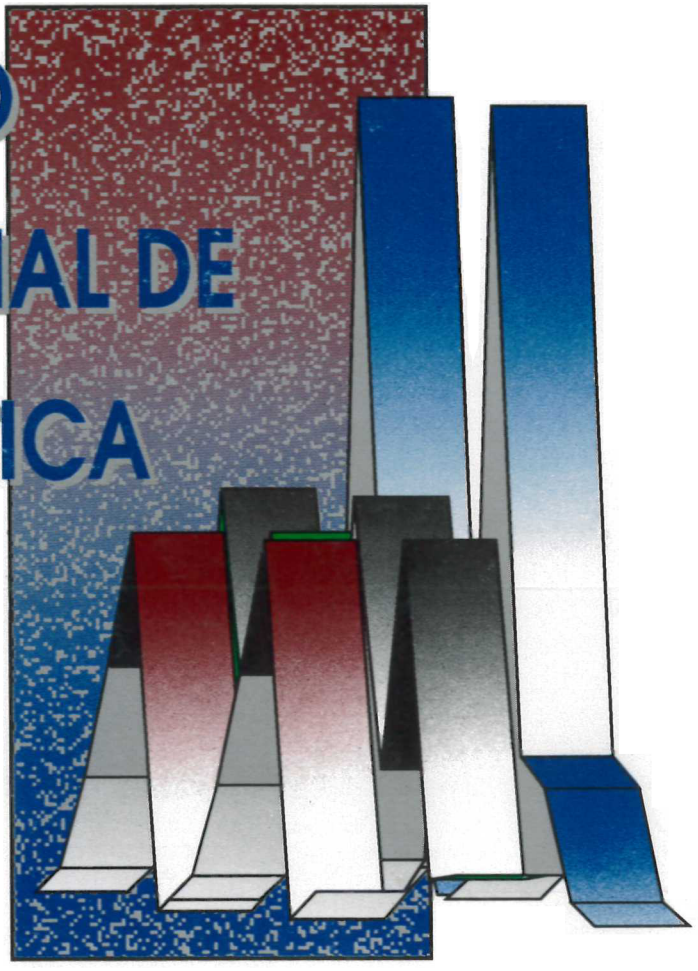
MEMORIA DEL IX FORO NACIONAL DE ESTADISTICA



UNIVERSIDAD AUTONOMA DE COAHUILA

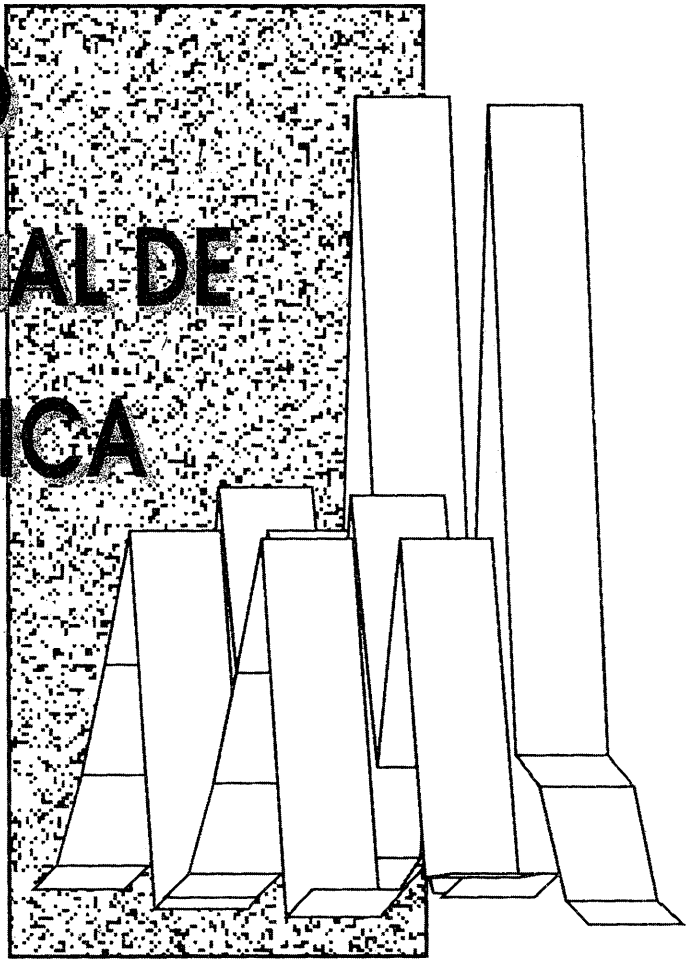
SALTILLO, COAHUILA, DEL 26 AL 30 DE SEPTIEMBRE DE 1994

INEGI

INSTITUTO NACIONAL DE ESTADISTICA GEOGRAFIA E INFORMATICA

AME

MEMORIA DEL IX FORO NACIONAL DE ESTADISTICA



UNIVERSIDAD AUTONOMA DE COAHUILA

SALTILLO, COAHUILA, DEL 26 AL 30 DE SEPTIEMBRE DE 1994

INEGI

INSTITUTO NACIONAL DE ESTADISTICA GEOGRAFIA E INFORMATICA

AME

Memoria del IX Foro Nacional de Estadística

Impreso en México
ISBN 970-13-0799-2

Esta publicación consta de 300 ejemplares y se terminó de imprimir en el mes de septiembre de 1995 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

Presentación

Durante la semana del 26 al 30 de septiembre de 1994 tuvo lugar, en la ciudad de Saltillo, Coahuila, el IX Foro Nacional de Estadística, teniendo como institución sede a la Universidad Autónoma de Coahuila. Los trabajos que se presentaron en dicho evento abordaron temas muy diversos que reflejan lo que se está haciendo, en diferentes lugares del país, con respecto al desarrollo y aplicación de métodos estadísticos.

En estas Memorias se recopilan, en versión resumida, algunos de los trabajos que se presentaron ya sea como conferencia invitada, o como contribución libre. Es conveniente mencionar que la convocatoria para reunir trabajos para estas Memorias solicitó explícitamente resúmenes extendidos, y no se realizó un arbitraje estricto. El orden en que aparecen los mismos corresponde al apellido del primer autor.

El Comité Editorial desea agradecer, muy especialmente, a todas las personas que colaboraron con sus trabajos escritos. También agradece a Alejandra Macías por la labor tipográfica, así como a Miguel Ancona, Jorge Argaez, René Castro, Sergio Nava, y María Guadalupe Russell por su revisión.

Tambien agradecemos a las diferentes instituciones que brindaron su apoyo e hicieron posible la realización del Foro y la publicación de estas Memorias: la Asociación Mexicana de Estadística, la Universidad Autónoma de Coahuila, el Centro de Investigación en Matemáticas, el Instituto Tecnológico Autónomo de México, y el Instituto Nacional de Estadística, Geografía, e Informática.

El Comité Editorial
Alejandro Alegría
Miguel Nakamura
Enrique Villa
Agosto de 1995

Contenido

Teoría de credibilidad y sus aplicaciones

Alejandro Alegría Hernández
Instituto Tecnológico Autónomo de México

Introducción

Uno de los problemas mas importantes que se puede plantear una compañía de seguros es determinar la prima que deben pagar sus asegurados. Es un hecho reconocido que al asignar primas lo mejor que se puede hacer es buscar un valor que se encuentre entre lo que nos dice la experiencia particular del asegurado y el comportamiento de todo el grupo de asegurados. Esto se puede justificar de varias formas. Es evidente que una prima de cero, para un asegurado que no ha tenido reclamos en algún período de tiempo, no es un valor factible por muchas razones. En el otro extremo, si una persona tuvo un período muy desfavorable (es decir, presentó muchos reclamos), tampoco es una política adecuada el castigar a esta persona, con una prima más alta, tomando en cuenta solamente su experiencia particular. El problema es entonces determinar en qué forma se debe ponderar la información que sobre un asegurado se tiene y la información de todo el grupo de asegurados. Esto es, muy someramente, de lo que trata la Teoría de Credibilidad.

El problema antes mencionado fue inicialmente tratado por Whitney (1918), y más recientemente por Bühlmann (1967), Goovaerts and Hoogstad (1987), Klugman (1992), entre otros. En el presente trabajo se exponen los aspectos fundamentales de la Teoría de Credibilidad, y se hace notar que los métodos bayesianos nos proporcionan la forma más natural de tratar el problema que nos ocupa.

En la siguiente sección se plantea de manera más formal el problema de Credibilidad y se establece la conveniencia de usar métodos bayesianos. En la sección 3 se presentan algunos modelos y resultados asociados a éstos. Un ejemplo se presenta en la cuarta sección, y finalmente, se dan algunas conclusiones.

El problema de credibilidad

Con el paso de los años, la palabra Credibilidad ha tenido diferentes significados dependiendo de las restricciones que se deseen considerar. No obstante, lo que tienen en común todos los métodos que tratan de asignar una tarifa adecuada (con la cual el asegurado minimiza las consecuencias negativas de algún percance, y al mismo tiempo, el asegurador puede cubrir tales contingencias), consiste en el siguiente problema de estimación: "Basándose en observaciones de elementos de algún grupo de interés y en observaciones de miembros de otros grupos, se desea obtener la distribución de los reclamos futuros para los elementos del grupo de interés". La variable que representa el monto de los reclamos futuros es claramente una variable aleatoria, e influyen sobre ella una serie de factores, como son las características individuales del asegurado, las características del grupo al cual pertenece el asegurado, factores externos (casi siempre de tipo económico), y la naturaleza aleatoria del evento que puede causar el reclamo.

Se debe mencionar que en este trabajo no se toman en cuenta variables de tipo económico, como por ejemplo la inflación, así que, el efecto de tales variables se estudiaría después de haber realizado el análisis de credibilidad. Consideremos entonces k grupos de personas (asegurados) y t observaciones de cada uno de estos grupos. Sea x_{ij} la j-ésima observación del grupo i. También supongamos que la distribución de probabilidad de x_{ij} depende de un parámetro θ_i, y que dados los parámetros $\theta_1, \theta_2, \ldots, \theta_k$, las observaciones son independientes. En la práctica, las observaciones podrían corresponder a montos reclamados o bien al número de reclamos. Lo que nos interesa conocer es el comportamiento de la variable $x_{i,t+1}$, la siguiente observación en el grupo i. En particular es interesante el poder determinar el valor esperado de $x_{i,t+1}$, es decir,

$$E(x_{i,t+1}/\theta_i) = m(\theta_i).$$

Un estimador natural para $m(\theta_i)$ es $\overline{x}_i = \sum_{j=1}^{t} x_{ij}/t$.

Sea $\delta(\mathbf{x})$ un estimador de $m(\theta_i)$, en donde x denota el total de las observaciones. El error cuadrático medio de este estimador es

$$E_{\mathbf{x}|\theta}\{[\delta(\mathbf{x}) - m(\theta_i)]^2\},$$

cuyo valor depende del parámetro θ_i. Si se cuenta con una distribución para los parámetros $\theta_1, \theta_2, \ldots, \theta_k$, es posible aplicar procedimientos bayesianos, en donde el problema a resolver sera encontrar $\delta(\mathbf{x})$ tal que

$$L(\delta(\mathbf{x})) = E_\theta E_{\mathbf{x}|\theta}\{[\delta(\mathbf{x}) - m(\theta_i)]^2\},$$

sea mínimo. La solución a este problema se obtiene fácilmente y está dáda por

$$\delta(\mathbf{x}) = E_{\theta|\mathbf{x}}[m(\theta_i) \mid \mathbf{x}],$$

que es el valor esperado a posteriori de la cantidad de interés.

Consideremos ahora el siguiente estimador $\delta_C(\mathbf{x})$, el cual establece un compromiso entre $\overline{x}_i$ y $m = E_\theta[m(\theta_i)]$,

$$\delta_c(\mathbf{x}) = Z\overline{x_i} + (1 - Z)\, m.$$

La expresión correspondiente a $L(\delta_c(\mathbf{x}))$ resulta ser

$$L(\delta_c(\mathbf{x})) = sZ^2/t + (1 - Z)^2 v,$$

en donde

$$s = E_\theta[s(\theta_i)] = E_\theta[\mathrm{Var}(x_{ij}|\theta_i)],\ v = \mathrm{Var}_\theta[m(\theta_i)].$$

Al minimizar $L(\delta_C(\mathbf{x}))$ con respecto a Z, se obtiene que el mínimo se alcanza cuando Z es igual a

$$Z = \frac{v}{v + (s/t)}.$$

Con este valor de Z resulta que

$$L(\delta_C(\mathbf{x})) = \frac{sv/t}{v + (s/t)}.$$

Como v y s son no negativos, $L(\delta_C(\mathbf{x}))$ es menor que el correspondiente valor para $\delta(\mathbf{x}) = \overline{x}_i$, el cual es igual a s/t. Apliquemos los resultados anteriores al caso en donde las observaciones se obtienen de un proceso Poisson con parámetro θ y a su vez el parámetro θ sigue una distribución Gamma con parámetros α y β. En la notación utilizada esto significa que $m(\theta) = \theta, m = E(\theta) = \alpha\beta, v = Var(\theta) = \alpha\beta^2, s(\theta) = \theta$ y $s = E(\theta) = \alpha\beta$. El valor óptimo de Z está dado por

$$Z = \alpha\beta^2/(\alpha\beta^2 + \alpha\beta/t) = \beta t/(\beta t + 1),$$

por lo que el estimador $\delta_C(\mathbf{x})$ estaría dado por,

$$\delta_C(\mathbf{x}) = \frac{\beta t}{\beta t + 1}\overline{\mathbf{x}}_i + \frac{1}{\beta t + 1}\alpha\beta.$$

Usando la metodología Bayesiana en la situación anterior resulta que la distribución a posteriori, $p^*(\theta|\mathbf{x})$, del parámetro θ está dada por

$$p^*(\theta|\mathbf{x}) \propto \frac{e^{-t\theta}\theta^{\sum x_i}}{\prod x_i!} \propto \theta^{\alpha+\sum x_i-1}e^{-(t+1/\beta)\theta},$$

así que, dado $\mathbf{x}$, θ sigue una distribución Gamma con parámetros $\alpha + \sum x_i$ y $\beta/(1 + t\beta)$. Si la función de pérdida para estimar el parámetro θ es cuadrática, el estimador óptimo de θ resultara ser el valor esperado de la distribución a posteriori de θ, es decir, $(\alpha + \sum x_i)\beta/(1 + t\beta)$, el cual puede reescribirse como

$$\frac{\beta t}{\beta t + 1}\overline{x}_i + \frac{1}{\beta t + 1}\alpha\beta.$$

Este último es uno de los varios casos en donde un estimador bayesiano (con pérdida cuadrática) y el estimador de credibilidad, $\delta_C(\mathbf{x})$, coinciden.

Desde el punto de vista bayesiano, lo que nos interesa es encontrar la distribución predictiva de la variable de interés, es decir, $p(x)$. Esta distribución está dada por

$$p(x) = \int f(x \mid \theta)p^*(\theta \mid \mathbf{x})\, d\theta$$

y la predicción del valor futuro de x, usando pérdida cuadrática, resulta ser $E(X)$. Este valor esperado se puede expresar de la siguiente forma,

$$E(X) = E_\theta E_x\{x|\theta\}.$$

Al ser $E_x\{x|\theta\}$ una función de θ, el problema que se tiene es entonces encontrar el valor esperado, a priori o a posteriori, de alguna función de θ.

Modelos jerárquicos

Los modelos presentados en la sección anterior resultan ser muy simples para representar adecuadamente una realidad. Es por esta razón que ahora se presentarán los llamados modelos jerárquicos, con los cuales se pretende lograr una mayor generalidad al permitir

que los parámetro que definen la distribución a priori, también sean desconocidos, y por lo tanto habrá que asignar a éstos una distribución inicial. El modelo jerárquico considera las siguientes distribuciones,

$$f(x|\theta, F), \qquad p_1(\theta|\mu, G), \qquad p_2(\mu, F, G).$$

En el primer nivel se indica cómo la distribución de la variable que se observa depende del parámetro de interés (θ), y de otro parámetro de ruido (F). El segundo nivel indica cómo el parámetro de interés vara en la población. Finalmente el tercer nivel corresponde al comportamiento de todos los parámetros de ruido.

