y** y **x**.

Con el objetivo de estimar B_U, mediante un diseño muestral $p(S|Z_U^*)$, basado en z^* $(Z_U^*=[z_i^*]_{N,q^*})$, no informativo y monoetápico, se toma una muestra **S** de tamaño n de la población finita **U**.

Como lo señalan Nathan y Holt (1980), si la variable z^* está correlacionada con las variables **y** y **x**, y el diseño es no autoponderado, el estimador de mínimos cuadrados ordinarios presenta problemas de sesgo que incrementan de manera importante su error cuadrático medio. Existen varios estimadores para esta situación, (algunos de ellos se consideran en el estudio de simulación), el de uso más frecuente es el propuesto por Kish y Frankel: $\hat{B}_{KF}=(X_S' W_S X_S)^{-1} X_S' W_S Y_S$, donde X_S y Y_S son los equivalentes muestrales de X_U y Y_U, $W_S=\mathrm{Diag}(w_i)_{n,n}$ y $w_i=1/\Pr(i\in S)$.

3. El estimador $\hat{B}_{RG}$ de B_U

La idea detrás del estimador B_{KF} es estimar las entradas de las matrices $X_U'X_U$ y $X_U'Y_U$ mediante el uso de estimadores tipo Horvitz-Thompson. Se toma por tanto en cuenta el procedimiento de selección muestral, pero se omite la información sobre la relación entre la variable diseño y las del modelo de regresión.

Dentro del contexto del muestreo, los estimadores de razón y los de regresión se utilizan porque permiten incorporar la información auxiliar dis-

ponible. En nuestro caso, se construirá un estimador que estime las entradas de las matrices anteriores mediante el estimador de regresión generalizado (ERG) propuesto por Cassel et al (1976). Para ello se supondrá que las relaciones entre $x_{ij}x_{ik}$ y z_i^*, por un lado, y entre $x_{ij}y_i$ y z_i^* por el otro, se pueden aproximar mediante el siguiente modelo de regresión lineal:

$$x_{ij}x_{ik}=[1\ \ g(z_i^*)]\begin{bmatrix}\theta_{0(jk)}\\ \theta_{g(jk)}\end{bmatrix}+\varepsilon_{i(jk)}=z_i\theta_{(jk)}+\varepsilon_{i(jk)}\ ,\ j,k=1,..,p$$

$$x_{ij}y_i=[1\ \ g(z_i^*)]\begin{bmatrix}\theta_{0(jY)}\\ \theta_{g(jY)}\end{bmatrix}+\varepsilon_{i(jY)}=z_i\theta_{(jY)}+\varepsilon_{i(jY)}\ ,\quad j=1,..,p$$

donde $z_i=[1,g(z_i^*)]$, la función $g:\mathbb{R}^{q^*}\to\mathbb{R}^{q-1}$ es conocida y $\theta_{(jk)}$ y $\theta_{(jY)}\in\mathbb{R}^q$ son parámetros a estimar.

Si $z^*\in\mathbb{R}$, dos aproximaciones a considerar son la lineal: $g(z^*)=z^*$ y la cuadrática: $g(z^*)=(z^*,z^{*2})$, ambas estudiadas numéricamente en este trabajo.

Una vez que se define la función g, los ERG de $X_U'X_U$ y $X_U'Y_U$, son:

$$\hat{X_U'X_U}=\sum_{i\in S}w_ih_{iS}x_i'x_i=\sum_{i\in S}w_{iS}^*x_i'x_i=X_S'W_S^*X_S\quad y\quad \hat{X_U'Y_U}=\sum_{i\in S}w_ih_{iS}x_i'y_i=\sum_{i\in S}w_{iS}^*x_i'y_i=X_S'W_S^*Y_S$$

donde $h_{iS}=1_U'Z_U(Z_S'W_SZ_S)^{-1}z_i$, $w_{iS}^*=h_{iS}w_i$ y $W_S^*=Diag(w_{iS}^*)_{n,n}$

Finalmente, el estimador de B_U, basado en el ERG, es $\hat{B}_{RG_S}=(X_S'W_S^*X_S)^{-1}X_S'W_S^*\ Y_S$

4. Propiedades estadísticas asintóticas basadas en diseño de $\hat{B}_{RG}$

Puede mostrarse que bajo ciertas condiciones (Félix, 1993), el estimador $\hat{B}_{RG}$ satisface:

$$\hat{B}_{RG}-B_U=O(n^{-1/2})\quad y\quad [V_p(\hat{B}_{RG_S})]^{-1/2}(\hat{B}_{RG_S}-B_U)\xrightarrow{D}N_p(0,I)$$

donde

$$V_p(\hat{B}_{RG_S})=(X_U'X_U)^{-1}V_p\Big[\sum_{i\in S}w_{iS}^*(y_i-x_iB_U)x_i'\Big](X_U'X_U)^{-1}=(X_U'X_U)^{-1}V_p[X_S'W_S^*E_S](X_U'X_U)^{-1}$$

y $E_S=[e_i]_n=[y_i-x_iB_U]_n$

Para obtener $V_p(\hat{B}_{RG_S})$, se debe calcular $V_p[X_S'W_S^*E_S]$. El vector aleatorio $X_S'W_S^*E_S$ tiene la estructura de un ERG, por tanto, su matriz de dispersión depende del diseño muestral. En un diseño con reemplazo y monoetápico:

$$V_p[X_S'W_S^*E_S]=[\sum_{i=1}^{N} n^{-1}(d_i - d.)'(d_i - d.)p_i]$$

donde p_i es la probabilidad de tomar la unidad i en cada extracción, $d_i=(e_ix_i-z_i\phi_U)p_i^{-1}$, $\phi_U=(Z_U'Z_U)^{-1}Z_U'(Diag(e_i))X_U$ y $d.=\sum_{i=1}^{N} d_ip_i$. Un estimador de $V_p(\hat{B}_{RG_S})$ se presenta en Félix (1993).

Cabe aclarar que el estimador B_{RG} presenta el problema de que algunos de los ponderadores $w_{i_S}^*=h_{i_S}w_i$ pueden ser negativos o nulos. Dos problemas derivados de éste hecho son que $X_S'W_S^*X_S$ no es necesariamente positiva definida y que no es posible el uso del software tradicional de mínimos cuadrados para el cálculo de $\hat{B}_{RG}$. Por el momento se ignora la existencia de otros problemas. Afortunadamente, si $X_S'W_SX_S$ es no singular, la estimación $V_p(\hat{B}_{RG_S})$ que se presenta en Félix (1993) es positiva definida.

5. Estudio de simulación

Se generaron doce poblaciones, mediante la consideración de diferentes relaciones entre las variables Y y X con Z, y diferentes distribuciones de las variables. En todos los casos la relación entre Y y X fue la de una regresión lineal simple. De cada población se tomaron muestras mediante diseños muestrales estratificados monoetápicos, tanto autoponderados, como no autoponderados. Los estimadores que se compararon fueron de dos clases: **No ponderados:** EMCO y máximo verosímil de DeMets y Halperin. **Ponderados:** estimador de Kish y Frankel, máximo verosímil ponderado de Nathan y Holt y, tres variantes del ERG definidas de acuerdo a la función g: cuadrática, lineal y puntajes (un único valor de Z para todos los elementos de un mismo estrato).

Los resultados del estudio se presentan en Félix (1993). Las principales conclusiones son las siguientes:

- El EMCO es inadecuado en diseños no autoponderados
- El estimador máximo verosímil es el de mejor comportamiento en las poblaciones con relaciones lineales y homocedásticas. Con relaciones cuadráticas o con perturbaciones heterocedásticas es inadecuado.
- El comportamiento de los estimadores ponderados es bastante similar y aceptable en todos los diseños y poblaciones. Son menos eficientes que los no ponderados bajo las condiciones en que éstos se derivaron.
- El uso de información auxiliar en el estimador $\hat{B}_{RG}$ no se manifiesta en un mejor comportamiento con respecto al de Kish y Frankel. Asimismo, no se perciben diferencias entre las tres variantes del estimador $\hat{B}_{RG}$.

De los resultados obtenidos se hace la siguiente recomendación práctica: Cuando se tenga la certeza de que las relaciones entre las variables del modelo de regresión y las diseño son lineales y homocedásticas úsese el estimador máximo verosímil, en caso contrario, el propuesto por Kish y Frankel.

6. Referencias

- Cassel C.M., Särndal C.E. and Wretman J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika 63 615-20.
- Félix M.M. (1993). Un estimador para ajustar modelos de regresión lineal con datos de muestras complejas, basado en el estimador de regresión generalizado: construcción y características. Tesis de maestría (no publicada), UNAM.
- Kish L. and Frankel M.R. (1974). Inference from complex samples (with discussion). Journal of the Royal Statistical Society B 36 1-37.
- Nathan G. and Holt D. (1980). The effect of survey design on regression analysis. Journal of the Royal Statistical Society B 42 377-86.
- Pfeffermann D. and Holmes D.J.(1985). Robustness considerations in the choice of method of inference for the regression analysis of survey data. Journal of the Royal Statistical Society A 148 268-78.