Es usual imponer ciertas restricciones en el modelo anterior, a saber, normalidad y linealidad en los parámetros. Con lo anterior se tiene la siguiente especificación,

$$x|\theta, F \sim N(A\theta, F), \qquad \theta|\mu, G \sim N(B\mu, G), \qquad p_2(\mu, F, G).$$

Este tipo de modelos fue originalmente introducido por Lindley y Smith (1972), bajo el nombre de Modelos Lineales Jerárquicos, y han sido aplicados con éxito en varias situaciones.

Un caso particular del modelo anterior, el modelo con un criterio de clasificación, corresponde al analizado por Bühlmann y Straub (1972) y también por Meyers (1984). En el primer nivel se cuenta con n_i observaciones en cada uno de k grupos. Sea x_{ij} la observación j-ésima correspondiente al grupo i. Entonces,

$$x_{ij}|\theta_i, \sigma^2 \sim N(\theta_i, \sigma^2/P_{ij}) \ , \ i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, n,$$

en donde las $N = \sum n_i$ observaciones son condicionalmente independientes dados $\theta_1, \theta_2, \ldots,$ θ_k y σ^2. Las P_{ij} son valores conocidos que son proporcionales a la exposición que produjo la observación. Sea ahora x el vector de observaciones $x = (x_{11}, x_{12}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{kn_k})$. La matriz A en este caso está dada por

$$A = \begin{bmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \cdots & \cdots & \cdots & \cdots \\ 0_{n_k} & 0_{n_k} & \cdots & 1_{n_k} \end{bmatrix}$$

en donde 1_m es un vector $m \times 1$ de unos y 0_m es un vector $m \times 1$ de ceros. La matriz de covarianzas F es diagonal con elementos σ^2 / P_{ij} en la diagonal.

El segundo nivel está dado por

$$\theta_i|\mu, \tau^2 \sim N(\mu, \tau^2), \ i = 1, 2, \ldots, k$$

donde $\theta_1, \theta_2, \ldots, \theta_k$ son independientes dados μ y τ^2. En términos del modelo lineal jerárquico basta tomar $B = 1_k$ y $F = \tau^2 I_k$, donde I_k es la matriz idéntica de dimensión k.

Lo que hace el modelo anterior es que a cada uno de los k grupos se les ha asignado un valor que nos indica qué tan propenso es el grupo i a producir pérdida, este valor es θ_i.

Usualmente, el tercer nivel queda especificado por alguna distribución no informativa.

Otros modelos conocidos también son caso particular del modelo lineal jerárquico. Tal es el caso del modelo con dos criterios de clasificación, regresión lineal, Filtro de Kalman, series de tiempo. El problema de graduación, muy conocido en el ámbito actuarial, también se puede plantear en los términos de un modelo jerárquico.

Ejemplo

A manera de ejemplo se presentan los siguientes datos correspondientes a los montos reclamados por 50 personas. En este caso se tiene $k = 10$ grupos y $n_i = 5$ observaciones en cada grupo. Las exposiciones son $P_{ij} = 1$ para todas las observaciones. Esto último facilita todos los cálculos. Se suponen distribuciones normales en los primeros niveles con varianza en el primer nivel igual a $\sigma^2 = 100$ y varianza en el segundo nivel igual a $\tau^2 = 225$. El valor esperado de θ_i es igual a $\mu = 100$.

i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}
1	124.93	110.67	106.93	104.05	101.60
2	97.16	89.28	102.88	111.10	105.78
3	103.57	86.82	92.49	87.99	83.8
4	119.53	125.92	98.05	117.57	94.17
5	95.68	110.43	83.59	110.54	101.51
6	102.04	93.60	106.12	98.7	108.55
7	112.71	101.64	106.50	111.71	100.59
8	119.16	111.54	127.24	115.02	115.06
9	81.71	90.45	94.51	84.74	88.95
10	102.51	123.81	113.83	94.97	90.51

Con estos datos resulta que

$$E(x_i|\mathbf{x}, Z_i) = Z_i\overline{x}_i + (1 - Z_i)103.1842$$

en donde $Z_i = 5/(5 + \delta)$ y $\delta = \sigma^2/\tau^2$.

El valor que nos interesa es $E(x_i|\mathbf{x})$, el cual se puede expresar como

$$E(x_i|\mathbf{x}) = E_{Z_i}E(x_i|\mathbf{x}, Z_i) = E(Z_i)\overline{x} + (1 - E(Z_i))103.1842.$$

Considerando la transformación $\alpha = \sigma^2, \delta = \sigma^2/\tau^2$, se obtiene la distribución final $p(\alpha, \delta \mid \mathbf{x})$, de donde la distribución marginal $p(\delta \mid \mathbf{x})$ está dada por,

$$p(\delta \mid \mathbf{x}) = \int_0^\infty p(\alpha, \delta \mid \mathbf{x})\, d\alpha.$$

Para el ejemplo que nos ocupa,

$$p(\alpha, \delta \mid \mathbf{x}) \propto \alpha^{-23.5}\delta^{2.5}(5 + \delta)^{-4.5} \exp\left\{\frac{-3382.82}{2\alpha} - \frac{5(737.703)}{2\alpha(5 + \delta)}\right\},$$

$$p(\delta \mid \mathbf{x}) \propto \delta^{2.5}(5 + \delta)^{-4.5}\left[1691.41 + \frac{1844.2575\delta}{5 + \delta}\right]^{-22.5}\Gamma(22.5),$$

así que,

$$E(Z_i \mid \mathbf{x}) = \int_0^\infty 5(5 + \delta)^{-1} p(\delta \mid \mathbf{x})\, d\delta = 0.8218.$$

Finalmente, el valor esperado de x_i es

$$E(x_i|\mathbf{x}) = 0.8218\,\overline{x}_i + 0.1882\,(103.1842) = 0.8218\overline{x}_i + 18.387.$$

A continuación se presentan los valores esperados y las varianzas a posteriori de x_i para cada uno de los diez grupos considerados.

i	1	2	3	4	5	6	7	8	9	10
$E(x_i \mid \mathbf{x})$	108.480	101.586	93.117	109.646	100.855	102.048	106.016	115.034	90.271	104.780
$V(x_i \mid \mathbf{x})$	16.075	15.646	17.304	16.304	15.694	15.624	15.737	17.961	18.402	15.645

Conclusiones

La solución que originalmente se planteó, para la determinación de la prima que un asegurado debe pagar por cierto período de tiempo, se ha visto substancialmente enriquecida con la aplicación de los métodos bayesianos. La distribución predictiva que se obtiene, después de incorporar la información disponible, nos permite tomar la decisión de cuál debe ser la prima adecuada. No obstante, varios aspectos quedan aún por investigar. En primer lugar, el utilizar un valor esperado, a posteriori, como el estimador óptimo de la prima a pagar, supone una función de pérdida cuadrática. Valdrá la pena investigar qué implicaciones actuariales tiene el usar otro tipo de funciones de pérdida. Otro aspecto que últimamente ha cobrado importancia, es estudiar el efecto que tienen sobre los estimadores, observaciones consideradas discrepantes (outliers). Autores como Kunsh (1992) y Gisler y Reinhard (1993), han propuesto el uso de estimadores robustos. En el problema de Credibilidad también es importante incluir, desde un principio, variables de tipo económico. En este caso, los modelos econométricos pueden ser de gran utilidad.

Referencias

Bühlmann, H. (1967) "Experience Rating and Credibility", *Astin Bulletin*, **4**, 199–207.

Bühlmann, H. and Straub, E. (1972) "Credibility for Loss Ratios", *ARCH*, **1972.2.**

Gisler, A. and Reinhard, P. (1993) "Robust Credibility", *Astin Bulletin*, **23**, 1.

Goovaerst, M. and Hoogstad, W. (1987) "Credibility Theory", *Survey of Actuarial Studies*, **4**.

Klugman, S. A. (1972) *Bayesian Statistics in Actuarial Science*, Kluwer Academic Publishers.

Kunsh, H. R. (1992) "Robust Methods for Credibility", *Astin Bulletin*, **22**, 1.

Lindley, D. V. and Smith, A. (1972) "Bayes Estimates for the Linear Model", *JRSS*, **34**, 1–41.

Meyers, G. (1984) "Empirical Bayesian Credibility for Workers' Compensation", *Proceedings of the Casualty Actuarial Society*, **71**, 96–121.

Whitney, A. (1918) "The Theory of Experience Rating", *Proceedings of the Casualty Actuarial Society*,**71**, 96–121.

Una aplicación de la estadística en el área de genética forestal

ALEJANDRO ARMAS
Laboratorio de Investigación y Asesoría Estadística, Facultad de Estadística, Universidad Veracruzana

Introducción

Se presenta un análisis de la variabilidad en el crecimiento de las progenies de 43 familias ubicadas en un área semillera de Pinus Patula. Este procedimiento se basa en la técnica de Componentes de la Varianza, y usa como datos los registros mensuales de diámetro y altura.

En México, son contados los estudios de progenies realizados a las áreas semilleras, por lo que el grado de mejoramiento genético obtenido de éstas es poco conocido.

Actualmente se han diseñado programas de mejoramiento genético de bosques, en los cuales la estadística, a través de su metodología, ha permitido obtener resultados satisfactorios en beneficio del área forestal. Componentes de la varianza, análisis de varianza, análisis de regresión, correlación lineal, y algunas técnicas multivariadas forman parte de la metodología estadística que se utiliza en la solución de problemas en esta área.

En este estudio se persigue analizar los datos de crecimiento de 43 familias de Pinus Patula de un área semillera ubicada en el ejido Ingenio El Rosario, Municipio de Xico, Ver., los cuales fueron obtenidos por la bióloga Virginia Rebolledo. Se pretende evaluar y comparar la variabilidad y crecimiento de la progenie de 43 familias.

Materiales y métodos

La semilla empleada en está investigación fue colectada de 43 árboles ubicados dentro del área semillera en estudio, de la siguiente manera: 16 árboles de la zona central, 9 árboles de la zona de protección y 18 árboles de la zona periférica.

A cada árbol se le denomina familia, por lo que en total se tienen 43 familias. En cada zona los árboles seleccionados fueron elegidos por presentar en ese momento una buena producción de conos de donde se obtuvieron las semillas. La colecta de estos conos se realizó con la ayuda de equipo adecuado para trepar árboles y con garrochas para cortar los conos. Posteriormente se procedió al beneficio de la semilla, esto abarca desde el colocar los conos al sol para que desprendan la semilla, recoger y cuidar de no revolver la semilla de cada árbol, hasta que finalmente se le quita el ala y se limpia de basuras; todo este proceso se realizó en el Banco Central de Germoplasma Forestal ubicado en Los Molinos, municipio de Perote, Ver.

Una vez obtenida la semilla se trasladó a las instalaciones del Centro de Genética Forestal en Jalapa Ver., para su evaluación.

Metodología estadística

Los datos fueron capturados en los paquetes estadísticos SOLO y SYSTAT, posteriormente se realizó un análisis preliminar, en el cual en primera instancia se obtuvieron gráficas de cajas y alambres, las cuales permitieron hacer un estudio visual y rápido de la forma de la distribución de los datos. Asimismo estas gráficas permitieron obtener información general sobre simetría y dispersión en las diferentes partes de la escala de los datos dividida por los cuartiles. Cabe mencionar que al hacer uso de estás gráficas se pudieron detectar algunos valores extremos, los cuales fueron estudiados con detalle. Posteriormente se llevó a cabo un análisis descriptivo, para lo cual se obtuvieron las medias y desviaciones estándar mensuales para cada familia evaluando la altura y el diámetro, a fin de elaborar una tabla de medias que mostrara el crecimiento mensual.

Con el objetivo de verificar la similaridad que existe entre grupos de familias en cuanto a sus características físicas de crecimiento como los son la altura y el diámetro, se realizaron varios análisis cluster o de agrupamiento, utilizando diferentes métodos, tales como el método del vecino más cercano, el método del vecino más lejano y el método del centroide, usando como distancia la euclidiana. El propósito general de este análisis fue encontrar el "agrupamiento natural" de objetos o individuos similares. Estos resultados no se incluyen aquí por razones de espacio, pero aparecen en Armas (1995).

Modelo de Componentes de la Varianza para el caso de una variable de clasificación

Se presenta como uno de los modelos más sencillos del modelo de regresión Lineal en dos niveles, y se utiliza en la etapa exploratoria inicial para estimar el coeficiente de correlación intraclase, y con esto saber si existe evidencia de heterogeneidad entre los grupos (Bryk y Raudenbush, 1992; Ojeda, 1993). Este modelo es, quizá, el modelo jerárquico más conocido y que más aplicaciones tiene asociadas. Podemos mencionar su gran utilidad en genética (Henderson, 1986), en fases de inferencia analítica en muestreo (Ojeda, 1992), en aplicaciones a control de calidad (Montgomery, 1992) y en estudios educativos (Aitkin y Logford, 1986).

En este modelo las ecuaciones en forma escalar se establecen como sigue:

Modelos al nivel 1 :

$$Y_{ij} = \beta_{oi} + \xi_{ij};\ i = 1, 2, \ldots, g;\ j = 1, 2, \ldots, n_g$$

Modelos al nivel 2 :

$$\beta_{oi} = \beta_o + U_i$$

Asociados a estos modelos se tienen las siguientes suposiciones:

$$\begin{array}{ll} E(\xi_{ij}) = 0 & \mathrm{Var}(\xi_{ij}) = \sigma^2 \\ \mathrm{Cov}(\xi_i, \xi_i) = 0 & \text{para } i \neq j \text{ y/o } j \neq j \\ E(U_i) = 0 & \mathrm{Var}(U_i) = \sigma_g^2 \\ \mathrm{Cov}(U_i, U_j) = 0 & \text{para } i \neq j \\ \mathrm{Cov}(\xi_i, U_i) = 0 & \text{para cualquiera } i, j. \end{array}$$

Para este modelo se toma como parámetros en el primer nivel, las medias dentro de las unidades correspondientes al nivel 2. Como parámetro del segundo nivel se toma la media general de todas las unidades.

Resultados

Se usó el modelo presentado, el cual se ajustó. La hipótesis a probar es:

$H_o : \sigma_g^2 = 0$ H_o : No hay variabilidad entre las 43 familias
$H_o : \sigma_g^2 > 0$ H_o : Sí hay variabilidad entre las 43 familias

Después se procedió a realizar el análisis de heterogeneidad entre las familias, el cual se llevó acabo a través del paquete estadístico ML3 (Proser et al., 1990), por medio del cual se obtuvieron los coeficientes de correlación intra-clase $\hat{p}$, que se define como:

$$\hat{p} = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}$$

donde

σ_g^2: Representa a la variabilidad que existe entre las familias.
σ_e^2: Representa a la variabilidad que existe entre los árboles.

Los resultados se presentan a continuación en las Tablas 1,2,3 y 4.

	Variable Diámetro				Variable Altura		
Mes	σ_g^2	$\hat{\sigma}_e^2$	$\hat{p}$	Mes	σ_g^2	$\hat{\sigma}_e^2$	$\hat{p}$
1	.39200	.64300	0.3787	1	.0421	.0833	0.3400
2	.48100	.54800	0.4674	2	.0963	.3569	0.2126
3	.00008	.00006	0.5714	3	.1570	.2470	0.3886
4	.00008	.00016	0.3333	4	.3510	.5870	0.2905
5	.00009	-00024	0.2727	5	.6230	1.039	0.3748
6	.00012	.00027	0.3077	6	1.035	1.774	0.3684
7	.00010	.00025	0.2857	7	1.327	2.181	0.3783
8	.00011	.00026	0.2973	8	1.706	2.303	0.4255
9	.00008	.00045	0.1509	9	2.997	3.414	0.4674

Tabla 1. Componentes de varianza y correlación intraclase estimados para las 43 familias.