"INDUSTRIALIZACION DEL Rhizobium sp., UNA SOLUCION A TRAVES DE TECNICAS DE SUPERFICIES DE RESPUESTA"

Edgar Guadiana Ordaz
Yuria Cardel Sánchez
Esteban Burguete
Departamento de Ingeniería Industrial
Universidad de las Américas-Puebla
Sta. Catarina Mártir 72820, Puebla.

Abstract. Las bacterias del género Rhizobium se unen a las células de las raíces de las leguminosas para fijar el nitrógeno atmosférico. En esta condición pueden llegar a fijar hasta 100 kg/Ha de nitrógeno aprovechable por las plantas. Lo que es un indicativo de su enorme importancia económica.

El medio de cultivo necesario para la industrialización de Rhizobium sp. que se ha usado tradicionalmente como fuente de carbono es el manitol, cuya desventaja es su elevado costo. Tres alternativas más económicas (glicerol, sacarosa y glucosa) son estudiadas en este trabajo, a través de un experimento diseñado para comparar niveles de estas tres alternativas con manitol. Finalmente, direcciones de máxima pendiente para búsqueda de un óptimo económico son encontradas y se presentan propuestas para investigaciones posteriores.

1. Introducción.

El nitrógeno es un bioelemento fundamental para los seres vivos. Gran cantidad de este elemento se encuentra en la atmósfera en forma no metabolizable para ellos.

La fuente de nitrógeno para la síntesis de aminoácidos y proteínas es el amonio del suelo y el agua. El amonio es absorbido por las plantas y pasa a formar parte de los aminoácidos y proteínas. Las plantas pueden ser ingeridas por los animales, que a su vez emplean los aminoácidos de las proteínas vegetales para sintetizar sus propias proteínas.

Las bacterias del género Rhizobium no son capaces de fijar el nitrógeno atmosférico ellas solas, por lo que se combinan con células de las raíces de leguminosas como chícharos y habas. Las bacterias invaden las raíces y estimulan la formación de nódulos radiculares.

La unión entre la célula de la leguminosa y la bacteria puede fijar el nitrógeno atmosférico (lo que no podrían hacer ninguna de las dos por separado). Estas bacterias en los nódulos pueden fijar anualmente de 50 a 100 kg. de nitrógeno por hectárea, en tanto que las bacterias libres del suelo fijan hasta 12 kg. por hectárea.

2. Justificación del experimento.

Primeramente se aisló y reprodujo la cepa. Posteriormente en una sesión de tormenta de ideas, se estudiaron como posibles factores de la producción a:

1) Material de laboratorio.
2) Equipo de laboratorio.
3) Reactivos para el medio de cultivo, y
4) Mano de obra.

Se determinó, como área de oportunidad a los reactivos para el medio de cultivo. Dentro de los reactivos para el medio de cultivo se utiliza una fuente de carbono conocida como manitol, esta fuente es muy bien aceptada por la bacteria pero presenta una desventaja: su precio y disponibilidad. De esta manera se sugirió la utilización de otras fuentes de carbono más económicas las cuales deberían ser probadas. Se empleó un diseño 2^3 de efectos fijos, donde los factores a considerar fueron las fuentes de carbono en 2 concentraciones:

FACTORES:	NIVELES	
	Bajo (-)	Alto (+)
A: GLICEROL	0.075ml	0.375ml
B: SACAROSA	0.075g.	0.375g.
C: GLUCOSA	0.075g	0.375g.

(Todo en 75ml de medio)

3. Resultados y Discusión.

Las lecturas se hicieron con la ayuda de un espectrofotómetro el cual registra el nivel de absorbancia a 640 Nm. Las lecturas fueron:

1	372	314
a	344	319
b	353	343
ab	301	322
c	294	307
ac	245	310
cb	325	324
abc	283	317
control	398	389

Al realizar el Anova, resulta que C (glucosa) es significativo al 5% y A (glicerol) es significativo al 10%, con todos los demás efectos no significativos. Al graficar en papel normal los residuos resulta en una recta, además, los residuos estandarizados se encuentran entre -2 y 2 por lo que aceptamos el modelo como adecuado.

En búsqueda de las cantidades óptimas para las fuentes de carbono se modeló una superficie de respuesta lineal:

$$W = -79.583A - 109.583\,C + 359.625$$

La máxima pendiente nos llevaría a descender en los niveles de A y C (glicerol y glucosa), sin embargo este resultado no es biológicamente recomendable debido a que se afectaría la etapa de conservación de la bacteria. De esta manera es factible pensar que el óptimo se encuentra entre las concentraciones usadas por lo que se recomienda proceder con un diseño central compuesto.

4. CONCLUSIONES

Analizando las mediciones en el espectrofotómetro vemos que el manitol no puede ser reemplazado por ninguna otra fuente de carbono en lo que respecta a

velocidad de crecimiento. Sin embargo, tanto el glicerol como la glucosa exhiben un buen comportamiento en el caso que se busque la conservación de la cepa sin la necesidad de un crecimiento muy acelerado.

Los niveles de absorbancia resultan mayores en la concentración baja tanto para la glucosa como el glicerol. El incremento de la concentración trajo consigo una disminución en el crecimiento de las bacterias, esto se debe a que al aumentar las concentraciones del glicerol y de la glucosa producen un fenómeno conocido como presión osmótica, el cual inhibe el crecimiento de la bacteria. La aportación de la sacarosa no resultó significativa, sin embargo, el efecto de ésta es el único positivo. Lo que indica que a mayor concentración se espera un mayor crecimiento.

La interacción de glucosa y glicerol es considerada nula. La interacción del glicerol y la sacarosa solo puede ser considerada como pequeña en concentraciones altas, debido a que para ambos en su nivel alto el crecimiento es menor. La interacción sacarosa-glucosa presenta un incremento en el crecimiento cuando se combina el nivel alto de sacarosa con el bajo de glucosa.

Se sospecha de una posible reacción (especialmente entre la glucosa y el glicerol) en el medio de cultivo al usar las concentraciones altas. Dicha reacción trae como consecuencia una disminución del crecimiento bacteriano. De esta forma es explicado el hecho de haber observado mayor crecimiento a concentraciones bajas de las fuentes de carbono glicerol y glucosa.

Así pues, se propone el uso de un diseño central compuesto para dar seguimiento a la investigación, considerando los siguientes intervalos:

Glicerol	[0.075g. 0.250g]
Sacarosa	[0.075g. 0.375g.]
Glucosa	[0.075g. 0.200g.]

BIBLIOGRAFIA.

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

Arms, K. and Camp, P. S. (1991) A Journey into Life. Saunders College Pub. Co. Fort Worth, TX.

Análisis Geoestadístico de la difusión de aguas de desecho en Altamira, Tamaulipas

Claudia Lara Pérez Soto
IIMAS, UNAM CC, IBM

Introducción

Uno de los grandes problemas en la actualidad es la contaminación ambiental, la cual ha ido creciendo a través del tiempo de manera desproporcionada.

Hoy en dia se están llevando a cabo planes para controlar la contaminación, siendo necesario analizar su comportamiento para planear acciones que ofrezcan soluciones rápidas y confiables.

Para aplicar estas acciones se han hechos, estudios dando prioridad a la contaminación del aire, sin embargo, es importante tambien considerar la contaminación del agua que ha llegado a representar otro grave problema para la sociedad.

Uno de los principales factores de contaminación proviene de los desechos industriales, causantes de un marcado desequilibrio ecológico en las áreas donde se localizan, tal es el caso del corredor industrial que se encuentra en Altamira, Tamaulipas en donde gran cantidad de desperdicios petroquímicos son desechados en la costa del Golfo de México.

Estudios de impacto ambiental han sido llevados a cabo, para determinar que tan afectada se encuentra la zona, esto, con varios propósitos, uno en el cual el órgano regidor de protección ambiental determine los límites de tolerancia permitidos y otro que permita saber que tanto se está afectando la zona y como se distribuyen los contaminantes.

El corredor está formado por industrias que conducen sus desechos por medio de difusores, algunos de ellos de tipo marino y/o subterráneo. En particular se decidió trabajar en una zona por la que pasa un difusor marino, que se encuentra sumergido a lo largo de dos kms. desde la costa y siguiendo el contorno del fondo, contando con varias boquillas en su último tramo de un km.

El estudio de la zona se esta llevando acabo por la Secretaria de Marina con el apoyo del Centro Científico de IBM para lo cual propusieron 3 muestreos, los cuales dependen de la situación climatológica del lugar: secas, lluvias y nortes; esto se hizo con la idea de establecer el comportamiento de difusión en cada una de las situaciones.