Dada la teoría que justifica ésta técnica y tomando como base un criterio heurístico, si $\rho > .10$ se puede decir que al menos hay una familia que no es igual que las demás en el tiempo dado y para la variable dada. Así también el $\hat{\rho}$ es un estadístico que permite describir la importancia relativa de las diferencias entre familias con respecto a las diferencias entre individuos, en este caso en particular entre los pinos.

Cabe mencionar también que cuando la variabilidad entre familias es muy pequeña en relación a la variabilidad dentro de las familias el $\hat{\rho} \approx 0$. También se realizó este análisis para las zonas central, protección y periférica.

	Variable Diámetro				Variable Altura		
Mes	σ_g^2	$\hat{\sigma}_e^2$	$\hat{p}$	Mes	σ_g^2	$\hat{\sigma}_e^2$	$\hat{p}$
1	.00003	.00007	0.3000	1	.0318	.0759	0.2952
2	.00005	.00005	0.5000	2	.0823	.1340	0.3805
3	.00012	.00003	0.8000	3	.1650	.2850	0.3666
4	.00013	.00014	0.4814	4	.2410	.5030	0.3239
5	.00013	-00028	0.3171	5	.3770	.9730	0.2792
6	.00014	.00028	0.3555	6	.4100	1.697	0.1945
7	.00014	.00028	0.3333	7	.3800	2.146	0.1504
8	.00006	.00026	0.1875	8	.2780	2.369	0.1050
9	.00009	.00052	0.1475	9	.4610	3.493	0.1166

Tabla 2. Componentes de varianza y correlación intraclase estimados para las familias que delimitan la zona central.

	Variable Diámetro				Variable Altura		
Mes	σ_g^2	$\hat{\sigma}_e^2$	$\hat{p}$	Mes	σ_g^2	$\hat{\sigma}_e^2$	$\hat{p}$
1	.00005	.00007	0.4166	1	.03040	.08990	0.2527
2	.00006	.00006	0.5000	2	.04990	.12900	0.2789
3	.00004	.00006	0.4000	3	.11500	.22800	0.3353
4	.00004	.00017	0.1904	4	.28600	.72000	0.2843
5	.00009	-00023	0.2812	5	.28600	.72000	0.2843
6	.00009	.00026	0.2571	6	1.0070	1.9610	0.3393
7	.00008	.00023	0.2580	7	1.3770	2.3750	0.3670
8	.00008	.00026	0.2352	8	1.9000	2.4030	0.4415
9	.00008	.00040	0.1666	9	2.7820	3.5580	0.4388

Tabla 3. Componentes de varianza y correlación intraclase estimados para las familias que delimitan la zona de protección.

	Variable Diámetro				Variable Altura		
Mes	σ_g^2	$\hat{\sigma}_e^2$	$\hat{p}$	Mes	σ_g^2	$\hat{\sigma}_e^2$	$\hat{p}$
1	.00005	.00007	0.4166	1	.0304	.0899	0.2527
2	.00006	.00006	0.5000	2	.0499	.1290	0.2789
3	.00004	.00006	0.4000	3	.1150	.2280	.3353
4	.00004	.00017	0.1904	4	.2860	.7200	0.2843
5	.00009	-00023	0.2813	5	.2860	.7200	0.2843
6	.00009	.00027	0.2500	6	1.007	1.961	0.3393
7	.00008	.00023	0.2580	7	1.377	2.375	0.3670
8	.00008	.00028	0.2222	8	1.900	2.403	0.4415
9	.00008	.00040	0.1666	9	2.782	3.558	0.4388

Tabla 4. Componentes de varianza y correlación intraclase estimados para las familias que delimitan la zona periférica.

Conclusiones

Al evaluar el crecimiento de la altura y diámetro de las 43 familias se observó que:

Los pinos que conforman la zona central no presentan un comportamiento homogéneo en su crecimiento de altura y diámetro, sucediendo lo mismo con los pinos que conforman la zona periférica. En cada zona al menos existe una familia cuyo crecimiento de altura y diámetro no es igual al de los demás durante los nueve meses de evaluación. Cabe mencionar que al evaluar el diámetro, el mes en el que se presentó la mayor variabilidad entre las familias fue el tercero y que al evaluar la altura fue el noveno.

Al aplicar la técnica de Componentes de la Varianza para cada una de las zonas, se observó que para la zona central al evaluar el diámetro, el mes en el que se presentó mayor variabilidad entre familias fue el tercero y al evaluar la altura fue en el segundo. Para la zona periférica al evaluar el diámetro, el mes en el que se presentó mayor variabilidad entre las familias fue en el segundo y al evaluar la altura fue el octavo.

Referencias

Armas, A. (1995) *Una aplicación de la Estadística en el área de Genética Forestal,* L. I. N. A. E., Fac. de Estadística, U.V.

Aitkin, M. and Logford, N. (1986) "Statistical modelling issues in school effectiveness studies"; *Jour., Roy. Stat. Soc.*, A, **149**, 1–43.

Bryk A. S. and Raudenbush S. W. (1992) *Hierarchical linear models: applications and data analysis methods*, Newbury Park, Sage Publications.

Henderson C.R. (1986) "Recent developments in variance and covariance estimation", *Journal of Animal Sciences*, **63**, 208–216.

Montgomery D.C. (1992) *Diseño y análisis de experimentos*; Grupo Editorial Iberoamérica, México.

Ojeda M.M. (1992) *Aspectos teóricos y metodológicos de la modelación de datos en encuestas complejas*; Tesis Doctoral, Departamento de Matemática Aplicada, Universidad de la Habana, Cuba.

Ojeda M.M. (1993) "Multilevel Modelling Strategies for complex samples", paper presented at *The 1993 European Meeting of The Psichometric Society*, Barcelona, Spain.

Modificaciones en la construcción de las cartas de control p y np para reducir la frecuencia de falsas alarmas

Osvaldo Camacho Castillo & Humberto Gutiérrez Pulido
Centro Universitario de Ciencias Exactas e Ingeniería, Universidad de Guadalajara

Introducción

En las cartas de control de Shewhart es usual aplicar ocho pruebas para detectar cambios especiales (Nelson, 1984). En Camacho y Gutiérrez (1994 y 1995) se demuestra que no es del todo correcto aplicarlas a las cartas **p** y **np**, ya que éstas generan una mayor cantidad de falsas alarmas que las que se esperaran bajo el supuesto de normalidad (ver tabla 1 y figura 1). Lo anterior, como se documenta en Camacho y Gutiérrez (1994 y 1995), contradice a afirmaciones que sobre la aplicación de estas pruebas en cartas **p** y **np** se hacen en algunas referencias clásicas del control de calidad (ver por ejemplo Besterfield, 1990; Western Electric, 1958; y Nelson, 1984) y también cuestiona las omisiones o falta de recomendaciones que sobre el particular se da en otros textos clásicos (ver por ejemplo Montgomery, 1991).

Prevenir el exceso de falsas alarmas en la aplicación de las cartas de control es un aspecto importante, ya que no hacerlo lleva a que el practicante de control de calidad declare con mayor frecuencia que el proceso estuvo fuera de control estadístico, cuando en realidad no fue así.

En este trabajo se presentan modificaciones en la construcción de las carta **p** y **np**, mediante las cuales se logra reducir la cantidad de falsas alarmas cuando se aplican las pruebas estándar a tales cartas. Tales modificaciones conservan la sencillez de la carta, por lo que se considera que es una alternativa fácil de llevar a la práctica.

Las ocho pruebas estándar que se aplican a las cartas de Shewhart son las siguientes (Gutiérrez, 1992):

Prueba 1. Un punto fuera de los límites de control.

Prueba 2. Dos de tres puntos consecutivos en la zona A.

Prueba 3. Cuatro de cinco puntos consecutivos en la zona B o más allá, sin salirse de los límites de control.

Prueba 4. Ocho puntos consecutivos de un sólo lado de la línea central, sin salirse de los límites de control.

Prueba 5. Seis puntos consecutivos en aumento (o en disminución).

Prueba 6. Catorce puntos consecutivos alternando entre altos y bajos.

Prueba 7. Ocho puntos consecutivos a ambos lados de la línea central con ninguno en la zona C.

Prueba 8. Quince puntos consecutivos en la zona C, arriba o abajo de la línea central.

Estas pruebas se derivan a partir del supuesto de normalidad e independencia de los datos generados por el proceso (Western Electric, 1958).

Modificación en la construcción de las cartas p y np

Los límites de control de la carta **p** se calculan típicamente de la siguiente manera:

$$LCS = \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \text{ y } LCI = \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}},$$

y los límites de la carta **np** están dados por

$$LCS = n\overline{p} + 3\sqrt{n\overline{p}(1-\overline{p})} \text{ y } LCI = n\overline{p} - 3\sqrt{n\overline{p}(1-\overline{p})}.$$

Nuestra propuesta consiste en modificar ligeramente la obtención de los límites de control de las cartas **p** y **np**, dejando igual la línea central.

Proponemos que en la carta **np** se calcule el límite de control superior de la siguiente manera:

$$LCS^* = EM[LCS],$$

donde $EM[x]$ es la función mayor entero o igual que x, y LCS es el límite de control superior típico de la carta **np**. Así para calcular el LCS^* lo que se hace es calcular el LCS típico y a éste se le aplica la función mayor entero o igual; por ejemplo si $LCS = 17.872$, entonces $LCS^* = EM[17.872] = 18$. Si el límite superior típico para una carta **np** es un entero, entonces éste coincide con el límite de control modificado, LCS^*. La línea central se calcula de la forma tradicional, y el límite de control inferior modificado (LCI^*) se obtiene de la siguiente manera:

$$LCI^* = n\overline{p} - 3S^*,$$

donde S^* se obtiene al dividir entre tres la distancia entre la línea central y el LCS^*, es decir,

$$S^* = (LCS^* - n\overline{p})/3.$$

De esta manera las zonas de la carta de control tendrán una amplitud igual a S^*.

Los límites de control para una carta **p**, se calculan de la misma manera, dividiendo entre el tamaño de muestra a los límite de la carta **np**. De aquí que el límite de control superior modificado (LCS') para una carta **p** está dado por:

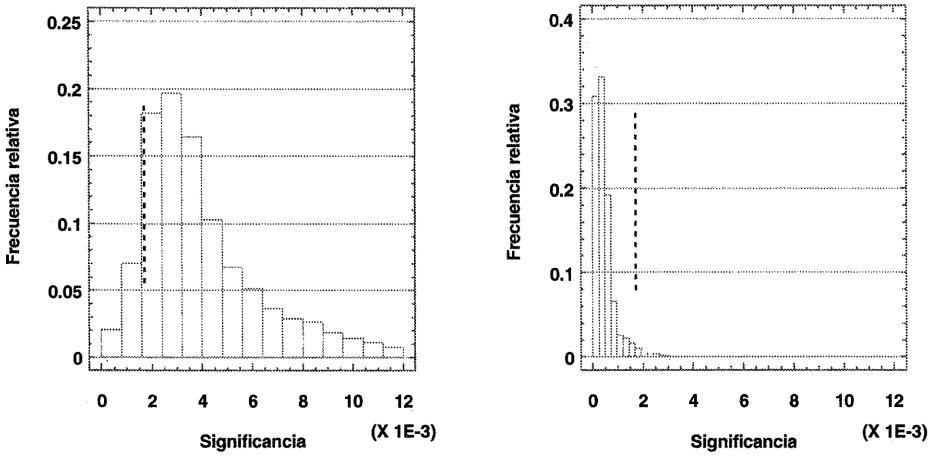
$$LCS' = LCS^*/n.$$

La línea central es la media muestral de p y el límite de control inferior modificado (LCI') está dado por

$$LCI' = LCI^*/n.$$

Para evaluar la viabilidad de la modificación se calculó la significancia de las ocho pruebas para 2000 cartas de control diferentes, tanto las típicas como las modificadas, con $p = 0.001$, 0.002, 0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.05, 0.06, 0.08, 0.1, 0.13, 0.15, 0.16, 0.19, 0.2, 0.22, 0.25, 0.3, 0.35, 0.4, 0.45, y 0.5; y todos los enteros n entre 10 y 500, con la

Figura 1: La prueba 1 en el lado superior de las cartas. La línea vertical punteada se ubica en la significancia bajo normalidad.



(a) Cartas p y np típicas.

(b) Cartas p y np modificadas.

cota superior para $np \leq 20$. De los np obtenidos entre 0 y 1 sólo se evaluó una muestra aleatoria de 40. La significancia de las pruebas 5 y 6 se calculó por el método Monte Carlo, y dado que éstas no se ven afectadas por el cambio no hay necesidad de analizarlas en las cartas modificadas.

Conclusiones

La síntesis de los resultados obtenidos se muestran en la tabla y figura 1, de donde se puede concluir que:

1. Aplicar las ocho pruebas a las cartas **p** y **np** lleva a tener para ciertos valores de np, tanto pequeños como grandes, una mayor frecuencia de falsas alarmas que bajo el supuesto de normalidad. Sobre todo en las pruebas 1 y 2 del 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 específicas). En los casos descritos antes, a medida que np es más pequeño el riesgo se incrementa.

2. Con la modificación propuesta se reduce el riesgo de falsas alarmas en las más de las pruebas, exceptuando la ocho donde ocurre lo contrario. La prueba cuatro se hace más confiable si en lugar de considerar ocho puntos consecutivos de un solo lado de la línea central se consideran nueve.

3. En particular, en la figura 1 se comparan la distribución de frecuencia de las significancias para la prueba 1 del lado superior aplicada a la carta **p** típica y a la carta **p**

modificada, con lo que se aprecia las ventajas de ésta última. Que aunque generalmente es un poco más conservadora que bajo normalidad (ver línea punteada en figura 1), esto se ve compensado por una mayor potencia de la prueba en la carta **p** que en la carta de medias, y por una mejor protección contra las falsas alarmas.

4. Como se demuestra en Camacho y Gutiérrez (1994 y 1995) seguir aplicando algunas o varias de las ocho pruebas a las cartas **p** típicas lleva a tener una mayor posibilidad de riesgo de falsas alarmas, agudizándose lo anterior debido a que muchas de las cartas **p** que se usan en la actualidad tienen valores de n y p pequeños. Por lo que, o se evalúa la significancia específica de cada carta y se busca un valor adecuado de n, o se opta por medidas generales que reduzcan el error tipo I. Nuestra propuesta se orienta en el último sentido, con resultados satisfactorios como se muestra en la tabla y figura 1.

Referencias

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

Camacho Castillo, O. y Gutiérrez Pulido H. (1994) *Caracterización de las pruebas para causas especiales en las cartas p y np*, Memorias del VIII Foro Nacional de Estadística, INEGI, Aguascalientes.

Camacho Castillo, O. y Gutiérrez Pulido H. (1995) "Estudio de la significancia de las pruebas para detectar causas especiales de variación en las cartas p y np", *La Estadística*, INEGI-AME, Aguascalientes.

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

Montgomery, D.C. (1991) *Introduction to Statistical Quality Control*, second edition, Wiley, Singapore.

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

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

Tabla 1: Significancias de las 8 pruebas. CI y CS cuartil inferior y superior. Resultados de 2000 valores de *np* entre 0 y 20.

Prueba	Lado de la carta	Significancia bajo distribución normal	Significancia en cartas p y np	Significancia en cartas p y np modificadas
1	Superior	0.00135	$\overline{\alpha}$ = 4.36766E-3 S = 4.14322E-3 CI= 2.32401E-3 CS= 5.18214E-3	$\overline{\alpha}$ = 4.74281E-4 S = 4.62582E-4 CI= 1.9121E-4 CS= 5.87E-4
	Inferior	0.00135	$\overline{\alpha}$ = 1.42037E-4 S = 2.66915E-4 CI= 0 CS= 1.9158E-4	$\overline{\alpha}$ = 5.93753E-5 S = 1.31221E-4 CI= 0 CS= 5.196E-5
2	Superior	0.001075	$\overline{\alpha}$ = 2.62586E-3 S = 2.66628E-3 CI= 1.15372E-3 CS= 3.22299E-3	$\overline{\alpha}$ = 1.70205E-3 S = 1.54724E-3 CI= 7.8167E-4 CS= 2.12647E-3
	Inferior	0.001075	$\overline{\alpha}$ = 4.78194E-4 S = 6.13491E-4 CI= 0 CS= 8.0476E-4	$\overline{\alpha}$ = 2.357E-4 S = 3.85112E-4 CI= 0 CS= 3.4156E-4
3	Superior	0.0027	$\overline{\alpha}$ = 3.06442E-3 S = 3.03198E-3 CI= 1.18787E-3 CS= 4.20799E-3	$\overline{\alpha}$ = 1.98874E-3 S = 1.87155E-3 CI= 6.9466E-4 CS= 2.85938E-3
	Inferior	0.0027	$\overline{\alpha}$ = 5.65633E-3 S = 8.62424E-3 CI= 1.2844E-3 CS= 5.86092E-3	$\overline{\alpha}$ = 2.62508E-3 S = 4.18724E-3 CI= 5.4355E-4 CS= 3.11966E-3
4	Superior	0.003822	$\overline{\alpha}$ = 1.47611E-3 S = 1.68029E-3 CI= 2.1484E-4 CS= 2.19284E-3	$\overline{\alpha}$ = 1.59908E-3 S = 1.80334E-3 CI= 2.6762E-4 CS= 2.34172E-3
	Inferior	0.003822	$\overline{\alpha}$ = 0.0112063 S = 0.0599405 CI= 1.09681E-3 CS= 6.01864E-3	$\overline{\alpha}$ = 0.0102149 S = 0.0518734 CI= 1.13658E-3 CS= 6.09204E-3
5	Toda la carta	$\hat{\alpha}$ = 0.000335	$\overline{\alpha}$ = 2.64151E-5 S = 5.71744E-5 CI= 0 CS= 0	
6c	Toda la carta	$\hat{\alpha}$ = 0.00253	$\overline{\alpha}$ = 1.00863E-3 S = 7.77592E-4 CI= 4E-4 CS= 1.4E-3	
7	Toda la carta	0.000103	$\overline{\alpha}$ = 4.50695E-4 S = 2.084E-3 CI= 3.746E-5 CS= 3.5988E-4	$\overline{\alpha}$ = 8.21919E-5 S = 1.5522E-4 CI= 1.338E-5 CS= 7.535E-5
8	Toda la carta	0.003254	$\overline{\alpha}$ = 0.0110611 S = 0.0539998 CI= 9.5054E-4 CS= 7.28966E-3	$\overline{\alpha}$ = 0.0259902 S = 0.0822885 CI= 2.14314E-3 CS= 8.48143E-3

La comparación de varios promedios con el promedio general

ALBERTO CASTILLO MORALES
Colegio de Postgraduados

Introducción

En aplicaciones hacia la industria puede ser mas fácil de interpretar la comparación de cada marca con el promedio general que las comparaciones directas entre marcas. Esto ocurre cuando se desea conocer la situación de una marca en relación con el promedio del mercado.

Se dispone de una muestra aleatoria independiente de tamaño n_i de cada una de k marcas: X_{ij}; $j = 1, \ldots, n_i$, $i = 1, \ldots, k$. El planteamiento corresponde a un diseño completamente al azar con diferente número de repeticiones. Suponiendo normalidad e igual varianza las X_{ij} se distribuyen normal con media μ_i, varianza σ^2 y son independientes.

Notación

Se usará la letra Y con los subíndices apropiados para denotar los promedios correspondientes de la variable X, con la notación usual de puntos para representar sumas sobre los subíndices. Con respecto a los parámetros, para la media general se usará μ para la media sin ponderar y μ_p para la media ponderada por los tamaños de muestra. Tanto para los estimadores como para los parámetros, cuando se obtenga el promedio sobre los valores que quedan después de quitar lo correspondientes a la marca o al subíndice i, se denota por el subíndice $(-i)$ que se coloca después del punto que indica sobre que subíndice se sumó, por ejemplo:

$$\mu_{(-i)} = \sum_{h \neq i}^{k} \mu_h / \ (k-1)$$

es la media sin incluir la marca i.

Soluciones directas

La comparación que interesa al usuario se basa en la diferencia

$$Y_{i.} - Y_{..} \tag{1}$$

Se puede construir una prueba de t para la diferencia, pero debe notarse que la media (Esperanza) de $Y_{i.} - Y_{..}$, es

$$E(Y_{i.} - Y_{..}) = \mu_i - \mu_p = \frac{n_{.} - n_i}{n_{.}} - \frac{1}{n_{.}} \sum_{h \neq i}^{k} n_h \mu_h.$$

La sutileza radica en que la esperanza de la diferencia se refiere al promedio ponderado de las medias; no coincide con lo que se intentaba en el inicio, esto es, en comparar μ_i con

μ . Para comparar una media contra las demás, desde el principio se puede plantear el contraste

$$(n_{\cdot} - n_i)Y_{i\cdot} - \sum_{h \neq i}^{k} n_h Y_{h\cdot} \, .$$

Este contraste da una idea muy clara de la hipótesis que se prueba. Se compara la media μ_i con el promedio ponderado de las demás medias, dando mayor peso a aquellas con mayor tamaño de muestra.

En cada problema se debe analizar si es conveniente usar la media ponderada por los tamaños de muestra o no. Cuando interesa comparar medias de marca con la media general en el mercado, se debe ver si los tamaños de muestra de las marcas son proporcionales a su porción de mercado. Si son proporcionales, la ponderación permite una visión mas realista del comportamiento del mercado.

En el caso de que los tamaños de muestra no sean proporcionales a la porción de mercado de cada marca, ni siquiera de manera aproximada, la ponderación no es apropiada y conviene utilizar una ecuación que relacione a los promedios muestrales de manera que su esperanza sea $\mu_i - \mu$; $i = 1, \ldots, k$. Una elección que cumple con la condición deseada es la diferencia

$$Y_{i\cdot} - Y_{(-i)} \tag{2}$$

ya que

$$E(Y_{i\cdot} - Y_{(-i)}) = \mu_i - \mu_{(-i)} = \mu_i - \frac{1}{k-1} \sum_{h \neq i}^{k} \mu_h.$$

La diferencia de la comparación en (2) con la que se plantea en (1) consiste en que en (2) se compara μ_i con la media sin ponderar de los promedios de las marcas restantes.

Nótese que en el caso de que los tamaños de muestra sean iguales, se obtiene el mismo resultado con ambos esquemas, sólo se debe observar que tanto el numerador como el denominador de la t en el segundo caso están multiplicadas por $(k-1)/k$.

La prueba de verosimilitud

Definiendo los espacios paramétricos

$$\Omega = \{-\infty < \mu_i < +\infty;\ i = 1, ..., k,\ 0 < \sigma^2 < +\infty\}$$

y

$$\Omega_0 = \{-\infty < \mu_i < +\infty;\ i = 1, ..., k,\ 0 < \sigma^2 < +\infty;\ \mu_i = (\sum_{h \neq i}^{k} n_h \mu_h)/(n_{\cdot} - n_i)\},$$

se puede obtener la relación de verosimilitud al dividir el máximo de la verosimilitud para Ω_0 sobre el máximo para Ω. La verosimilitud está dada por la densidad conjunta de las observaciones.

La maximización sobre Ω sigue el camino usual y produce la SCE, esto es, el numerador del estimador de la varianza sin restricciones. La maximización con la restricción dada por la hipótesis nula en Ω_0 produce $\text{SCE}_0 = n_{\cdot}n_i(Y_{i\cdot} - Y_{\cdot\cdot})^2/(n_{\cdot} - n_i)$.

Haciendo los cambios usuales para llegar a un análisis de varianza, se obtiene la relación

$$F_c = \left[n_{\cdot}n_i(Y_{i\cdot} - Y_{\cdot\cdot})^2/(n_{\cdot} - n_i)\right]/\text{CME},$$

con 1 y $n_{\cdot} - k$ grados de libertad.

Conviene presentar la esperanza del estimador de varianza bajo la hipótesis nula:

$$E(\text{SCE}_0) = \sigma_2 + (\mu_i - \frac{1}{n_{\cdot} - n_i}\sum_{h\neq i}^{k} n_h\mu_h)^2.$$

Análisis de varianza y pruebas simultáneas de Scheffe y Bonferroni

La hipótesis nula conjunta $\mu_i = \sum_{h\neq i}^{k} n_h\mu_h\,/(n_{\cdot} - n_i)$; $i = 1, ..., k$, coincide con la hipótesis nula de igualdad de medias. Dicho de otra manera, las k comparaciones de media de marca con la media general ponderada por los tamaños de muestra, tienen el mismo rango que la hipótesis de igualdad de medias y que el espacio de los parámetros μ_i; $i = 1, ..., k$. La prueba es equivalente a la que se hace con el análisis de varianza para igualdad de medias de marcas.

Para conocer la situación de cada una de las k marcas con relación al promedio general, se puede utilizar el método de Scheffè. Para revisar cada una de las marcas, en el caso (1) se compara la diferencia de cada media y el promedio general $Y_{i\cdot} - Y_{\cdot\cdot}$ con el valor dado por la ecuación de Scheffè:

$$\left\{(k-1)F_{\alpha,k-1,n_{\cdot}-k}\text{CME}(n_{\cdot} - n_i)/\,(n_{\cdot}n_i)\right\}^{1/2}.$$

Si se desea utilizar la media general sin ponderar por el tamaño de muestra (2), para la marca i se compara $Y_{i\cdot} - Y_{(-i)}$ con el valor correspondiente de Scheffè:

$$\left\{(k-1)F_{\alpha,k-1,n_{\cdot}-k}\text{CME}\left(\frac{1}{n_i} + \frac{1}{(k-1)^2}\sum_{h\neq i}^{k}\frac{1}{n_h}\right)\right\}^{1/2}$$

El procedimiento de Bonferroni tiene la ventaja de ser muy fácil de usar, dando una buena aproximación a la significación general cuando a es pequeño. Todo lo que debe hacerse es cambiar el valor tabulado para las t que se construyeron en la primera sección a partir de las diferencias (1) y (2), y hacer las comparaciones para cada una de las marcas usando $\alpha/[2(k-1)]$ en lugar de $\alpha/2$.

Pruebas simultáneas. Una propuesta específica.

Si $W' = (W_1, ..., W_m)$ tiene distribución normal multivariada con media cero y matriz de varianzas y covarianzas $\sigma^2 R$, donde R es la matriz de correlaciones y tiene inverso, y si

$gCME/\sigma^2$ es una χ^2 con g grados de libertad independiente de W, el vector dado por $T' = (t_1, ..., t_m)$, donde $t_i = W_i\, CME^{-1/2}$; $i = 1, ..., m$, tiene como función de densidad a

$$f(T) = (g\pi)^{-k/2}\Gamma[(m+g)/2]\{\Gamma(g/2)\}^{-1}\{\det R\}^{-1/2}\{1 + T'R^{-1}T/g\}^{-(m+g)/2}.$$

En el caso de comparaciones usando la media ponderada por los tamaños de muestra, se tienen $k-1$ comparaciones con la media general que producen la matriz R no singular con elementos

$$\rho_{ij} = -\{(\frac{n}{n_i} - 1)(\frac{n}{n_j} - 1)\}^{-1/2} \text{ ; para } i \neq j.$$

Si $n_i = n$ para $i = 1, ..., k$, los valores de las correlaciones se reducen a $\rho_{ij} = -(k-1)^{-1}$.

La distribución está tabulada para el caso de tamaños de muestra iguales y matriz $R = I$, que corresponde a la prueba honesta de Tukey.

La prueba de Dunnett para comparar tratamientos con un testigo produce

$$\rho_{ij} = (\frac{n_0}{n_i} + 1)(\frac{n_0}{n_j} + 1)^{-1/2} \text{ ; para } i \neq j.$$

Sus tablas se refieren al caso de igual tamaño de muestra, con $\rho_{ij} = 1/2$.

La distribución que ocurre en las comparaciones de promedio de marca con el promedio general, es diferente de las tabuladas. Para construir una prueba que mejore a los métodos de Scheffè y de Bonferroni se debe elaborar la tabla, por lo menos para tamaños de muestra iguales, dando después, si es posible, lineamientos para los casos donde los tamaños de muestra son diferentes.

Referencias

Miller, R. G. Jr., *Simultaneous Statistical Inference,* McGraw-Hill, New York.

Análisis exploratorio de la infertibilidad en México

Ma. de Lourdes De la Fuente & Tatiana Fernández-Naranjo
Instituto Tecnológico Autónomo de México

Durante los últimos 15 años se han logrado grandes avances en el estudio de la infertilidad. Estos avances se refieren al conocimiento de los mecanismos que intervienen en el proceso de reproducción, tanto en personas normales, como en aquellas en estado patológico. Sin embargo, aún no se cuenta con una visión lo suficientemente completa que permita el desarrollo de nuevas y mejores técnicas para el diagnóstico y tratamiento de este padecimiento. Este trabajo tiene por objeto describir ciertas características de la población de parejas con problemas de infertilidad en México. El propósito es caracterizar esta población, identificando los rasgos comunes de las parejas y explorando las posibles relaciones de éstos con las causas que provocan el trastorno. Los datos analizados en este estudio provienen de los archivos actual y muerto de casos estudiados por un especialista en endocrinología de la reproducción en México. Se cuenta con un total de 309 expedientes.

La infertilidad debe entenderse como un padecimiento relativo al tiempo y a la pareja, esto es, no debe ser considerado un término absoluto-individual. La Sociedad Estadounidense para la Fertilidad la define como "un matrimonio debe considerarse infecundo cuando no ha ocurrido el embarazo después de un año de coito sin contracepción". Son diversas las causas tanto psicológicas como fisiológicas que provocan este trastorno, dentro de las causas fisiológicas se tiene a la anovulación, la endometriosis, el factor masculino, el factor tuboperitoneal, el factor uterino y la interacción moco cervical espermatozoide. Estas causas pueden incidir de manera aislada o en conjunto, por lo que un estudio integral y global de la pareja es recomendable. Por su naturaleza la infertilidad puede ser primaria, cuando no existe evidencia de fertilidad previa, o secundaria, cuando hay fertilidad previa comprobada. Son varios los tratamientos que se han desarrollado con el fin de remover la condición de infertilidad en la pareja, siendo el más común el de la "inducción de ovulación".

La selección de las características de interés se sujeto al juicio y sugerencia del especialista, resultando en tres grupos: datos correspondientes a la mujer, datos al hombre y a la pareja. Para la mujer la edad, el peso conjuntamente con la estatura, y la condición de fumadora o no-fumadora fueron estudiadas. Para el hombre la edad y ocupación, medida en grados de exposición al riesgo. En tanto para la pareja el tipo, tiempo y causa de infertilidad, la frecuencia de relaciones sexuales, al mes, si se indujo o no la ovulación, si hubo o no embarazo, y, en caso de haberlo, si llegó éste al término del periodo gestacional, algunos resultados de interés son los siguientes: La clasificación y codificación de estas características se presentan en la siguiente tabla.