La zona elegida para el muestreo consta de un área de 5 x 5 kms. alrededor del difusor, se establecieron 30 estaciones de trabajo : 5 en costa y 25 en la superficie restante , en estas últimas se considero la profundidad de 0, 4, 8, 12, y 16 mts. cuando se pudo medir. Cada estación cuenta con un km de distancia entre una y otra. Se consideraron además estaciones testigo que van en función de la dirección de las corrientes y varian en número dependiendo de la climatologia.

En los puntos muestreados se considero importante medir 3 tipos de factores:

i) Biológicos : Nutrientes.

ii) Físicos : Temperatura, Corrientes, Tiempo.

iii) Químicos Metales, Oxígeno Disuelto, PH, Salinidad.

para analizar su comportamiento ante la presencia del difusor.

Dentro del estudio se utilizaron diferentes metodologías para realizarlo. Estos métodos son : Estudios en Organismos Bentónicos, Procesamiento Digital de Imagenes y Visualización de Datos.

Dentro de los objetivos principales se considero importante proponer modelos de difusión para los factores afectados con la idea de saber si el grado de variación cambia en exceso y si estos cambios afectan biológicamente tanto a la flora como a la fauna marina, a corto y largo plazo.

Análisis Geoestadístico

Este tipo de problemas son estudiosdos por Geoestadística basado en la Teoría de Variables Regionalizadas (TVR), donde se considera una variable regionalizada como una variable aleatoria que toma diferentes valores de acuerdo a su posición dentro de una zona determinada. Además la TVR esta basada en la Teoría de Funciones Aleatorias, gracias a esto la función puede ser representada como un modelo con un componente determinístico y otro de tipo estocástico que representa la dispersión del modelo.

Las aplicaciones en Geoestadística son básicamente estimaciones de la variable de interés en un punto o en un bloque no muestreado, a partir de la función propuesta en los puntos muestreados.

Otros factores importantes de la estructura de un fenómeno son la continuidad y regularidad espacial, los cuales estan relacionados con el comportamiento del variograma y tienen mucho que ver con las estimaciones que se hagan.

Esto permitirá construir modelos de dispersión para cada uno de los muestreos y ver hacia donde se dispersan los contaminantes valorando si la zona esta siendo afectada o no.

Variograma

En Geoestadística los dos primeros momentos son los de mayor utilidad, el primero es la esperanza $E\{Z(x)\} = m(x)$ mientras que para el segundo son tres los que son tomados en cuenta:

i) Varianza que se define como el momento de segundo orden alrededor de $m(x)$, es decir

$$Var\{Z(x)\} = E\{[Z(x) - m(x)]^2\}$$

ii) Covarianza que para los puntos x_1 y x_2 se define

$$C(x_1, x_2) = E\{[Z(x_1) - m(x_1)][Z(x_2) - m(x_2)]\}$$

iii) Variograma que se define como la varianza del incremento $\{Z(x_1) - Z(x_2)\}$ y esta dada por

$$2\gamma(x_1, x_2) = \mathrm{Var}\{Z(x_1) - Z(x_2)\}$$

La función $\gamma(x_1, x_2)$ es llamado el *semivariograma.*

Para el variograma la distancia y/o la dirección es importante porque de esta manera muestran cambios en el rango o la meseta, esto se debe a que en la zona la variable en estudio no se igual para todos los puntos. Si **x** representa las coordenadas, **h** es el vector del modulo $|\mathbf{h}|$ y α la dirección, entonces $\gamma(|\mathbf{h}|, \alpha)$ representa el conjunto de semivariogramas con dirección lo cual será de utilidad para detectar *anisotropía* y por tanto sus ejes. Esto se hace para determinar direcciones correspondientes al rango máximo y mínimo.

La distribución espacial de los datos y los ejes anisotrópicos esta respresentados en la Fig. 1, los ejes estan dados por varigramas de tipo gaussiano que es :

$$\gamma(h) = 1 - \exp(-\frac{3h^2}{a^2})$$

donde h es el rezago y a es el rango del variograma con valor de meseta igual a 1, las direcciones de los ejes son 60 y 140 grados respectivamente(Fig. 2). Para construir la superficie generada por los ejes se considero la extención del variograma a dos dimensiones en donde el modelo del variograma es expresado como

$$\gamma(h) = \gamma(h_x, h_y) = \gamma_1(h_1)$$

y donde h_1 esta dada por

$$h_1 = \sqrt{\left(\frac{h_x}{a_x}\right)^2 + \left(\frac{h_y}{a_y}\right)^2}$$

donde h_x es la componente de h a lo largo del eje X y h_y es la del ejes Y.

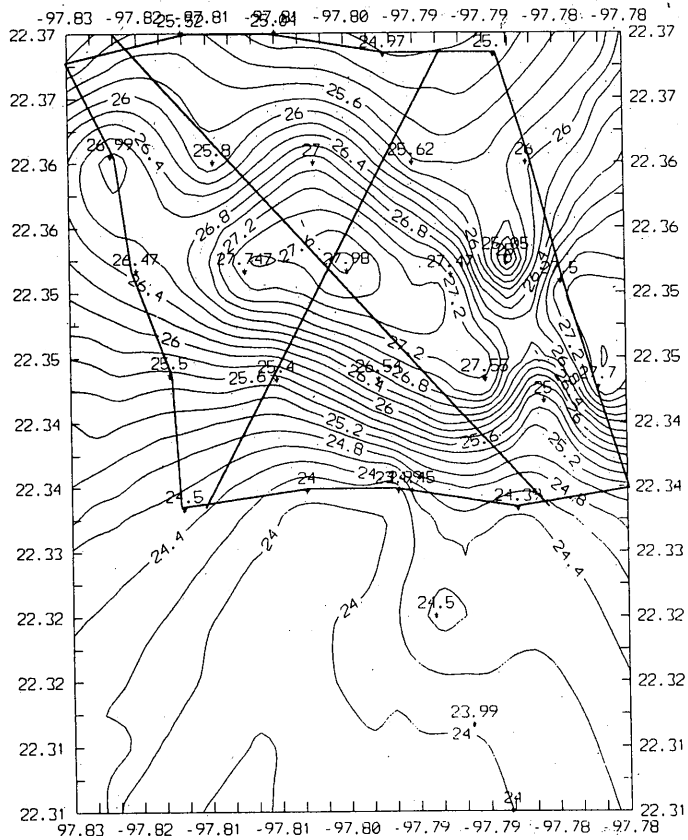


Fig. 1 Curvas de Nivel de la Variable Temperatura en Epoca de Secas, Malla de muestreo y Ejes de Anisotropía de 60 y 140 grados.

Es importante hacer notar que los ejes no necesariamente son ortogonales , la siguente matriz es la que hace la traslación de ejes

$$T = \begin{bmatrix} \dfrac{\operatorname{sen}\theta}{\operatorname{sen}(\theta-\varphi)} & \dfrac{-\cos\theta}{\operatorname{sen}(\theta-\varphi)} \\ \dfrac{\cos\varphi}{\operatorname{sen}(\theta-\varphi)} & \dfrac{-\operatorname{sen}\varphi}{\operatorname{sen}(\theta-\varphi)} \end{bmatrix}$$

donde θ y φ son los angulos. La superficie resultante se encuentra en la Fig. 3.

Median Polish

La palabra *Kriging* se refiere básicamente a hacer inferencias para los valores no observados en el proceso aleatorio dado que $\{Z(x): x \in D \subset R^d\}$ son tomadas como puntos localizados espacialmente y conocidos.

El elegir un buen predictor depende de la geometría y la localización de la región del espacio donde la predicción es deseada y donde esta es el proceso Z.

Tomando la descomposición

$$Z(x) = m(x) + \delta(x)$$

donde $\delta(x)$ es la estructura del error. Si m(x) es conocida se puede proporcionar al predictor óptimo lineal, pero no siempre es conocida. En dimesiones mayores se asume que el m(x) puede descomponerse en forma aditiva dentro de los componentes direccionales Por ejemplo, en R^2 se considera

$$m(s) = a + c(x) + r(y)$$

en donde $s = (x, y)^t \in D$, entonces una notación lógica es que $s_i = (x_l, y_k)^t$ implica que

$$m(s_i) = a + c_k + r_l$$

en donde c_k y r_l son el efecto columna y renglón. Esto puede ser visto como una tabla de contingencia, en donde los nodos son las entradas de la tabla y los estimadores por mínimos cuadrados estan dados por

$$a = \sum_{i=1}^{n} Z(s_i) / n$$

$$r_k = \{\sum_{N(y_k)} Z(s_i) / q\} - a$$

$$c_l = \{\sum_{N(x_l)} Z(s_i) / p\} - a$$

donde $k = 1,\ldots,p$ y $l = 1,\ldots,q$,

$$N(y_k) \equiv \{i: s_i = (., y_k)^t, i = 1,...,n\}$$
$$N(x_l) \equiv \{i: s_i = (x_l, .)^t, i = 1,...,n\}$$

La ventaja del estimador para m(x) es que es no paramétrica, espacialmente en su escala y continuidad. Su desventaja es que los residuos conducen a estimadores sesgados para la dependencia espacial desconocida en el proceso del error.