	Característica	Clasificación (variable)	Codificación
MUJER	Edad	Cuantitativa Discreta	(0,...,19,20)
	Peso+estatura	Cualitativa ordinal	Bajo peso 1- En peso 0 Sobre peso 1+
	Fuma	Cualitativa Nominal	No fuma 0 Fuma 1
HOMBRE	Edad	Cuantitativa Discreta	(0,...,19,20)
	Ocupación	Cualitativa Ordinal	Riesgo=F(Estrés, Calor, Activ. sedentaria, subst. químicas) Bajo 1 Medio 2.3 Alto 4
PAREJA	Frecuencia de relaciones sexuales al mes	Cuantitativa Continua	$\mathbf{R}^+$
	Tipo de infertilidad	Cualitativa Nominal	Primaria 1 Secundaria 2
	Tiempo de infertilidad	Cualitativa Continua	$\mathbf{R}^+$
	Causa (1)	Cualitativa Nominal	En estudio 0 Simple 1 Múltiple 2
	Causa (2)	Cualitativa Nominal	Simple 1,....6 Doble 12, 13....,56 Séxtuple 123456
	Inducción	Cualitativa Nominal	No Inducción 0 Sí Inducción 1
	Embarazo	Cualitativa Nominal	No embarazo 0 Sí embarazo 1
	# Embarazo	Cuantitativa Discreta	(0,1,2,3)
	# Embarazo al término	Cuantitativa Discreta	(0,1,2,3)

Los resultados del análisis univariado de las características permitieron establecer un perfil[1] para la mujer, el hombre y la pareja infértil.

El perfil de la mujer infértil resultó:

(i)	Edad	grupo quinquenal: 25-29 Puntualmente	años años
(ii)	Peso+Estatura	sobre peso	
(iii)	Fuma	no fumadora	

El perfil del hombre infértil resultó:

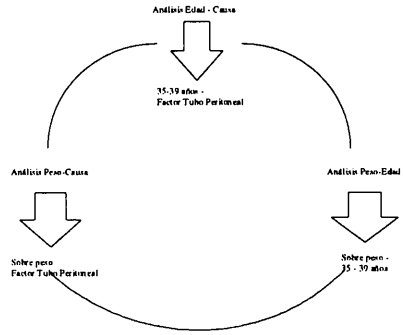
[1] Los rasgos que determinan dicho perfil resultan ser los valores modales de cada una de las características estudiadas.

(i)	Edad	Grupo quinquenal: 30-34 Puntualmente: 32	años años
(ii)	Ocupación	riesgo medio	

El perfil de la pareja infértil resultó:

(i)	Frecuencia de Relaciones sexuales al mes	5-10 veces
(ii)	Tiempo de Infertilidad	0-2 años
(iii)	Tipo de Infertilidad	Primaria
(iv)	Causa (1)	Simple
(v)	Causa Simple	Anovulación
(vi)	Inducción	Sí
(vii)	Embarazo	No

El análisis bivariado de las características permitió establecer relaciones interesantes entre algunos de los rasgos estudiados y la causa y tipo de infertilidad. Adicionalmente, se obtuvieron resultados al relacionar la edad de la mujer con su peso o la inducción de embarazo con la incidencia de éste. La variable causa se cruz con las variables de edad (mujer y hombre), peso, fuma y tipo de infertilidad, obteniendo los siguientes resultados. La causa modal por grupo quinquenal de edad fue anovulación, salvo para las mujeres entre 35 - 39 años, para las cuales el factor tubo-peritoneal resultó ser la causa modal. Con respecto al peso, las mujeres en bajo peso presentaron anovulación y endometriosis con mayor frecuencia, en tanto que aquellas en sobrepeso lo hicieron con factor tubo-peritoneal. Cabe agregar en este punto que la mayor porción de mujeres en sobrepeso se ubicó entre los 35 y 39 años de edad, por lo que se observa que:



Siguiendo en la misma línea, las fumadoras presentaron interacción moco cervical espermatozoide en tanto que las no fumadoras no presentaron ni un solo caso. Finalmente, la infertilidad primaria se vio caracterizada en su mayoría por anovulación, y la secundaria por endometriosis y factor tubo-peritoneal. En relación al estudio del tipo de infertilidad, se vio que la infertilidad primaria se presenta como mayor frecuencia en mujeres en bajo peso, no-fumadoras y mujeres entre los 20 y 34 años de edad: mientras que la secundaria lo hace en mujeres en peso y sobre peso, fumadoras y con 40 años y más. Cabe mencionar además que las mujeres en bajo peso fueron quienes lograron el mayor número de embarazos (máximo fumadoras: 2; máximo no-fumadoras: 3), pero el menor número de embarazos al término, mientras que aquellas en peso o sobre peso no lograron un gran número de embarazos, pero si terminaron la mayoría de los que iniciaron.

Cartas de control robustas: una aplicación

CONSUELO DÍAZ TORRES
Universidad Autónoma Metropolitana, Unidad Iztapalapa

Introducción

[1] Se presenta un resumen de los métodos alternativos a las tradicionales cartas de control de Shewart. Estos métodos, que han sido propuestos por diferentes autores, se basan en el cálculo de estadísticas resistentes para generar procedimientos robustos. Finalmente se presenta un ejemplo de aplicación.

La carta $\bar{X} - R$ propuesta por Shewart en 1924 ha sido usada tradicionalmente en la industria para la evaluación y control de procesos productivos. Para su elaboración se usan como medidas descriptivas de los subgrupos la media y el rango y para calcular los límites de control se usan los promedios de las medias y de los rangos. La gran aceptación que ha tenido esta carta se debe a la sencillez en su construcción e interpretación y a que permite detectar cambios significativos en el proceso, debido a la utilización de estadísticas sensibles.

Una desventaja de este procedimiento es que se basa en la suposición de que la distribución de las observaciones de las muestras es normal. En muchas ocasiones sucede que el conjunto de observaciones con que se cuenta tiene una distribución sesgada y/o con observaciones extremas, lo que ocasiona que los límites de control calculados en la forma tradicional sean muy amplios y no sea posible detectar fácilmente situaciones fuera de control o cambios en el proceso.

Algunos autores han propuesto como alternativa el uso de medidas resistentes para generar procedimientos robustos. En la sección 2 se presenta una breve descripción de algunos de estos procedimientos y en la sección 3, un ejemplo de aplicación a un caso que se presentó en la industria.

Cartas de control alternativas

Los procedimientos propuestos como alternativa a las cartas de control tradicionales se basan en el cálculo de estadísticas resistentes.

Una estadística es *resistente* si no es gravemente afectada por unas cuantas observaciones extremas. Las estadísticas resistentes se definen a partir de las estadísticas de orden de los subgrupos, denotadas por

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}.$$

Ferrel (1953) y Clifford (1959) fueron los primeros en proponer el uso de medidas resistentes en la construcción de cartas de control. Ferrel propuso usar medianas y rangos medios. Clifford extiende el trabajo de Ferrel y sugiere calcular la mediana en vez del rango medio y la mediana del rango en vez del promedio de los rangos.

[1] Nota de los editores: la ausencia de algunas gráficas del presente trabajo se debe a problemas técnicos generados por los archivos enviados por la autora.

Langenberg e Iglewicz (1986) proponen el uso de medidas resistentes solamente para calcular los límites de control y la media y el rango para describir los subgrupos.

Para calcular los límites utilizan la media recortada definida por Hoaglin, Mosteller y Tuckey (1983) como

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

donde

$\alpha \in [0,1]$: proporción de observaciones que se recorta;
n: tamaño de subgrupo;
$r = [\alpha n]$;
$t = \alpha n - r$.

Si $\bar{\bar{X}}(\alpha)$ y $\bar{R}(\alpha)$ son las medias recortadas de las medias y de los rangos de los subgrupos, respectivamente, entonces los límites de control se calculan como se muestra en el cuadro 1.

Debido al uso de medias recortadas para calcular los límites, éstos son igual o más estrechos que en la carta de control tradicional, con lo que se obtiene un procedimiento robusto, capaz de detectar cambios repentinos en el proceso y puntos fuera de control más eficientemente.

En forma independiente White y Schroeder (1987) y Hoaglin e Iglewicz (1987) proponen el uso de diagramas de caja para la construcción de cartas de control calculando también medidas resistentes.

White y Schroeder (1987) sugieren calcular la mediana y la Q-dispersión dentro de cada subgrupo y el promedio de estas estadísticas para calcular los límites de control.

La Q-dispersión es una versión simplificada del rango intercuartílico, en vez de usar los valores interpolados, la Q-dispersión es la diferencia entre los valores inmediatos interiores, es decir,

$$R_Q = \begin{cases} X_{(n-f-1)} - X_{(f)} & \text{si } f \text{ es entero} \\ X_{(n-i)} - X_{(i+1)} & \text{si } f \text{ no es entero} \end{cases}$$

donde

$$f = \frac{\left[\frac{n+3}{2}\right]}{2}, \qquad i = [f].$$

Los límites de control para esta carta se reportan también en el cuadro 1.

Una de las críticas a este procedimiento es que no permite detectar cambios significativos en el proceso, debido a que la mediana no es sensible a observaciones extremas.

Hoaglin e Iglewicz (1987) proponen el uso de estadísticas más eficientes como la F-media y la F-dispersión. Proponen dos procedimientos: la carta $M - R_F$, es decir, mediana y F-dispersión y la carta $\bar{X}_F - R_F$, es decir, F-media y F-dispersión.

La F-media denotada por $\bar{X}_F$ se define para $n \geq 4$ como sigue:

$$\bar{X}_F = \begin{cases} \overline{\overline{X}} & \text{para } n = 4 \\ \left[\sum_{j=f}^{n+1-f} X_{(j)}\right] / (n+2-2f) & \text{para } f \text{ entero} \\ \left[X_{(f)} + X_{(n+1-f)} + \sum_{j=i+2}^{n-i-1} X_{(j)}\right] / (n-2i) & \text{para } f \text{ no entero} \end{cases},$$

donde f e i se definen como antes.

La F-dispersión se define como

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

La diferencia entre la Q-dispersión y F-dispersión es que la primera evita la interpolación.

Los límites de control para estas cartas se calculan a partir de los promedios de las estadísticas de los subgrupos y se reportan en el cuadro 1.

Los autores compararon la eficiencia relativa de la mediana y la F-media con respecto a la media y obtuvieron que la F-media es la más eficiente. Con respecto a la dispersión, compararon la eficiencia de la Q-dispersión y la F-dispersión con el rango y encontraron que la F-dispersión tiene una eficiencia mayor o igual a la Q-dispersión. Los autores recomiendan la carta $\bar{X}_F - R_F$.

Autores	Cartas	Límites de control	
		Localización	Dispersión
Langenberg e Iglewicz	$\bar{X}(\alpha) - \bar{R}(\alpha)$	$\bar{X}(\alpha) \pm A_2 k\bar{R}(\alpha)$	$D_3 k\bar{R}(\alpha), D_4 k\bar{R}(\alpha)$
White y Schroeder	$M - R_Q$	$\bar{M} \pm A_{2Q}\bar{R}_Q$	$D_{3Q}\bar{R}_Q, D_{4Q}\bar{R}_Q$
Hoaglin e Iglewicz	$M - R_F$	$\bar{M} \pm A_{2M}\bar{R}_F$	$D_{3F}\bar{R}_F, D_{4F}\bar{R}_F$
	$\bar{X}_F - R_F$	$\bar{\bar{X}}_F \pm A_{2F}\bar{R}_F$	$D_{3F}\bar{R}_F, D_{4F}\bar{R}_F$

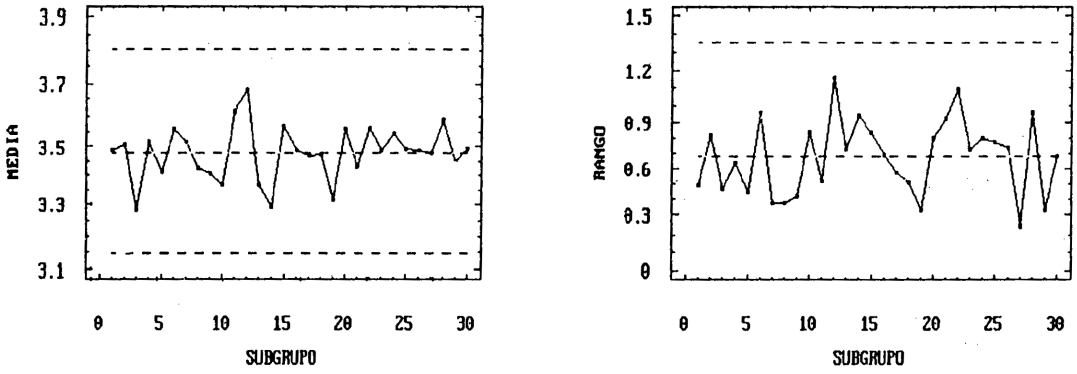
Cuadro 1. Límites de control para los diferentes tipos de cartas.

Las constantes A_2, D_3 y D_4 son las usuales. Las constantes A_{2Q}, A_{2M}, A_{2F}, D_{3Q}, D_{4Q}, D_{3F} y D_{4F} son tabuladas por Hoaglin e Iglewicz (1987) para tamaños de subgrupos de 4 a 15. Las constantes A_{2Q}, D_{3Q} y D_{4Q} también son tabuladas por White y Schroeder (1987) como M_2, Q_3 y Q_4 respectivamente para tamaños de subgrupos de 2 a 20.

Ejemplo

En una empresa automotriz se tiene una máquina de control numérico con 6 estaciones para una operación de pulido del diámetro interior de una pieza. Las piezas llegan a la línea de producción con un diámetro interior inicial de 53.0 ± 0.5 mms. La especificación del diámetro interior después de la operación de pulido debe ser 53.5 ± 0.5 mms. Después de este proceso la pieza pasa al área de ensamble.

Gráfica 1: Carta $\bar{X} - R$ para el diámetro interior



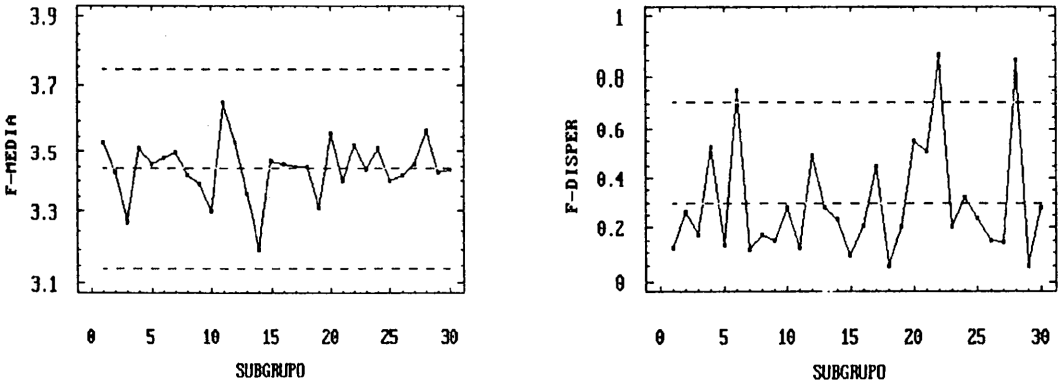
Durante mucho tiempo para este proceso se llevó una carta $\bar{X} - R$, en la cual nunca se observaron puntos fuera de los límites, lo que hacía pensar al personal responsable del proceso que éste estaba dentro de control, sin embargo se recibían quejas del área de ensamble por mal ajuste o fugas en un alto porcentaje de piezas ensambladas por tener un diámetro interno muy grande.

Se realizó un análisis exploratorio de las observaciones registradas y se encontró que su distribución es sesgada, con aproximadamente un 5% de valores extremos en la parte superior de la distribución y un 1.6% de valores fuera de especificación. .Por tanto se sugirió emplear medidas resistentes como la F-media y la F-dispersión y construir una carta $\bar{X}_F - R_F$.