Actualmente se esta implementando esta técnica con la idea de aplicarlo en este problema, además es necesario generalizarlo para k dimensiones, pensando en la idea de que este problema será atacado para un mayor número de variables, es decir, pensando en *Cokriging*.

Bibliografia

Cressie, N.A. 1991, "Statistics for Spatial Data". Wiley, N.Y.

Hughes, J. P. and D. P. Letttenmainer. 1981. "Data requirements for kriging: Estimation and Network Design " . Water Res. Research, 17 (6): 1641-1650.

Isaaks, E. and Srivastava. 1990. "An Introduccion to Applied Geostatistics ". Oxford University Press, N.Y.

Journel, A.G. and CH. J. Huijbregts. 1978. "Mining Geostatistics". Academic Press, London.

Rippley, B.D. 1981. "Spatial Statistics". Wiley, N.Y.

Rippley, B.D. 1988. "Statistical Inference for Spatial Processes". Cambridge University Press

Robinson, G. K. 1990. "A Role for Variograms". Austral. J. Statist. 32 (3). 327-335.

Velleman, P.F and D. C. Hoaglin. 1981. "Applications, Basics, and Computing of Exploratory Data Analysis". Boston Massachusettes: Duxbury Press.

OPTIMIZACIÓN EN SUPERFICIES DE RESPUESTA SUJETA A p RESTRICCIONES LINEALES

Blanca Rosa Pérez Salvador
IIMAS, UNAM - UAM-I

Federico J. O'Reilly
IIMAS, UNAM

1 Introducción

En la industria, la medicina, la agricultura y en muchas otras áreas, surgen problemas cuya solución requiere encontrar la combinación de factores que optimiza la respuesta de un proceso específico. Si el proceso es descrito mediante la función

$$\eta(\mathbf{x}) = \beta_0 + \mathbf{b}^T\mathbf{x} + \mathbf{x}^T B\mathbf{x}$$

donde β_0, es un real; $\mathbf{b}$ es un vector de dimensión m y B es una matriz de orden $m \times m$ negativa definida; el valor de la respuesta máxima es:

$$\eta_{op} = \beta_0 - \frac{1}{4}\mathbf{b}^T B^{-1}\mathbf{b}.$$

y el punto en donde ésta se alcanza es:

$$\mathbf{x}_{op} = \frac{1}{2}B^{-1}\mathbf{b}.$$

Razones de operatividad como el alto costo o la escasez de un producto, y razones de incapacidad del sistema como no contar con un instrumento que emita cierta cantidad de radiación; limitan el dominio de la función de respuesta. Así, la necesidad de conocer el óptimo de $\eta(\mathbf{x})$ sujeto a restricciones en la localización del diseño, puede surgir naturalmente en situaciones prácticas y por lo tanto resulta de interés estudiar este problema.

Los trabajos reportados son escasos y en todos consideran a las restricciones como un conjunto de ecuaciones. Por ejemplo: Stablein, Carter y Wampler(1983) encontraron una región de confianza para el vector $\mathbf{x}$ en donde $\eta(\mathbf{x})$ alcanza el valor máximo sujeta a $g_i(\mathbf{x}) = C_i$; $i = 1, 2, ..., k$. Más tarde, Carter, Chinchilli, Myers y Campbell (1986) proponen un método para obtener los intervalos de confianza para las coordenadas del vector $\mathbf{x}_{op}$ y para los valores propios de la matriz B, sujetos a $\mathbf{x}^T\mathbf{x} = r^2$.

En la mayoría de las aplicaciones, un conjunto de ecuaciones no modela adecuadamente las restricciones. Ejemplos de restricciones son: la temperatura máxima que un horno puede alcanzar, la cual se puede representar por $T \leq T_0$; o el costo de dos (o más) substancias, que no debe exceder a un valor fijo c_0, y se puede representar por $c_1x_1 + c_2x_2 \leq c_0$. Aquí se consideran las restricciones como un conjunto de p desigualdades lineales, con lo cual se modela de modo más realista.

2 El óptimo restringido de $\eta(x)$

El problema de optimizar a $\eta(\mathbf{x})$ sujeto a p restricciones lineales se establece así:

$$\boxed{\text{Maximizar} \quad \eta(\mathbf{x}) = \beta_0 + \mathbf{b}^T\mathbf{x} + \mathbf{x}^T B\mathbf{x} \quad \text{sujeta a} \quad R\mathbf{x} \leq \mathbf{r}}$$

en donde R es una matriz de orden $p \times m$ de rango p $(p \leq m)$; $\mathbf{r}$ es un vector de dimensión p; y $\mathbf{x}$ es de dimensión m. Se puede probar que la solución está dada por:

$$\eta_{op_r} = \eta_{op} + (\mathbf{r} - R\mathbf{x}_{op} + \omega)^T(RB^{-1}R^T)^{-1}(\mathbf{r} - R\mathbf{x}_{op} + \omega) = \eta_{op} + (\mathbf{r} - R\mathbf{x}_{op})^T\upsilon$$

y $\mathbf{x}_{op_r} = \mathbf{x}_{op} + B^{-1}R^T(RB^{-1}R^T)^{-1}(\mathbf{r} - R\mathbf{x}_{op} + \omega) = \mathbf{x}_{op} + B^{-1}R^T\upsilon$

donde ω y $\upsilon \in R^p$ cumplen las condiciones de complementaridad lineal; esto es:

$$\omega^T\upsilon = 0; \qquad \omega \leq 0, \upsilon \leq 0 \qquad \text{y} \qquad \mathbf{r} - R\mathbf{x}_{op} + \omega = -RB^{-1}R^T\upsilon.$$

Observe que cuando $\upsilon = 0$, $\mathbf{x}_{op}$ cumple todas las restricciones y por lo tanto $\mathbf{x}_{op_r} = \mathbf{x}_{op}$. Por otro lado, cada coordenada de ω igual a 0 indica que la correspondiente restricción es activa; y por lo tanto, si $\omega = 0$, significa que todas las restricciones son activas.

El problema de complementaridad lineal se puede resolver utilizando el algoritmo propuesto por Quintana, O'Reilly y Gómez (1987), con el cual se determina las coordenadas de ω y υ que son diferentes de cero, reduciendose el problema a resolver un sistema de ecuaciones lineales.

3 Estimadores de η_{op_r} y de x_{op_r}

Considere que las observaciones son de la forma

$$Y = \eta(\mathbf{x}) + \varepsilon$$

con $\varepsilon \sim N(0, \sigma^2)$ independientes. Los estimadores de máxima verosimilitud para β_0, $\mathbf{b}$ y B, se encuentran con el método de mínimos cuadrados y por la propiedad de invarianza, los estimadores de máxima verosimilitud de η_{op_r} y de $\mathbf{x}_{op_r}$ son:

$$\hat{\eta}_{op_r} = \hat{\eta}_{op} + (\mathbf{r} - R\hat{\mathbf{x}}_{op} + \hat{\omega})^T(R\hat{B}^{-1}R^T)^{-1}(\mathbf{r} - R\hat{\mathbf{x}}_{op} + \hat{\omega}) = \hat{\eta}_{op} + (\mathbf{r} - R\hat{\mathbf{x}}_{op})^T\hat{\upsilon} \qquad (3.1)$$

$$\text{y } \hat{\mathbf{x}}_{op_r} = \hat{\mathbf{x}}_{op} + \hat{B}^{-1}R^T(R\hat{B}^{-1}R^T)^{-1}(\mathbf{r} - R\hat{\mathbf{x}}_{op} + \hat{\omega}) = \hat{\mathbf{x}}_{op} + \hat{B}^{-1}R^T\hat{\upsilon} \qquad (3.2)$$

donde $\hat{\omega}$ y $\hat{\upsilon} \in R^p$ son dos vectores tales que:

$$\hat{\omega}^T\hat{\upsilon} = 0; \qquad \hat{\omega} \leq 0; \qquad \hat{\upsilon} \leq 0 \qquad \text{y} \qquad \mathbf{r} - R\hat{\mathbf{x}}_{op} + \hat{\omega} = R\hat{B}^{-1}R^T\hat{\upsilon}.$$

Los métodos de estimación del óptimo basados en el clásico ascenso por pendiente máxima consideran acercamientos paulatinos hacia el óptimo y en caso de existir restricciones, los acercamientos estarán condicionados por las mismas. Se considera dos casos: en uno se permite violar las restricciones en la etapa experimental, en el otro no.

$\hat{\eta}_{op_r}$ y $\hat{\mathbf{x}}_{op_r}$ no son funciones lineales de $\hat{\beta}_0$, $\hat{\mathbf{b}}$ y $\hat{B}$; por lo tanto, sus propiedades estadísticas son difíciles de estudiar analíticamente y para sondearlas se realizó un estudio de simulación con las siguientes características:

- Se consideraron tres funciones de respuesta que dependen de dos factores, x_1 y x_2; esto es, $m = 2$.

  1. $\eta(\mathbf{x}) = (-x_1^2 - x_2^2 + x_1x_2 + 20x_1 + 20x_2)/4$
  2. $\eta(\mathbf{x}) = (-x_1^2 - x_2^2 - x_1x_2 + 60x_1 + 60x_2)/12$
  3. $\eta(\mathbf{x}) = (-x_1^2 - x_2^2 + 40x_1 + 40x_2)/8$

  En los tres casos, la respuesta máxima es $\eta_{op} = 100$, el vector donde se alcanza es $\mathbf{x}_{op} = (20, 20)$ y $\eta(0, 0) = 0$.

- Para estimar la tendencia lineal se utilizó el diseño factorial 2^2 con cuatro repeticiones en el centro y para la tendencia cuadrática, se usó un diseño compuesto.

- Se asignaron tres valores para σ: $\boxed{\sigma = 1,\ 5,\ \text{y}\ 10;}$. Con esto se tiene un coeficiente de variación en la respuesta óptima de 1% a 10%.

- Se consideraron tres tipos de restricciones,

$$\textbf{1)}\ \mathbf{x} \leq \begin{pmatrix} 19 \\ 19 \end{pmatrix} \quad \textbf{2)}\ \mathbf{x} \leq \begin{pmatrix} 10 \\ 10 \end{pmatrix} \quad \textbf{3)}\ \mathbf{x} \leq \begin{pmatrix} 100 \\ 10 \end{pmatrix}$$

  En **1)** y **2)**, ambas restricciones son activas. En **3)**, aunque formalmente son dos restricciones, el valor igual a 100, es "grande" respecto a 20 (tomando en cuenta a σ^2) por lo que se puede tratar para todo efecto práctico, como si realmente fuera una única restricción.

- Se consideraron los dos manejos de las restricciones en la etapa de experimentación:

  1. Se respetan en la experimentación.
  2. Se pueden violar durante la misma.

En la tabla siguiente se muestra un fragmento de los resultados, CRE significa con restricciones durante la experimentación, SRE significa sin restricciones durante la experimentación.

Se reportan los resultados empíricos de las desviaciones estándares y los coeficientes de variación de $\hat{\eta}_{op}$ y de $\hat{\mathbf{x}}_{op}$, el sesgo de $\hat{\eta}_{op}$, el tamaño muestral promedio para alcanzar el óptimo y la frecuencia de aparición de las restricciones para las 100 corridas.

Tabla 4.1. Resultados empíricos con base en 100 corridas.

Las restricciones son $x_1 \leq 10$ y $x_2 \leq 10$.

Primera función. $\mathbf{x}_{op_r} = (10, 10)$, $\eta_{op_r} = 75$

Varianza		$\sigma = 10$		$\sigma = 5$		$\sigma = 1$	
Método		CRE	SRE	CRE	SRE	CRE	SRE
restric-	0	1	0	0	0	0	0
ciones	1	0	0	0	0	0	0
activas	2	99	100	100	100	100	100
$s_{\hat{\eta}_{op_r}}$		9.4840	2.2093	1.4262	1.0518	0.2531	0.2023
$s_{\hat{\mathbf{x}}_{op_r}}$		1.4362	0.0000	0.0000	0.0000	0.0000	0.0000
$sesgo(\hat{\eta}_{op_r})$		-1.4596	0.0437	-0.1869	-0.0201	-0.0192	0.0042
$CV(\hat{\mathbf{x}}_{op_r})$		0.1016	0.0000	0.0000	0.0000	0.0000	0.0000
$CV(\hat{\eta}_{op_r})$		0.1265	0.0294	0.0190	0.0140	0.0034	0.0026
$\hat{n}$		25	30	25	30	25	30

Segunda función. $\mathbf{x}_{op_r} = (10, 10)$, $\eta_{op_r} = 75$

Varianza		$\sigma = 10$		$\sigma = 5$		$\sigma = 1$	
Método		CRE	SRE	CRE	SRE	CRE	SRE
restric-	0	0	0	0	0	0	0
ciones	1	1	0	0	0	0	0
activas	2	99	100	100	100	100	100
$s_{\hat{\eta}_{op_r}}$		5.3513	2.3299	1.3153	1.1096	0.2516	0.1980
$s_{\hat{\mathbf{x}}_{op_r}}$		1.0522	0.0000	0.0000	0.0000	0.0000	0.0000
$sesgo(\hat{\eta}_{op_r})$		-1.3470	0.1217	-0.1598	0.0966	-0.0213	-0.0140
$CV(\hat{\mathbf{x}}_{op_r})$		0.0744	0.0000	0.0000	0.0000	0.0000	0.0000
$CV(\hat{\eta}_{op_r})$		0.0713	0.0311	0.0175	0.0148	0.0034	0.0026
$\hat{n}$		25	30	25	30	25	30

Tercera función. $\mathbf{x}_{op_r} = (10, 10)$, $\eta_{op_r} = 75$

Varianza		$\sigma = 10$		$\sigma = 5$		$\sigma = 1$	
Método		CRE	SRE	CRE	SRE	CRE	SRE
restric-	0	0	0	0	0	0	0
ciones	1	3	0	0	0	0	0
activas	2	97	100	100	100	100	100
$s_{\hat{\eta}_{op_r}}$		6.1261	1.7913	1.4622	0.9763	0.2694	0.2070
$s_{\hat{\mathbf{x}}_{op_r}}$		1.1844	0.0000	0.0000	0.0000	0.0000	0.0000
$sesgo(\hat{\eta}_{op_r})$		-1.6936	-0.3793	-0.2220	0.1235	-0.0178	0.0124
$CV(\hat{\mathbf{x}}_{op_r})$		0.0837	0.0000	0.0000	0.0000	0.0000	0.0000
$CV(\hat{\eta}_{op_r})$		0.0817	0.0239	0.0195	0.0130	0.0036	0.0028
$\hat{n}$		25	30	25	30	25	30

Cuando se eliminan las restricciones en la etapa experimental: 1) se obtienen estimadores más precisos. 2) se determina mejor el número de restricciones activas.

Conforme σ es menor, las estimaciones son más precisas.

Se tienen estimadores más precisos cuando en la realidad son dos las restricciones activas. Incluso se pueden tener una estimación exacta de $\mathbf{x}_{op_r}$.

5 Conclusiones

Los resultados de la simulación indican que los métodos de estimación propuestos son bastante precisos y no se requiere de una muestra muy grande.

BIBLIOGRAFIA.

1. Box, G. E. P. and Wilson, K. B. (1951), "On the Experimental Attainment of Optimum Conditions." Journal of the Royal Statistical Society, Ser. B. 13, p1-45.

2. Stablein, D. M., Carter, W. H. Jr. y Wampler G. L. (1983), "Confidence Regions for Constrained Optima in Response Surface Experiments." Biometrics, 39, p759-763.

3. Carter, W. H. Jr., Chinchilli, V. M., Myers, R. H. y Campbell, E. D. (1986), "Confidence Intervals and an Improved Ridge Analysis of Response Surfaces." Technometrics, vol. 28, no. 4, p339-346.

4. Myers, R. H. and Khuri, A. I. (1979) "A New Procedure for Steepest Ascent." Communications in Statistics, Part A-Theory and Methods, 8. p1359-1376.

5. Myers, R. H., Khuri, A. I. and Carter, W. H. Jr. (1989), "Response Surface Methodology: 1966-1988." Technometrics, V, 31, no. 2, p137-157.

6. Pérez, B. R. y O'Reilly, F. J. (1993), "Sobre el ascenso en pendiente máxima en superficies de Respuesta" Reportes de Investigación, IIMAS, UNAM Vol. 3, No. 26.

7. Quintana, J. M., O'Reilly, F. J. y Gómez, S. (1987) "Least Squares with Inequality Restrictions: A symmetric Positive-Definite Linear Complementery Problem Algorithm" J. Statist Comput Simul., V. 28, p127-143.

EVALUACION DE GENOTIPOS EN SERIES DE EXPERIMENTOS: DIFERENCIAS PARAMETROS GENETICOS GENERADOS EN DOS MODELOS

Jaime Sahagún Castellanos[1]

INTRODUCCION. La evaluación de los genotipos a través del tiempo y del espacio es una condición indispensable para estimar objetivamente su auténtico potencial agronómico y de rendimiento. Sin embargo, el análisis de varianza de la información producida ha sido realizado de acuerdo con modelos diferentes, pudiendo producir resultados diferentes también.