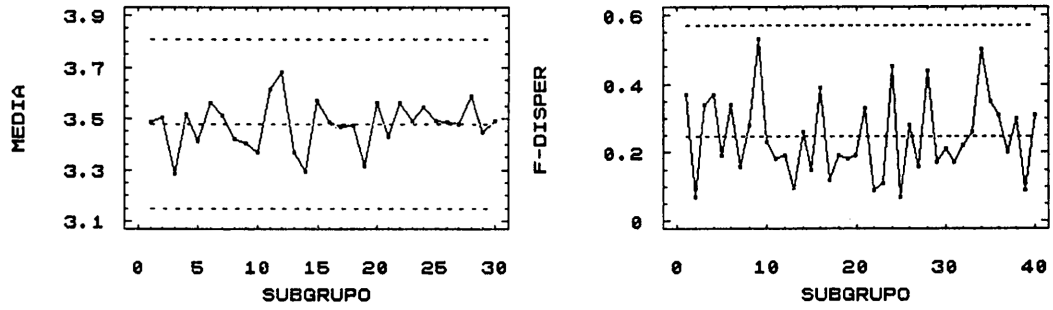
Se tomó una muestra de 30 subgrupos y se calcularon los límites de acuerdo al procedimiento seleccionado. En la gráfica 1 se muestra la carta $\bar{X} - R$ para estos datos y en la gráfica 2 se muestra la carta $\bar{X}_F - R_F$. Comparando estas dos gráficas se puede observar que en la gráfica de medias existe adhesión a la línea central, en la gráfica de F-media los límites de control son ligeramente más estrechos y la línea central está un poco más abajo que en la de medias. En la gráfica para F-dispersión los límites son mucho más estrechos que en la de rangos, encontrándose 3 puntos fuera de los límites.

Se recalcularon los límites de control, después de eliminar los subgrupos correspondientes a los puntos fuera de éstos y con éstos se estableció la nueva carta. En la gráfica 3 se muestra esta nueva carta con los límites recalculados y con datos registrados durante el proceso normal. En ésta se observa, para el subgrupo 9 un punto fuera de control en la gráfica de F-medias con una F-dispersión muy alta. Inmediatamente se hizo un análisis técnico y se encontró y corrigió la falla.

Gráfica 2: Carta $\bar{X}_F - R_F$ para el diámetro interior



Gráfica 3: Carta $\bar{X}_F - R_F$ durante el proceso



Comentarios

Los procedimientos mencionados en este trabajo son tan sencillos de implementar como el tradicional y dan mucho mejor resultado en los casos de desviación de la normal y en presencia de valores extremos.

Se recomienda que antes de iniciar el control del proceso mediante alguna carta se haga un análisis exploratorio de los datos para determinar su naturaleza y poder elegir adecuadamente el tipo de carta.

Referencias

Barrios, E., y Domínguez, J., (1990) "Cartas de Control Robustas", *Memorias del IV Foro de Estadística.*

Iglewicz, B., and Hoaglin, D. (1987) "Use of boxplots for process evaluation", *Journal of Quality Technology,* **19**, 180–190.

Hoaglin, D., Mosteller, F. and Tuckey, J. W. (1983) *Understanding robust and exploratory data analysis*, John Wiley and Sons, Inc., New York.

Langenberg, P., and Iglewicz B. (1987) "Trimmed mean $\bar{X}$ and R charts", *Journal of Quality Technology,* **18**, 152–161.

Tuckey, J. W. (1977) *Exploratory data analysis*, Addisson Wesley, Reading.

White, E. and M., Schroeder, R. (1987) "A simultaneous control chart", *Journal of Quality Technology,* **19**, 1–10.

Comparación de los parámetros de localización de dos distribuciones exponenciales

JESÚS ARMANDO DOMÍNGUEZ MOLINA
Centro de Investigación en Matemáticas, A.C. y Facultad de Matemáticas, Universidad de Guanajuato

Introducción

[1] En 1994 Nassarallah y Saleh consideraron el problema de hacer inferencia sobre la diferencia de los parámetros de localización de dos muestras exponenciales de dos parámetros $\sigma^{-1}\exp\left[(x-\mu)/\sigma\right]$ cuando los parámetros de escala son desconocidos y diferentes. Ellos realizan la inferencia sobre la diferencia $\delta = \mu_1 - \mu_2$ utilizando metodología Bayesiana. Todo su trabajo se enfoca en obtener la distribución a posteriori de δ dado los datos. Nassarallah y Saleh afirman que puesto que σ_1 y σ_2 son desconocidas y diferentes, el problema no da pie a una solución no-Bayesiana razonable para muestras finitas. Es por eso que ellos toman el recurso de la metodología Bayesiana para dar una solución "manejable". En contraposición, en este trabajo se trata el problema con el enfoque de la *inferencia pivotal*, obteniendo una solución exacta sin importar el tamaño de la muestra ni el hecho de que σ_1 y σ_2 sean desconocidas y diferentes. Una generalización de los resultados obtenidos y un desarrollo más completo se encuentra en Domínguez y Sprott (1995a).

Fundamentos

El término *pivotal* fue introducido por R. A. Fisher en 1945 como una función $t\left(\{y_i\},\theta\right)$ de los parámetros y las observaciones la cual tiene una distribución especificada, es decir, se puede calcular numéricamente la probabilidad de que t pertenezca a una región determinada, independientemente del valor de θ.

La inferencia pivotal consiste básicamente de dos pasos: 1) Hacer transformaciones uno a uno. 2) Condicionar sobre cantidades conocidas y cuya distribución es conocida. Más adelante veremos en que consisten los dos pasos anteriores.

De manera más compacta podemos decir que la inferencia pivotal consiste en encontrar una o varias estadísticas $T_1, T_2, ..., T_n$ y una expresión $h\left(T_1, ..., T_n, \theta\right)$ en la que aparezcan los T_j's y el parámetro desconocido (y quizás otras cantidades desconocidas) tal que la distribución de esta cantidad no dependa del parámetro desconocido (pivotal).

El modelo exponencial

Supongamos dos muestras de observaciones independientes $x_{(1)}, ..., x_{(r_1)}$ de una muestra aleatoria de n_1 artículos e $y_{(1)}, ..., y_{(r_2)}$ de una muestra aleatoria de n_2 artículos, con censura en la r_1-ésima falla y la r_2-ésima falla, respectivamente.

[1] Trabajo apoyado por CONACyT, proyecto 1858E9219.

Las x_i's provienen dèl modelo $\sigma_1^{-1}\exp\left[\left(x_i-\mu_1\right)/\sigma_1\right]$, $x_i>\mu_1$, $i=1,...,r_1$, y las y_j del modelo $\sigma_2^{-1}\exp\left[\left(y_j-\mu_2\right)/\sigma_2\right]$, $y_j>\mu_2$, $j=1,...,r_2$, $-\infty<\mu_1,\mu_2<\infty$, $0<\sigma_1,\sigma_2<\infty$. Por conveniencia denótese a σ por σ_1 y a σ_2 por $\rho\sigma$, donde $\rho=\sigma_2/\sigma_1$. La densidad conjunta de $x_{(1)},...,x_{(r_1)}$ y $y_{(1)},...,y_{(r_2)}$ está dada como

$$
\begin{aligned}
&g\left(x_{(1)},...,x_{(r_1)},y_{(1)},...,y_{(r_2)}\right)\\
&=\binom{n_1}{r_1}\binom{n_2}{r_2}\exp\left\{-\left[\sum_{i=1}^{r_1}\left(x_{(i)}-\mu_2\right)/\sigma+\left(n_1-r_1\right)\left(x_{(r_1)}-\mu_2\right)/\sigma\right.\right.\\
&\left.\left.+\sum_{j=1}^{r_2}\left(y_{(j)}-\mu_2\right)/\rho\sigma+\left(n_2-r_2\right)\left(y_{(r_2)}-\mu_2\right)/\rho\sigma\right]\right\}.
\end{aligned}
$$

Los estimadores de máxima verosimilitud de los parámetros μ_1,μ_2,σ,ρ, son los siguientes

$$\hat{\mu}_1=x_{(1)},\ \hat{\mu}_2=y_{(1)},$$

$$s_1=\frac{r_1}{n_1}\hat{\sigma}=\frac{1}{n_1}\left[\sum_{i=1}^{r_1}\left(x_{(i)}-x_{(1)}\right)+\left(n_1-r_1\right)\left(x_{(r_1)}-x_{(1)}\right)\right],$$

$$s_2=\frac{1}{n_2}\left[\sum_{j=1}^{r_2}\left(y_{(j)}-y_{(1)}\right)+\left(n_2-r_2\right)\left(y_{(r_2)}-y_{(1)}\right)\right].$$

La diferencia $\delta=\mu_1-\mu_2$ de localizaciones

En 1993 Barnard y Sprott prueban que la mejor cantidad para hacer inferencia sobre δ está dado por $w=\left(x_{(1)}-y_{(1)}-\delta\right)/s_1$.La distribución de w depende sólo de ρ, por consiguiente w es un pivotal únicamente si conocemos ρ.

La cantidad $v=\hat{\rho}/\rho$ es un pivotal que involucra sólo a ρ y dado que w contiene únicamente a δ, entonces calcularemos la densidad conjunta de w y v con motivo de considerar después la información que se tenga sobre ρ.

La densidad conjunta de (w,v) está dada por (véase Domínguez 1994 ó Domínguez y Sprott 1995)

$$
g\left(w,v\right)=\begin{cases}k_1v^{r_2-1}\left(n_1\tilde{\rho}+n_2v\right)^{-1}\left(n_1+n_2v+n_1w\right)^{-(r_1+r_2-1)}, & \text{si } w>0,\\ k_1v^{r_2-1}\left(n_1\tilde{\rho}+n_2v\right)^{-1}\left(n_1+n_2v-n_2vw/\tilde{\rho}\right)^{-(r_1+r_2-1)}, & \text{si } w<0,\end{cases}\qquad(1)
$$

donde $k_1=n_1^{r_1}n_2^{r_2}\left(r_1+r_2-2\right)!/\left(r_1-2\right)!\left(r_2-2\right)!$.

Con (1) podemos obtener la distribución relevante para hacer inferencias sobre w, considerando las siguientes dos situaciones:

Inferencia sobre δ cuando ρ es especificado ($\rho = \sigma_2/\sigma_1$)

Cuando $\rho = \rho_o$ se conoce numéricamente, se tiene que el pivotal v es observado como $v = v_o = \hat{\rho}_o/\rho_o$. En esta situación v es auxiliar, puesto que su distribución queda completamente determinada y no depende de parámetros desconocidos. Siguiendo el enfoque de la inferencia pivotal lo que haremos será condicionar sobre $v = v_o$ para obtener la distribución relevante de w para hacer inferencia sobre δ la cual es proporcional a (1) con $v = v_o$ fijo

$$g(w|v_o) = \begin{cases} k_2(v_o)(n_1 + n_2 v_o + n_1 w)^{-(r_1+r_2-1)}, & \text{si } w > 0, \\ k_2(v_o)(n_1 + n_2 v_o - n_2 v_o w/\hat{\rho}_o)^{-(r_1+r_2-1)}, & \text{si } w < 0, \end{cases}$$

donde $k_2 = \left[n_1 n_2 v_o (r_1 + r_2 - 2)(n_1 + n_2 v_o)^{r_1+r_2-2}\right](n_1\hat{\rho}_o + n_2 v_o)^{-1}$.

En este caso el análisis consiste en especificar un conjunto $\{\rho_o\}$ de valores posibles para ρ, los cuales se pueden encontrar con la distribución marginal de v. Entonces las inferencias sobre δ quedan como funciones del conjunto $\{\rho_o\}$, similar a lo presentado en la Sprott y Farewell (1993).

Inferencia sobre δ cuando ρ no es especificado (El problema de Behrens-Fisher)

En este caso no existen un pivotal que contengan sólo a δ. Por ejemplo, si ρ no se conoce, tampoco conocemos la distribución de $\hat{\rho}$. Por lo tanto la distribución de w es desconocida, puesto que ésta depende de $\hat{\rho}$. Sin embargo, usando (1) se puede calcular la distribución de w como función de v: $\Pr(w \geq w_o > 0, v)$ y $\Pr(w \leq w_o < 0, v)$.

Supondremos que la ausencia de conocimiento sobre ρ no cambia la distribución de v (lo que consideraremos como definición de "ausencia de conocimiento sobre ρ"). Cuando se observa $\hat{\rho} = \hat{\rho}_o$ se tiene una realización de $v = \hat{\rho}_o/\rho$, pero su distribución no cambia y no se conoce su valor numérico. El pivotal v contiene al parámetro desconocido ρ, entonces si se integra con respecto a v manteniendo $\hat{\rho} = \hat{\rho}_o$, se obtiene la distribución de w, la cual está dada por las siguientes dos ecuaciones:

$$\Pr(w \geq w_o > 0) = \int_{v=0}^{\infty} \Pr(w \geq w_o > 0, v)\, dv, \tag{2}$$

$$\Pr(w \leq w_o < 0) = \int_{v=0}^{\infty} \Pr(w \leq w_o < 0, v)\, dv. \tag{3}$$

Con las dos probabilidades anteriores se obtiene la distribución relevante de w para hacer inferencia sobre δ. Las expresiones exactas de las integrales (2) y (3) las puede encontrar en Domínguez y Sprott (1995b).

Conclusión

La solución mostrada en este trabajo para la diferencia de localizaciones es no-Bayesiana y quizás no clásica en el sentido de Nassarallah y Saleh (1994, p58) y proporciona una solución razonable para muestras finitas.

Si el lector quiere ahondar en este tema puede consultar las referencias al final de este trabajo.

Referencias

Barnard, G.A. y Sprott D.A. (1983) "The generalised Problem of the Nile: Robust Confidence Sets for Parametric Functions", *Annals of Statistics*, **11**, 104–113.

Domínguez, J.A. (1994) *Comparación de Parámetros de Localización*, Tesis de grado de Maestro en Estadística, Universidad de Guanajuato, Facultad de Matemáticas, Guanajuato, Gto., México.

Domínguez, J.A. y Sprott, D.A. (1995a) "Comparison of location parameter of two exponential distributions", por aparecer en *Statistical Papers*.

Domínguez, J.A. y Sprott, D.A. (1995b) "The difference of location parameter of two exponential distributions". Reporte Técnico del CIMAT.

Sprott, D. A. and Farewell, V.T. (1993) "The difference between two normal means", *The American Statistician*, **47**, 126–128.

Nassarallah, B. and Saleh, A.K. Ms. E. (1994) "Comparison of location parameters of two exponential distributions when the scale parameters are different and unknown", *Statistical Papers* **35**, 57–69.

El efecto de las observaciones agrupadas en la estimación de medias y proporciones poblacionales

MARTÍN HUMBERTO FÉLIX MEDINA
Escuela de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa
IGNACIO OSUNA RAMÍREZ
Facultad de Ciencias Químico-Biológicas, Universidad Autónoma de Sinaloa

Introducción

En los estudios epidemiológicos es común la obtención de muestras complejas (estratificadas, multietápicas, etc.), sin embargo, también es común que los datos se analicen como si la muestra fuera aleatoria simple (observaciones independientes e identicamente distribuidas). Como consecuencia de este análisis inadecuado, se incrementa el riesgo de obtener resultados erróneos.

En el presente trabajo se realiza un estudio "empírico" mediante el cual se comparan los resultados que se obtienen al tomar en cuenta el diseño muestral *versus* los que se obtienen al ignorarlo, cuando se comparan medias de diferentes subpoblaciones y se realizan pruebas de hipótesis sobre la independencia de dos variables dicotómicas.

Los datos que se analizan provienen de un estudio sobre parasitosis intestinal llevado a cabo por la Unidad de Investigaciones en Salud Pública de la Facultad de Ciencias Químico-Biológicas de la Universidad Autónoma de Sinaloa. La población objetivo estuvo integrada por N familias (conglomerados) de diferentes tamaños ($M_i = 1, ..., N$). De esta población se obtuvo una muestra aleatoria simple sin reemplazo de n conglomerados, cada individuo fue clasificado como parasitado o no parasitado, de acuerdo con los resultados de un análisis de parasitosis intestinal. Asimismo, a cada uno de ellos se le determinó el número de hematocritos por cm^3.

Aunque el estudio epidemilógico tuvo varios objetivos y por lo tanto se realizaron varios análisis estadísticos, aquí sólo se presentan los resultados sobre la comparación de las medias de hematocritos entre la subpoblación de parasitados y la subpoblación de no parasitados y, los resultados de la prueba de independencia, en cuanto a la presencia, de dos tipos de parásitos.