Cuando existe heterogeneidad en los agentes metereológicos dentro del área que cubre un programa de mejoramiento genético para cultivos anuales la consideración de que hay cruzamiento entre los factores "años" (A) y "localidades" (L) es errónea. En este caso la consideración de anidamiento de años en localidades seria más realista. Por supuesto, si por consideración natural, o por manejo, la incidencia de estos agentes ambientales en la variable de interés fuera muy similar a través de las localidades durante cada año los factores "años" y "localidades" serian cruzados. El objetivo de este estudio consiste en la evaluación, en términos de estimación de componentes de varianza y de heredabilidad, del efecto de considerar que A está anidado en L cuando en realidad A y L son factores cruzados.

[1] Profesor-Investigador, Departamento de Fitotecnia de la Universidad Autónoma Chapingo, CP 56290 Chapingo, Méx.

MODELOS. Cuando A y L son factores cruzados, el factor "bloques" (B) está anidado tanto en A como en L y los "genotipos" (G) se evaluan en cada combinación de año y localidad, la explicación de variable respuesta correspondiente al genotipo i en la repetición j de la localidad k durante el año m puede darse, omitiendo detalles obvios, en la forma

$$Y_{ijkm} = \mu + G_i + B_{j(km)} + L_k + (GL)_{ik} + A_m$$

$$+ (GA)_{im} + (GLA)_{ikm} + E_{ijkm.} \qquad (1)$$

Cuando A está anidado en L y B en A, en cambio, se tendrá el modelo

$$Y_{ijkm} = \mu + G_i + B_{j(km)} + A_{m(k)} + (GA)_{im(k)}$$

$$+ L_m + (GL)_{im} + E_{ijkm} .$$

Para el análisis, los factores G, L, A y B y E_{ijkm} serán considerados aleatorios e independientes y la estimación de componentes de varianza se realizará con el método de momentos.

RESULTADOS Y DISCUSION. Respecto a componentes de varianza, mientras que el modelo (2) no permite la estimación de los asociados a GLA y A GA, el modelo (1) la hace posible. Para cada uno de los factores comúnes a ambos modelos, G y GL, las estimaciones pueden ser diferentes; el estimador del componente de varianza asociado a GL cuando la evaluación se hace durante a años y cada experimento utiliza r repeticiones es, de acuerdo con el modelo (1), de la forma [CM(G) - CM(GAL)]/ar, en donde CM denota cuadro medio. Con el modelo (2) el estimador es equivalente a [CM(GL) - CM(GAL)]/ar - [CM(GA) - CM(GAL)]/lar en donde l es el número de localidades. Para el componente asociado a G la situación es la inversa: Mientras que en el modelo (2) el

estimador es [CM(G) - CM(GL)]/lar, en el modelo (1) toma la forma [CM(G) - CM(GL)]/lar - [CM(GA) - CM(GAL)]/lar. Cuando los factores A y L son cruzados, y dado que los componentes de varianza son no negativos, E[CM(GA)] ≥ E[CM(GAL)]; esto es, con el uso del modelo (1) se tendería a sobreestimar y a subestimar a los componentes de varianza asociados a G y a GL, respectivamente. En ambos casos el sesgo sería el valor absoluto de V(GA)/a, en donde V(GA) es la varianza de la interacción enter G y A, en el sentido del modelo (1).

Con respecto a la heredabilidad, concepto de enorme significado para el fitomejorador, los modelos (1) y (2) también ejercen efectos diferenciados. Definido este concepto como el cociente entre varianza genética (V(G)) y varianza fenotípica (varianza entre las medias fenotípicas) resulta que como esta varianza fenotípica tiene un mismo estimador en uno y otro modelo, si el modelo (1) fuera el indicado la adopción del modelo (2) produciría una sobreestimación.

CARACTERIZACION DE LOS MUNICIPIOS INDIGENAS CON LA TECNICA DE COMPONENTES PRINCIPALES

M. en C. Sergio de la Vega Estrada

En Noviembre de 1992, la Secretaría de Salud y el Instituto Nacional Indigenista realizaron el estudio "La Salud de los Pueblos Indígenas en México". Dentro de ese estudio, el Centro de Estudios en Población y Salud de la misma Secretaría se responsabilizó de varias cosas, entre ellas la asignación de Lengua predominante a cada uno de los 542 Municipios que conformaron el universo de estudio. Una vez que se encontró la Lengua predominante en los Municipios se realizó un análisis socioeconómico y demográfico de ellos.

En aquel estudio se buscó describir el comportamiento de municipios agrupados por la lengua hablada o por el Estado de pertenencia. El momento y características del estudio tan sólo permitieron apoyarse en el uso por separado de los porcentajes municipales o de las agrupaciones étnicas o estatales elegidas. Desde entonces existía la inquietud por encontrar un "orden" de los 542 Municipios que permitiera detectar diferencias entre ellos pero también semejanzas. Este orden se pretende explotar para construir Grupos de pertenencia y definir las características que les distinguen a cada uno de los Grupos formados. Si para cada Municipio se tiene una medida que permite su clasificación, cómo lograr medidas para ciertas agrupaciones "naturales" como son las Etnias, los Estados y las Etnias por Estado. Además de esto también existía la inquietud de establecer criterios para que dentro de municipios de muy altos niveles de marginación, se priorizara la atención en aquellos considerados como casos extremos.

VARIABLES

Con Componentes principales se logra reducir el número de variables y conservar la información inicial, es la transformación de un espacio de información en otro con mayores posibilidades de explicitación. Incluso con esto es detectable el nivel de influencia de cada una de las variables para el nuevo sistema de referencia y por lo tanto de no influencia. Esto es, que si se detectan las variables de mayor aportación para los resultados, también es factible detectar las variables de menor importancia, al grado de prescindir de algunas de ellas por el nivel de información desplegada.

Dieciseis variables fueron consideradas, nueve corresponden a datos sobre Población y siete sobre Vivienda:

p514nes	Porcentaje de Población de 5 a 14 años que no asisten a la escuela
phnv	Promedio de Hijos Nacidos Vivos de Mujeres de 25 a 29 años de edad
phfall	Porcentaje de Fallecimientos de HNV
panalfa	Porcentaje de Población de 6 años o más que no sabe leer o es analfabeta
pinstr	Porcentaje de Población de 15 años o más sin primaria o con primaria incompleta
ppea	Porcentaje de Población de económicamente activa
ppoboc1	Porcentaje de Población activa en el sector primario
pocviv	Promedio de Ocupantes por vivienda particular
poccrt	Promedio de Ocupantes por cuarto en viviendas particulares
pv1crt	Porcentaje de viviendas particulares con un cuarto
pvagent	Porcentaje de viviendas particulares con agua entubada
pvparlc	Porcentaje de viviendas particulares con paredes de lámina de cartón o materiales de desecho
pvteclc	Porcentaje de viviendas particulares con techos de lámina de cartón o materiales de desecho
pvpistrr	Porcentaje de viviendas particulares con piso diferente a tierra
pvdren	Porcentaje de viviendas particulares con drenaje
pvelec	Porcentaje de viviendas particulares con electricidad

INDICE DE MARGINALIDAD

Las características de medición señaladas determinan que el valor de la Primera Componente está relacionado con indicios de carencia o marginalidad dentro del conjunto de 542 Muncipios indígenas considerando las variables seleccionadas. Por ser un valor que puede ordenar los Municipios y por su escala de intervalo que permite comparaciones, es factible definir a este valor como un Indice de Marginalidad.

Los valores de la Primera Componente, con las variables estandarizadas conservan un orden sin importar la magnitud, orden que muestra las diferencias que prevalecen entre los Municipios. El valor de la función *ind* que combina las dieciseis variables para los 542 municipios encamina a la obtención de este índice deseado. La expresión queda de la siguiente forma:

$$\begin{aligned} ind = {} & 0.68\ p514nes + 0.60\ phnv + 0.54\ phfall + 0.82\ panalfa + 0.77\ pinstr \\ & + 0.11\ ppea + 0.53\ ppoboc1 + 0.37\ pocviv + 0.78\ poccrt + 0.65\ pv1crt \\ & - 0.64\ pvagent + 0.18\ pvparlc + 0.37\ pvteclc - 0.70\ pvpistrr - 0.52\ pvdren \\ & - 0.73\ pvelec \end{aligned}$$

donde se puede deducir que: la variable con más peso para los 542 municipios es la de Población analfabeta o que no sabe leer, por tener 0.824 como coeficiente. La segunda en importancia es la de Ocupantes por cuarto, con 0.783 La variable que menos contribuye a distinguir es la de

Población económicamente activa, por tener 0.113, la que le sigue como segunda de menor importancia es la de Viviendas con paredes de lámina de cartón o materiales de desecho con 0.175. La varianza para la primera componente es del 35.8%, la segunda tan sólo alcanza 10.1%

Con estos coeficientes y las variables estandarizadas, los tres primeros municipios con peores condiciones socioeconómicas son de habla mixteca, ellos son San Martín Peras y San Simón Zahuatlán del Estado de Oaxaca y Metlatonoc del Estado de Guerrero. Por el otro lado, las tres mejores condiciones existen en municipios zapotecos, todos ellos de Oaxaca: Guelatao de Juárez, El Espinal y Unión Hidalgo.

Cuadro 1. Municipios por Estado según grupo asignado

	MB	B	M	A	MA	Total
Campeche		2	1			3
Chiapas			7	14	16	37
Chihuahua			1	2		3
Durango				1		1
Guerrero			2	3	12	17
Hidalgo	1	3	10	3	2	19
Jalisco				1		1
México		1				1
Michoacán		2	2			4
Nayarit					1	1
Oaxaca	21	30	106	63	42	262
Puebla	2	6	18	12	14	52
Quintana Roo		1	2			3
San Luis Potosí			11	2		13
Veracruz			13	11	14	38
Yucatán	4	25	51	7		87
Total	28	70	224	119	101	542

CINCO ESTRATOS

Con la técnica de Dalenius aplicada al índice encontrado, se establecen cinco grupos, esta agrupación permite identificar semejantes y diferentes. Por la manera en que es construído el índice y por el hecho de realizar grupos con él, es posible caracterizar a los Municipios en lo general y por grupo. Con la medida que asigna pesos específicos a las variables originales se describe una estructura de los datos en general (índice) y, con la obtención de la medida por grupo, es factible establecer

una estructura de los datos al interior de cada uno de ellos. Esta descripción de las variables trascendentales y las que podrían no ser consideradas servirá para la caracterización de los Municipios, en lo general y por Grupo.

La nomenclatura de los cinco grupos trata de sintetizar el nivel socioeconómico: muy baja, baja, media, alta y muy alta marginación, el cuadro uno muestra esta división por Estado.

Cuadro 2. Población indígena por grupos étnicos según grupo asignado

	MB	B	M	A	MA	Total
AMUZGO			3688		15164	18852
CHATINO			11154	6513	11078	28745
CHINANTECO		1899	31524	50474	5855	89752
CHOCHO			521			521
CHOL			26872	102131	15005	144008
CORA					16619	16619
CUICATECO			2284	2359	5060	9703
HUASTECO			99876	24983		124859
HUAVE			2668	9416		12084
HUICHOL				8031		8031
MAYA	39473	207670	329189	37508		613840
MAZATECO		1143	54061	52057	53435	160696
MIXE			60059	35128	6609	101796
MIXTECO	2057	9686	122127	44980	131900	310750
NAHUATL		26653	579719	228306	142893	977571
OTOMI	38549	48804	16505		38257	142115
PAME DEL SUR				4168		4168
POPOLUCA					18822	18822
PUREPECHA		25732	14170			39902
TARAHUMARA			6998	24966		31964
TEPEHUA				5892		5892
TEPEHUAN				15989		15989
TLAPANECO					71806	71806
TOJOLABAL				50556		50556
TOTONACA		3219	43423	44097	57262	148001
TRIQUI					1686	1686
TZELTAL			15956	175436	111464	302856
TZOTZIL			18860	61095	146287	226242
ZAPOTECO	99020	45640	91144	67723	13173	316700
ZAPOTECO SUREÑO				6963	1820	8783
ZAPOTECO VALLISTA		2112				2112
ZOQUE			8487	4869	16249	29605
Total	179099	372558	1539285	1063640	880444	4035026

Cuando se trabaja con unidades de observación de Municipios, es deseable encontrar los valores correspondientes a unidades superiores como serían los Estados, las Etnias o alguna otra útil al estudio. El recurso más utilizado es hacer el mismo camino de cálculos pero para los valores de Estados, Etnias, etc. Es decir,repetir el proceso encontrando la estructura estatal o étnica y hacer a un lado lo obtenido. Se ha pensado que en pro de conservar el trabajo realizado, los valores del índice a nivel estatal o étnico han de ser obtenidos a través de los valores municipales. La propuesta es manejar un ponderador municipal para combinarlos y obtener el estatal o étnico. El ponderador está determinado por la proporción de población dentro del municipio en relación al Estado o a la Etnia, según sea el caso. El Cuadro dos muestra un resumen de los resultados obtenidos por Etnia.

CASOS CRITICOS

Por último, en virtud de la necesidad de un criterio inicial de selección de los casos críticos que deben priorizarse para ser atendidos, dentro del grupo de muy alta marginación se propone un algoritmo que involucra el tamaño de población con el índice encontrado. Con este algoritmo, los primeros cinco municipios que debieran recibir atención son: Metlatonoc de Guerrero, Chamula y Chilón, ambos de Chiapas, San Martín Peras y San Simón Zahuatlán, ambos de Oaxaca.

SINTESIS

El trabajo maneja cinco propuestas:

- construcción de un índice
- cómo agrupar en semejantes y diferentes (cinco grupos)
- cómo caracterizar a los Municipios analizando las constantes que forman el índice (dos etapas: índice y grupos)
- cómo encontrar valores índice de unidades de observación superiores (estados, etnias, etnia-estado)
- algoritmo de criterio inicial para seleccionar casos críticos

MAXIMAL ATTRACTION FOR MAXIMA AND MINIMA OF SAMPLES OF RANDOM SIZE

José A. Villaseñor
Colegio de Postgraduados, CEC
Carr. México-Texcoco Km. 35.5
Montecillo, México CP 56230

and

Barry C. Arnold
Department of Statistics
University of California
Riverside, CA92521. U.S.A.

Abstract

Consider the maximum (minimum) of a random number N_k of iid random variables with common distribution F_X. The tail behavior of its distribution is studied in terms of that of F_X and the local variation at 1 of the common generating probability function $P_N(s)$ of the N_k's. The results are applied to domains of maximal attraction.

AMS 1970 *Subject classifications*. Primary 60F05; secondary 60E05
Key words and phrases. Tail behavior, regular variation, domains of maximal attraction.

1. Introduction

Let $\{N_k\}$ be a sequence of independent identically distributed (iid) positive integer valued random variables (rv's) and consider a doubly infinite array of iid rv's $\{X_{kj}\}$, assumed independent of $\{N_k\}$. Define two sequences of iid rv's $\{Y_k\}$ and $\{Z_k\}$ by

$$(1) \qquad Y_k = \min_{j \le N_k} X_{kj} ,$$

$$(2) \qquad Z_k = \max_{j \le N_k} X_{kj} .$$

Let F_X, F_Y and F_Z be the common distribution function (df) of the X_{kj}'s, Y_k's and Z_k's respectively; also let P_N be the common generating probability function of the N_k's.

In this note we discuss the right hand side (rhs) tail behavior of F_Y and F_Z in terms of the rhs tail behavior of F_X and the local variation at 1 of P_N.

The right end point $\omega(F)$ of a df F plays an important rolle in a discussion such as this, it is defined as the point $x_o \equiv \omega(F)$ such that $F(x_o+\varepsilon) = 1 > F(x_o-\varepsilon)$ for any $\varepsilon>0$. This kind of discussion will also involve functions of regular variation which can be found in Feller (1971); an alternative reference which is very much inclined to the applications of the results we are seeking is De Haan (1970).

Random variables of the form (1) and (2) arise naturally in certain stochastic exceedance models (see e.g. Todorovic (1979)).

2. Tail Behavior of the Distributions of Minima and Maxima

The rhs tail behavior of the df of the minimum of samples of random size can be described without any conditions on the distribution of the N_k's (refer to equation (1)).

In the following let $p_i = P(N=i)$ and set $\overline{F}=1-F$ for a df F.

Theorem 1: Define $m = \min_{j \geq 1}\{j: p_j > 0\}$. Then as $y \uparrow \omega(F_Y)$,

$$\bar{F}_Y(y) \sim p_m\{\bar{F}_X(y)\}^m \quad . \tag{3}$$

Proof: First observe that $\bar{F}_Y(y) = P_N(\bar{F}_X(y))$. Hence, we may write,

$$\bar{F}_Y(y) = \{\bar{F}_X(y)\}^m\{p_m + \sum_{j=m+1}^{\infty} p_j[\bar{F}_X(y)]^{j-m}\} \quad .$$

Thus, using the Dominated Convergence Theorem, we conclude that (3) holds since $\omega(F_Y) = \omega(F_X)$.

The rhs tail behavior of the df of the maximum of samples of random size can be described under two cases; that is, when either $E\{N\} < \infty$ or $= \infty$.