Comparación de las medias de hematocritos de las subpoblaciones de parasitados y no parasitados

Análisis ignorando el diseño muestral

En este caso se supondrá que se obtuvo una muestra aleatoria simple de $m = \sum_{i=1}^{n} M_i$ individuos, por lo que las observaciones (x_i, y_i), $i = 1, ..., n$; son independientes e idénticamente distribuidas, donde $x_i = 1$ si el i-ésimo individuo está parasitado, $x_i = 0$ en caso contrario y, y_i indica el número de hematocritos por cm^3. El objetivo es estimar $\mu_1 - \mu_0$; es decir, la diferencia entre las medias de hematocritos de los parasitados y de los no parasitados.

El estimador es simplemente la diferencia entre las medias muestrales de hematocritos de los parasitados (subpoblación 1) y de los no parasitados (subpoblación 0), es decir $\hat{d} = \overline{y}_1 - \overline{y}_0$. Intervalos de confianza para $\mu_1 - \mu_0$ se obtienen considerando que ambas muestras (no parasitados y parasitados), de tamaños n_i, $i = 0, 1$, son independientes, por lo que $\hat{d}/\sqrt{\hat{V}(\hat{d})}$ tiene aproximadamente una distribución T con $n_1 + n_2 - 2$ grados de libertad.

Análisis tomando en cuenta el diseño muestral

Se denotará por (x_{ij}, y_{ij}) la observación asociada al j-ésimo individuo del i-ésimo conglomerado ($j = 1, ..., M_{ij}; i = 1, ..., n$). Las variables x y y se definen como en el caso anterior. De acuerdo con Cochran (1977, p. 181) un estimador de $\mu_1 - \mu_0$ es también $\hat{d} = \overline{y}_1 - \overline{y}_0$ (estimador de razón). Cochran también proporciona un estimador de la varianza y sugiere que intervalos de confianza para $\mu_1 - \mu_0$ se basen en la distribución normal asintótica de $\hat{d}/\sqrt{\hat{\mathbf{V}}(\hat{d})}$.

Resultados obtenidos de la aplicación a datos reales

La población muestreada estuvo integrada por 341 familias de las cuales se seleccionaron 56. Se presentan los resultados para tres casos (parásitos): Giardia lamblia, Hymenolepis nana y Entamoeba coli. Los resultados son los siguientes:

Caso	$\overline{y}_1 - \overline{y}_0$	Int. del 95% (ignorando diseño muestral)	Int. del 95% (considerando diseño muestral)	Efecto de liseño
Giardia lamb.	-1.736	$(-2.930, -0.542)$	$(-3.045, -0.427)$	1.20
Hymen. nana	-1.523	$(-2.640, -0.406)$	$(-2.617, -0.429)$	1.04
Entam. coli	-0.953	$(-2.305, 0.399)$	$(-2.250, 0.344)$	0.97

Los resultados que se obtienen con los dos tipos de análisis son practicamente iguales, esto se debe a que las observaciones no presentan correlación intraconglomerado, como lo indican los valores cercanos a 1 del efecto de diseño.

Análisis de tablas de contingencia

Análisis ignorando el diseño muestral

En este caso la observación correspondiente al i-ésimo individuo se denotará por $(x_{i(1)}, x_{i(2)})$, $i = 1, ..., m$; donde $x_{i(k)} = 1$ si el individuo presenta el parásito k ($k = 1, 2$) y $x_{i(k)} = 0$ en caso contrario. Si $p_{1\cdot}$ y $p_{0\cdot}$ denotan las probabilidades de la presencia y de la ausencia del parásito 1 respectivamente y, de manera similar se definen las probabilidades $p_{\cdot 1}$ y $p_{\cdot 0}$ para el parásito 2, entonces la hipótesis de interés se expresa mediante H_0: $p_{rs} = p_{r\cdot}p_{\cdot s}$ $r, s = 0, 1$ *vs* H_1: $p_{rs} \neq p_{r\cdot}p_{\cdot s}$ para algunos r, s; donde $p_{rs} = \Pr(x_{(1)} = r, x_{(2)} = s)$.

La estadística de prueba es la bien conocida χ^2 de Pearson, la cual, bajo H_0, se distribuye asintóticamente como una variable $\chi^2(1)$

Análisis tomando en cuenta el diseño muestral

La observación correspondiente al j-ésimo individuo del i-ésimo conglomerado se denota por $(x_{ij(1)}, x_{ij(2)})$, $j = 1, ..., M_i$, $i = 1, ..., n$; donde $x_{ij(k)} = 1$ si el individuo presenta el parásito k $(k = 1, 2)$, y $x_{ij(k)} = 0$ en caso contrario. Defínanse p_{rs}, $p_{r\cdot}$ y $p_{\cdot s}$ $r, s = 0, 1$ como en el caso anterior y considérese el mismo contraste de hipótesis. Rao y Thomas (1989) señalan que la estadística a usar es $\chi^2_p(\hat{\delta}_\cdot) = \chi^2_p/\hat{\delta}_\cdot$ donde χ^2_p es la χ^2 de Pearson ordinaria y $\hat{\delta}_\cdot$ es el efecto de diseño generalizado (la expresión de $\hat{\delta}_\cdot$ puede verse en Rao y Thomas (1989)). De acuerdo con ellos, bajo H_0, la estadística ajustada $\chi^2_p(\hat{\delta}_\cdot)$ tiene una distribución asintótica $\chi^2(1)$.

Resultados obtenidos de la aplicación a datos reales

Se presentan los resultados de la prueba de independencia para dos situaciones: Giardia lamblia-Hymenolepis nana y, Giardia lamblia-Ascaris lumbricoides. Los resultados son los siguientes:

Caso	χ^2_p	$\chi^2_p(\hat{\delta}_\cdot)$	$\hat{\delta}_\cdot$
Giardia lamb–Hym. nana	5.024*	3.160	1.595
Giardia lamb–Ascaris lumb	4.648**	3.296	1.416

(*) significante al 5%
(**) significante al 2.5%

Es claro que los resultados que se obtienen con cada tipo análisis son diferentes, ésto se debe a que en este caso la correlación intraconglomerado es distinta de cero, como lo indican los valores mayores que 1 del efecto de diseño.

Agradecedemos a la Maestra Sylvia Paz Díaz Camacho, responsable de la Unidad de Investigaciones en Salud Pública de la Facultad de Ciencias Químico Biológicas de la UAS, el habernos facilitado los datos que se utilizaron en este trabajo.

Referencias

Cochran W.G. (1977) *Sampling Techniques,* 3rd edn., Wiley, New York.

Rao J.N.K. and Thomas D.R. (1989) "Chi-square tests for contingency tables", *Analysis of Complex Surveys*, Skinner C.J., Holt D. and Smith. T.M.F., Chichester: Wiley, 89-114.

Regresión no paramétrica: Una alternativa

LETICIA GRACIA MEDRANO VALDELAMAR
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México

Introducción

La regresión no paramétrica consiste en una serie de técnicas para estimar una curva de regresión sin hacer supuestos fuertes sobre la forma de la verdadera función de regresión. Estas técnicas son más flexibles y resultan muy útiles en la construcción y verificación de los modelos paramétricos.

Existen dos enfoques para los modelos de regresión no paramétrica: uno el de penalización por aspereza de la curva del que resultan los splines y el otro el de estimación local del que resultan los estimadores de kernel y de vecinos cercanos.

Penalización por aspereza de la curva

Bajo este enfoque se tiene el siguiente modelo de regresión:

$$Y = g(t) + error,\ t \in [a, b].$$

La aspereza de la curva g se mide generalmente por:

$$\int_a^b [g''(t)]^2\, dt.$$

Esta cantidad no se ve afectada por la suma de constantes o de una función lineal.

Considerando a g como una función doblemente diferenciable en $[a, b]$, a $\alpha > 0$ como parámetro de suavizamiento y teniendo datos de la forma (t_i, y_i) con $i = 1, ..., n$, la suma de cuadrados penalizados queda expresada como:

$$S(g) = \sum_{i=1}^{n} [y_i - g(t_i)]^2 + \alpha \int_a^b [g''(x)]^2 dx.$$

En esta expresión la sumatoria da un costo de buen ajuste mientras que la integral da un costo por su aspereza o rugosidad.

La función $\hat{g}$ que estima a g es aquella que minimiza $S(g)$ sobre el conjunto de las funciones doblemente diferenciables. Si α es grande $\hat{g}$ tendrá poca curvatura y si es pequeña $\hat{g}$ tenderá a ser una interpolación. Para $n \geq 3$ y $\alpha > 0$, el spline cúbico natural $\hat{g}$ (con nodos en t_i y que interpola los puntos dados por el vector $\underline{g} = (I + \alpha K)^{-1}\underline{y}$ donde la matriz $K = QR^{-1}Q'$ siendo Q y R matrices que dependen exclusivamente de $h_i = t_{i+1} - t_i$ (ver Green y Silverman (1994)) cumple con la siguiente propiedad: $S(\hat{g}) \leq S(g)$ para cualquier g doblemente diferenciable en $[a, b]$.

El algoritmo de Reinsch encuentra este spline en cuatro pasos que realizan un número de operaciones algebraicas del orden de $O(n)$.

Para elegir el parámetro de suavizamiento α se requiere de algún criterio; el más utilizado es el de validación cruzada que consiste en calcular $\hat{g}^{-i}(t,\alpha)$ que es el spline mencionado antes pero sin considerar el dato (t_i, y_i) y se elige el α que minimiza

$$VC(\alpha) = n^{-1}\sum_{i=1}^{n}\{y_i - \hat{g}^{-i}(t,\alpha)\}^2.$$

Es importante comentar que no es necesario calcular n modelos y después calcular $VC(\alpha)$, sino que de manera similar a lo que ocurre en regresión: $\hat{g} = A(\alpha)\underline{y}$, donde $A(\alpha) = I - \alpha Q(R + \alpha Q'Q)^{-1}Q'$ y el valor de $VC(\alpha)$ puede escribirse como

$$n^{-1}\sum_{i=1}^{n}\left\{\frac{Y_i + \hat{g}(t_i)}{1 - A_{ii}(\alpha)}\right\}^2.$$

Estimación local

Los modelos más sencillos son versiones de los estimadores de localización (promedios locales). Se define como estimador de la función g en el punto t^* a:

$$\hat{g}(t^*) = \frac{1}{n^*}\sum_{t_i \in N(t^*)} y_i,$$

donde $N(t^*)$ es una vecindad del t^* y n^* es el número de puntos en la vecindad de t^* y puede reescribirse como:

$$\hat{g}(t^*) = g(t^*) + \frac{1}{n^*}\sum\{g(t_i) - g(t^*)\} + \frac{1}{n^*}\sum_{t_i \in N(t^*)} \varepsilon_i.$$

En esta expresión claramente puede verse que estos estimadores son sesgados pero tienen menor varianza que si se estimara con un solo punto. Los estimadores de kernel utilizan como vecindad de t^* bandas de longitud constante, mientras que los estimadores de vecinos cercanos utilizan bandas que contengan un número constante de puntos. Con el propósito de disminuir el sesgo se utilizan promedios ponderados, con pesos más pequeños para las observaciones lejanas al centro de la banda. Así la expresión para un estimador de kernel queda como:

$$\hat{g}(t^*) = \sum_{i=1}^{n}\frac{w_i y_i}{\sum_{i=1}^{n} w_i},$$

donde $w_i = K\{(t^* - t_i)/\lambda\}$, λ es el ancho de ventana y K es una función continua, acotada, simétrica y que integra uno. Para el caso de los estimadores de vecinos cercanos su expresión es similar pero con $w_i = K\{(r^* - r_i)/(\lambda - 1)\}$ donde r_i es el rango de los puntos t_i y r^* es el rango del punto t^* entre los puntos t^*. En ambos casos $\hat{g}(t) = \sum_{i=1}^{n}\gamma_i(t)y_i$ y se conoce a $\gamma_i(t)$ como pesos del kernel.

Al igual que en el enfoque de penalización por aspereza de la curva donde se requiere la elección de α, en este caso se requiere la elección del tamaño de ventana λ. El criterio de validación cruzada de forma similar puede utilizarse para la elección de λ.

Existe una equivalencia entre los splines y los estimadores kernel; la $\hat{g}$ que minimiza $S(g)$ es lineal en las y_i, y se puede encontrar una función $G(s,t)$ tal que

$$\hat{g}(s) = n^{-1} \sum_{i=1}^{n} G(s,t_i) y_i.$$

Aquí, $G(s,t_i)$ depende de las t_i y de α pero no de y_i.

Es importante mencionar que los métodos de regresión no paramétrica al igual que los métodos de regresión de mínimos cuadrados son sensibles a las observaciones discrepantes e influyentes y esto debe tomarse en cuenta cuando se utilicen.

Estos modelo de regresión no paramétrica son muy útiles para conocer la bondad de ajuste de un modelo. Por ejemplo si el modelo no paramétrico queda dentro de las bandas de error del modelo paramétrico, se considera que éste último se ajusta a los datos. También puede ajustarse un modelo no paramétrico a los residuales (del modelo paramétrico) y ver si siguen alguna tendencia que indique un mal ajuste por parte del modelo paramétrico (ver Altman (1992)). También pueden ser utilizados para sugerir la forma del modelo paramétrico.

Referencias

Altman (1992) "An introduction to kernel and nearest-neighbor nonparametric regression", *Annals of Statistics*, **46**, 3, 175–185

Eubank (1988) *Spline smoothing and nonparametric regression*, Marcel Dekker, Nueva York.

Green y Silverman (1994) *Nonparametric regression and generalized linear models*, Chapman y Hall, Londres.

Hardle (1990) *Applied Nonparametric regression*, University Press, Cambridge.

Hastie and Tibshirani (1990) *Generalized Additive Models*, Chapman y Hall, Londres.

Algunos análisis multivariados a variables antropométricas y de capacidades físicas de niños deportistas

María de Jesús Guijarro Soto & Martín Humberto Félix Medina
Escuela de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa

Introducción

En este estudio, llevado a cabo conjuntamente con la Dirección de Deportes de la U.A.S., se analiza un conjunto de variables antropométricas y de aptitudes físicas medidas en una muestra de niños deportistas (4 a 13 años) de diferentes disciplinas deportivas.

El objetivo de la investigación es caracterizar las disciplinas deportivas con base en las variables ya mencionadas, esto es, determinar si los practicantes de un mismo deporte presentan características antropométricas y de aptitudes físicas similares y diferentes a los practicantes de otros deportes. Con base en ésta caracterización, un futuro niño deportista podrá encausarse al deporte, que de acuerdo a su antropometría y capacidad física, se le pronostique una mayor probabilidad de éxito.

Las variables observadas fueron las siguientes:

Antropométricas:

Edad, peso, estatura, estatura sentado, longitud del brazo derecho, longitud de la pierna derecha, perímetro toráxico y longitud de la brazada.

Flexibilidad:

Flexibilidad de pie, flexibilidad sentado e hiperextensión.

El estudio se llevó a cabo con 53 niños. La clasificación por sexo y disciplina deportiva es la siguiente:

HOMBRES		MUJERES	
Deporte	Individuos	Deporte	Individuos
1. Atletismo	4	1. Atletismo	9
2. Basketball	2	4. Gimnasia	1
3. Baseball	2	5. Karate	1
5. Karate	12	6. Natación	10
6. Natación	6		
7. irregulares	7		

Los análisis estadísticos que hasta el momento se han llevado a cabo son los siguientes: 1).-Componentes principales 2).-Variables canónicas. Cabe mencionar que se tiene pensado incluir dos nuevas variables de aptitudes físicas, como son la fuerza y velocidad, para posteriores análisis.