Theorem 2: Suppose that $E\{N\} < \infty$. Then as $z \uparrow \omega(F_Z)$,

$$\bar{F}_Z(z) \sim E\{N\} \cdot \bar{F}_X(z) \quad . \tag{4}$$

Proof: First observe that $F_Z(z) = 1 - P_N(F_X(z))$. If we let $Q_N(s) = \sum_{j=0}^{\infty} P(N > j)s^j$ then $\{1-P_N(s)\}/(1-s) = Q_N(s)$ (cf. Feller (1968)). Hence

$$\bar{F}_Z(z)/\bar{F}_X(z) = \{1-P_N(F_X(z))\}/\bar{F}_X(z) = \sum_{j=0}^{\infty} P(N > j)\{F_X(z)\}^j \quad .$$

Thus, using the Dominated Convergence Theorem, we conclude that (4) holds, since $\omega(F_Z) = \omega(F_X)$.

When $E\{N\} = \infty$, a rather general condition on N is assuming P_N to be a function of regular variation at 1.

Theorem 3: Suppose that $1-P_N(1-x^{-1}) = x^{-\gamma}L(x)$ for $0 \leq \gamma \leq 1$ where L is a continuous function of slow variation at infinity. Then

$$\bar{F}_Z(z) = \{\bar{F}_X(z)\}^{\gamma}L(\{\bar{F}_X(z)\}^{-1}) \quad . \tag{5}$$

Proof: It follows by substitution.

Notice that in Theorem 3 the condition on P_N cannot hold with a value of $\gamma > 1$. In fact, since $Q_N(s) \uparrow$ as $s \uparrow$,

$$\frac{1-P_N(1-t^{-1}v^{-1})}{1-P_N(1-t^{-1})} = \frac{v^{-1}Q_N(1-t^{-1}v^{-1})}{Q_N(1-t^{-1})} \geq v^{-1} > v^{-\gamma}$$

for all $t > 0$, $v > 1$ and $\gamma > 1$. Then $1-P_N(1-x^{-1})$ cannot be regularly varying of order $\gamma > 1$.

3. Domains of Maximal Attraction for Y and Z

It is well known that there are only three possible types of limiting distributions for sample maxima (cf. Galambos (1978)). These are

$$\Phi_\alpha(x) = \exp\{-x^{-\alpha}\}, \quad x > 0 , \quad \alpha > 0$$
$$= 0 , \quad x \leq 0$$

$$\Psi_\alpha(x) = \exp\{-(-x)^{\alpha}\}, \quad x < 0 , \quad \alpha > 0$$
$$= 1 , \quad x \geq 0$$

$$\Lambda(x) = \exp\{-e^{-x}\}, \quad -\infty < x < +\infty .$$

These df's are called extreme distributions. If G is an extreme df, we will write $X \in \mathcal{D}(G)$ if there exist norming constants $\alpha_n > 0$ and β_n such that

$$\lim_{n\to\infty} P(\alpha_n \max_{i\leq n} X_i + \beta_n \leq x) = G(x) \quad \text{for all } x,$$

where the X_i's are iid copies of X. In this situation we say that X is maximally attracted to G.

Under this approach, we seek conditions on the distribution of the N_k's which enable the identification of the extreme df to which Y_k and Z_k (as defined in (1) and (2)) are maximally attracted, assuming we know the extreme df to which the X_{kj}'s are maximally attracted.

The following results will be very useful. From here on G will denote some extreme df.

Theorem 4: (De Haan (1970)). Let V have df F_V. $V \in \mathcal{D}(G)$ if and only if there exist functions $a: (-\infty,+\infty) \to (0,+\infty)$ and $b: (-\infty,+\infty) \to (-\infty,+\infty)$ such that for any fixed real x with $0 < G(x) < 1$ as $t \uparrow \omega(F_V)$

$$\lim \bar{F}_V(a(t)x+b(t))/\bar{F}_V(t) = -\log G(x)$$

Theorem 5: (Resnick (1971)). Let X_1 and X_2 have df's F_1 and F_2 with $\omega(F_1) = \omega(F_2)$. Suppose that $X_1 \in \mathcal{D}(G)$. Then $X_2 \in \mathcal{D}(G)$ if and only if as $x \uparrow \omega(F_1)$

$$\lim \bar{F}_1(x)/\bar{F}_2(x) = \ell, \quad 0 < \ell < \infty .$$

Notice that for a constant c, $G_c(x) = \exp\{-(-\log G(x))^c\}$ is an extreme df.

Theorem 6: $X \in \mathcal{D}(G)$ if and only if $Y \in \mathcal{D}(G_m)$ with m as in (3).

Proof: Let W have df $F_W(w) = 1-\{1-F_X(w)\}^m$. Note that by Theorems 1 and 5, $W \in \mathcal{D}(G_m) \iff Y \in \mathcal{D}(G_m)$ since $\omega(F_W) = \omega(F_Y)$. Assume $X \in \mathcal{D}(G)$. Then by Theorem 4 there exist functions $a(t) > 0$ and $b(t)$ such that as $t \uparrow \omega(F_X)$

$$\lim \bar{F}_X(a(t)x+b(t))/\bar{F}_X(t) = -\log G(x) \tag{6}$$

for any fixed real x with $0 < G(x) < 1$. Hence since $\omega(F_X) = \omega(F_W)$, as $t \uparrow \omega(F_W)$

$$\lim \bar{F}_W(a(t)x+b(t))/\bar{F}_W(t) = \{-\log G(x)\}^m .$$

Therefore by Theorem 4, $W \in \mathcal{D}(G_m)$; hence $Y \in \mathcal{D}(G_m)$. The proof of the converse is obtained following the steps backwards in the above argument.

Theorem 7: Suppose that $1-P_N(1-x^{-1}) = x^{-\gamma}L(x)$ for $0 < \gamma \leq 1$ where L(x) is a continuous function of slow variation at infinity. Then $X \in \mathcal{D}(G)$ if and only if $Z \in \mathcal{D}(G_\gamma)$

Proof: Sufficiency. By Theorem 4, (6) holds. Hence, since $\omega(F_Z) = \omega(F_X)$, by Theorem 3 and letting $x_t = a(t)x+b(t)$ we have that as $t \uparrow \omega(F_Z)$

$$\lim \bar{F}_Z(x_t)/\bar{F}_Z(t) = \lim\{\bar{F}_X(x_t)/\bar{F}_X(t)\}^\gamma L(\{\bar{F}_X(x_t)\}^{-1})/L(\{\bar{F}_X(t)\}^{-1})$$

$$= \{-\log G(x)\}^\gamma ,$$

since as $t \uparrow \omega(F_Z)$, $L(\{\bar{F}_X(x_t)\}^{-1})/L(\{\bar{F}_X(t)\}^{-1}) =$

$= L(\{\bar{F}_X(x_t)/\bar{F}_X(t)\}^{-1}\{\bar{F}_X(t)\}^{-1})/L(\{\bar{F}_X(t)\}^{-1}) \to 1$. Therefore, by Theorem 4, $Z \in \mathcal{D}(G_\gamma)$.

Necessity. Let $\psi(x) = 1-P_N(1-x^{-1})$, then $\psi^*(x) = \psi^{-1}(1/x)$ is a continuous function of regular variation at infinity of order $1/\gamma$. Hence $\psi^*(x) = x^{1/\gamma}L^*(x)$ where L^* is a continuous function of slow variation at infinity. Therefore, $\bar{F}_Z(z) = 1-P_N(F_X(z)) = \psi(\{\bar{F}_X(z)\}^{-1}) \Longleftrightarrow \{\bar{F}_X(z)\}^{-1} = \psi^*(\{\bar{F}_Z(z)\}^{-1})$. That is

$$\bar{F}_X(z) = \{\bar{F}_Z(z)\}^{1/\gamma}/L^*(\{\bar{F}_Z(z)\}^{-1}). \tag{7}$$

Thus by a reasoning similar to that used in the proof of the sufficiency part, replacing Theorem 3 by equation (7) we conclude that $X \in \mathcal{D}(G)$.

REFERENCES

De Haan, L. (1970). On regular variation and its application to the convergence of sample extremes. Mathemathical Centre Tracts, Vol. 32, Amsterdam.

Feller, W. (1968). An introduction to probability theory and its applications, Vol. I, Third Edition, Wiley, New York.

Feller, W. (1971). An introduction to probability theory and its applications. Vol. II, Second Edition, Wiley, New York.

Galambos, J. (1978). The asymptotic theory of extreme order statistics. Wiley, New York.

Resnick, S.I. (1971). Tail equivalence and applications. J. Applied Probability, 8: 136-156.

Todorovic, P. (1979). A probabilistic approach to analysis and prediction of floods. Proceedings 42nd Session ISI, Manila.

Esta publicación consta de 500 ejemplares y se terminó de
imprimir en el mes de septiembre de 1994 en los talleres gráficos del
Instituto Nacional de Estadística, Geografía e Informática
Av. Héroe de Nacozari Núm. 2301 Sur, Acceso 11, P.B.
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.
México