Análisis de componentes principales (variables antropométricas)

Siguiendo la recomendación de Krazanowski (1988), primero se realizó un análisis por separado para cada sexo y sólo para las variables antropométricas, ya que se deseaba obtener una descripción antropométrica de los diferentes deportes.

Hombres:

Comp.	Eigenvalor	% acum. de la var.
1	6.86	85.77
2	0.55	92.69
3	0.35	97.03

vars	Edad	Peso	Est.	E. sen.	L.brazo	L. pier	P. tor	Braza
c.eig.1	0.351	0.351	0.374	0.368	0.310	0.36 7	0.330	0.372
c.eig.2	-0.224	0.511	-0.175	0.015	-0.417	-0.213	0.649	-0.127

El primer componente es una medida de tamaño, valores grandes corresponden a individuos de tamaño grande. El segundo componente es un contraste entre las variables peso, perímetro, por un lado y edad, pie, brazo, pierna, brazada, por el otro. Refleja la "forma" del cuerpo, valores grandes corresponden a individuos bajos y gordos y valores pequeños a individuos altos y delgados. Los resultados para las mujeres son similares, por lo cual se realizó un análisis para la muestra combinada:

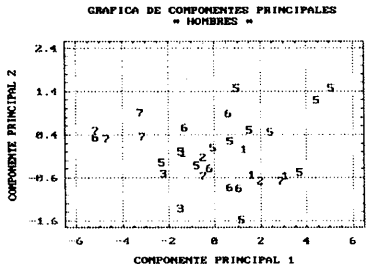
Componente	Eigenvalor	% acumumulado de la varianza
1	6.89	86.13
2	0.52	92.58
3	0.31	96.44

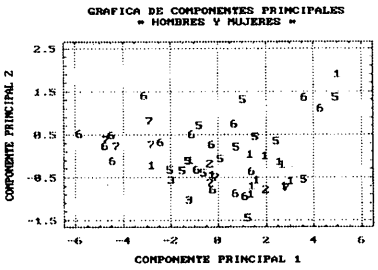
vars	Edad	Peso	Est.	E. sen.	L.brazo	L. pier	P. tor	Braza
c.eig.1	0.352	0.349	0.373	0.361	0.323	0.362	0.331	0.662
c.eig.2	-0.300	0.528	-0.197	-0.008	-0.314	-0.184	0.662	-0.144

La interpretación de los componentes es muy similar a la anterior.

Componentes principales con ambos grupos de variables: antropométricas y de flexibilidad

Hombres			Mujeres		
Comp.	Eigenvalor	% acum. de la var.	Comp.	Eigenvalor	% acum. de la var.
1	6.90	62.80	1	7.64	69.43
2	1.58	77.19	2	1.74	85.31
3	0.72	83.78	3	0.64	91.16





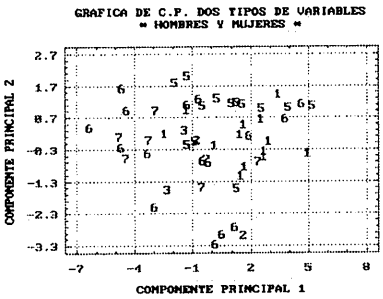
Hombres			Mujeres		
Variables	Coeficientes de los eigenvectores		Variables	Coeficientes de los eigenvectores	
	1	2		1	2
Edad	0.347	-0.203	Edad	0.340	-0.018
Peso	0.349	0.043	Peso	0.327	-0.064
Estatura	0.368	-0.074	Estatura	0.356	-0.006
Est. sent.	0.361	-0.075	Est. sent.	0.338	-0.033
L.brazo	0.307	-0.003	L.brazo	0.321	0.011
L.pierna	0.212	-0.186	L.pierna	0.341	-0.128
P.toráxico	0.332	0.075	P.toráxico	0.308	-0.119
Brazada	0.365	-0.063	Brazada	0.355	-0.001
F. pie	0.182	0.606	F. pie	0.129	0.650
F. sentado	0.002	0.718	F. sentado	-0.035	0.705
Hiperext.	0.274	0.132	Hiperext.	0.279	0.209

Hombres:

El primer componente está asociado al "tamaño", a la flexibilidad de pie, valores grandes corresponden a individuos de tamaño grande y con buena flexibilidad de pie. El segundo componente es una "medida de la flexibilidad", valores grandes corresponden a individuos con buena flexibilidad.

Mujeres:

El primer componente está asociado con el "tamaño", a la flexibilidad de pie y a la hiperextensión; valores grandes corresponden a individuos de tamaño grande y con buena flexibilidad. El segundo componente es una "medida de la flexibilidad", valores grandes corresponden a individuos con buena flexibilidad. Dado que los resultados son muy semejantes para ambos sexos, se realizó un análisis para la muestra combinada, se observaron



resultados semejantes y la interpretación de las componentes principales es similar a las anteriores.

Análisis de variables canónicas

En este análisis se consideran las variables antropométricas y las variables de flexibilidad; los resultados que se obtuvieron son los siguientes:

Hombres			Mujeres		
Comp.	Eigenval.	% acum. de la var.	Comp.	Eigenval.	% acum. de la var.
1	0.894	0.27			
2	0.802	0.50	1	0.972	0.38
3	0.664	0.70	2	0.863	0.72
4	0.570	0.86	3	0.725	1.00
5	0.483	1.00			

Hombres					Mujeres			
Vars.	Eigenval				Vars.	Eigenval		
	1	2	3	4		1	2	3
Edad	-1.23	-1.06	-0.86	0.17	Edad	0.75	0.03	0.82
Peso	-5.65	-1.50	-2.09	-1.19	Peso	-0.69	2.51	2.47
Est.	1.31	0.19	-1.65	0.13	Est.	8.35	1.12	-4.04
Esent.	0.26	3.43	1.11	0.64	Esent.	-4.74	1.37	3.24
Lbrazo	-0.70	0.49	1.36	-0.95	Lbrazo	0.56	0.83	0.27
Lpier	-0.26	0.64	-0.48	-0.43	Lpier	-2.95	-2.92	4.17
Ptórax	4.66	0.63	0.86	0.19	Ptórax	1.19	-1.93	-1.76
Braz.	1.35	-2.47	1.66	1.65	Braz.	-2.80	-1.31	1.17
Fpie	0.69	0.45	-0.23	0.07	Fpie	1.17	-1.82	0.91
Fsen	0.01	-0.28	0.09	0.38	Fsen	-0.25	1.45	-0.10
Hiper	0.75	-0.20	0.24	-0.72	Hiper	0.47	0.23	-0.10

Resulta difícil dar una interpretación intuitiva de estas variables.

Conclusiones

1.- El análisis de componentes principales proporcionó resultados satisfactorios en cuanto a la reducción de "dimensión" de los datos, ya que con dos componentes se explica aproximadamente el 90% de la variabilidad de éstos. Así mismo estos componentes tienen una clara interpretación.

2.-El análisis de variables canónicas no proporcionó resultados muy satisfactorios. En primer lugar, porque se requieren hasta cuatro variables canónicas para representar adecuadamente los datos, además de que éstas no tienen una clara interpretación.

Referencias

Krzanowwski, W. J. (1988) *Principles of Multivariate Analysis,* Clarendon Press, Oxford.

Sobre el análisis bayesiano del problema de clasificación estadística

ANA MARÍA MADRIGAL & MANUEL MENDOZA
Departamento de Estadística y Actuaría, Instituto Tecnológico Autónomo de México

El problema de clasificación ha recibido atención en la literatura estadística desde hace un buen número de años. La mayor parte de estos trabajos abordan el problema desde una perspectiva frecuentista y sólo en condiciones muy particulares garantizan que el problema se resuelva de manera óptima. El enfoque bayesiano reconoce este problema como uno de decisión y establece la existencia general de soluciones óptimas al minimizar la pérdida (costo) esperada.

El problema se presenta cuando se tiene un individuo (I) que:

i) Con seguridad proviene de una y sólo una de m poblaciones $\{\pi_i;\ i = 1, \ldots, m\}$,

ii) Presenta una colección de atributos $(X = X(I))$ cuyo comportamiento, en de cada población, se describe con un modelo de probabilidad distinto $\{P(X|\pi_i);\ i = 1, \ldots, m\}$.

y se desea responder la pregunta : ¿A qué población pertenece I?

Para proceder a la clasificación, con un enfoque bayesiano, es necesario asignar las probabilidades iniciales (o a priori) de que el individuo pertenezca a cada una de las m poblaciones $\{p_i = P(\pi_i);\ i = 1, \ldots, m\}$ *(distribución diagnóstica inicial)* y el costo (o pérdida) que se obtiene al clasificar un individuo que pertenece a la población j como de la población i. Para obtener la solución óptima se calcula la pérdida esperada final *(a posteriori)*,

$$l(i|X) = \sum_j P(\pi_i|X)C(i|j);\ i = 1, \ldots, m,$$

en donde

$$P(\pi_i|X)\alpha P(X|\pi_i)P(\pi_i),$$

son las probabilidades finales de las poblaciones *(distribución diagnóstica final)*. El criterio selecciona la $i - ésima$ población siempre que así se produzca la pérdida esperada mínima.

El problema queda resuelto de esta manera sin tener que hacer ningún supuesto acerca de la forma que tienen las distribuciones $P(X|\pi_i);\ i = 1, \ldots, m$, siempre que sean totalmente conocidas. Es importante notar que el problema puede resolverse también en el caso en que no se han observado los atributos en el individuo y sólo se cuenta con un estado de información inicial (a priori). En cualquier caso, para incorporar los atributos del individuo a la distribución diagnóstica es necesario establecer la forma precisa de la distribución que siguen éstos en cada una de las diferentes poblaciones. En ocasiones las distribuciones de los atributos se suponen completamente conocidas. En tal caso, y como ya se indicó, se obtiene la distribución diagnóstica final (dados los atributos), vía el teorema de Bayes, utilizando la relación:

$$P(\pi_i|X)\alpha P(X|\pi_i)P(\pi_i).$$

Por otra parte, si alguno de estos modelos no está totalmente especificado (el caso mas frecuente) entonces existe incertidumbre sobre el comportamiento de X en las diferentes

poblaciones a tratar. Se presupone que este comportamiento será distinto entre poblaciones, pero se puede pensar en algún modelo paramétrico común para describir este comportamiento de forma que el correspondiente parámetro (digamos θ) cambie dependiendo de la población de que se trate. En esta situación, se tiene una colección de distribuciones $\{P(X|\pi_i, \theta_i); i = 1, \ldots, m, \}$ que pertenecen a la misma familia paramétrica y en donde θ_i describe el comportamiento específico del atributo X en la población i.

Si la forma de las distribuciones se considera conocida, la incertidumbre sobre los modelos se reduce a la incertidumbre sobre el valor de los parámetros $\{\theta_i;\ i = 1, \ldots, m\}$ y desde una perspectiva bayesiana, es necesario asignarles algunas distribuciones de probabilidad

$$P(\theta_i \mid \pi_i);\ i = 1, \ldots, m,$$

que describan el conocimiento que se sobre ellos se tiene. En cualquier caso, para resolver el problema original, se obtiene una colección de distribuciones predictivas para la X:

$$P(X|\pi_i) = \int P(X \mid \pi_i, \theta_i) P(\theta_i \mid \pi_i) d\theta_i$$

donde $P(\theta_i|\pi_i)$ es una distribución a priori que se asigna a θ_i.

Es importante notar que la solución del problema depende crucialmente de la asignación de las distribuciones iniciales. Estas distribuciones deben reflejar el conocimiento que el tomador de decisiones posea sobre las distintas fuentes de incertidumbre y pueden actualizarse si se cuenta con información adicional. En particular, si se obtienen observaciones que provengan de las poblaciones involucradas en la clasificación, las distribuciones pueden modificarse para incorporar ese conocimiento. La obtención de observaeiones adicionales (Z) a través de una muestra puede ofrecer dos tipos de información:

i) Si se toma una muestra retrospectiva se obtiene información acerca del comportamiento de los parámetros θ_i en las diferentes poblaciones con lo cual se puede obtener $P(X|\pi_i, Z)$ que sustituya a $P(X|\pi_i)$;

ii) Por otra parte, si el muestreo es prospectivo se obtiene, además, información acerca de las prevalencias de las poblaciones $\pi_1, \ldots, \pi_m$, en la super población con lo que se puede obtener $P(\pi_i|Z)$ que sustituya a $P(\pi_i)$

La distribución diagnóstica a posteriori $\{P(\pi_i|X, Z);\ i = 1, \ldots, m\}$ sustituirá en cualquier caso a la distribución diagnóstica a priori $\{P(\pi_i|X);\ i = 1, \ldots, m\}$ para dar una solución al problema de clasificación. Si por cualquier razón no se conoce ningún atributo del individuo que ayude a la clasificación, pero se obtiene nueva información sobre las prevalencias, la distribución diagnóstica a utilizar para resolver el problema es $P(\pi_i|Z)$. Para obtener $P(X|\pi_i, Z)$, primeramente se obtiene una distribución a posteriori para θ_i:

$$P(\theta_i|\pi_i, Z_i) \alpha P(Z_i|\theta_i, \pi_i) P(\theta_i|\pi_i),$$

en donde $P(Z_i|\theta_i, \pi_i) = \prod_{j=1}^{n_i} P(X_{ij}|\theta_i)$. Finalmente, se obtiene $P(X|\pi_i, Z_i)$ de la expresión

$$P(X|\pi_i, Z_i) = \int P(X \mid \pi_i, \theta_i) P(\theta_i \mid \pi_i, Z_i)\, d\theta_i.$$

Como ya se indicó, un muestreo retrospectivo no proporciona información sobre la prevalencia de las distintas poblaciones, por lo que $P(\pi_i|Z) = P(\pi_i)$. Sin embargo, si la muestra contiene datos prospectivos, la información relevante adicional consiste en haber obtenido en la muestra $n_1, n_2, \ldots, n_m$ individuos $(n_1 + n_2 + \cdots + n_m = n)$ que provienen de las poblaciones $\pi_1, \pi_2, \ldots, \pi_m$ respectivamente (ver Bernardo, 1985).

La probabilidad de que el individuo (I) provenga de π_j se puede modelar siempre de la misma manera (sin importar la distribución de X) con ayuda de una variable auxiliar δ_j cuya distribución sea Bernoulli(ϕ_j) de tal manera que

$$\delta_j = \begin{cases} 1 \text{ si } I \text{ pertenece a } \pi_j \\ 0 \text{ en otro caso} \end{cases}$$

por lo que $\phi_j = P(\delta_j = 1|\phi_j) = P(\pi_j|\phi_j)$, en donde ϕ_j es un parámetro desconocido que representa la verdadera prevalencia de π_j en la superpoblación. Un modelo de este estilo permite incorporar fácilmente la información adicional. Así, ya que lo que se quiere tener es $P(\pi_j|Z) = P(\delta_j = 1|Z)$, es necesario encontrar la siguiente distribución predictiva

$$P(\delta_j \mid Z) = \int P(\delta_j \mid \phi_1, \ldots \phi_m)\, P(\phi_1, \ldots \phi_m \mid Z)\, \partial\phi_1, \ldots \partial\phi_m.$$

Para ello, debe obtenerse $P(\phi_1, \phi_2 \ldots \phi_m|Z)$ vía teorema de Bayes de la siguiente manera:

$$P(\phi_1, \phi_2, \ldots, \phi_m|Z) \propto P(Z|\phi_1, \phi_2, \ldots, \phi_m) P(\phi_1, \phi_2, \ldots, \phi_m).$$

Tomando en cuenta que $0 < \phi_i < 1$ $(i = 1, \ldots, m)$, y que la distribución conjunta de $n_1, \ldots, n_m$ se puede modelar con una distribución Multinomial $(\phi_1, \phi_2, \ldots, \phi_m)$, utilizando familias conjugadas se tiene que

$$P(\phi_1, \phi_2, \ldots, \phi_m) = \text{Dirichlet}_m(\phi_1, \phi_2, \ldots, \phi_m|\alpha_1, \alpha_2, \ldots, \alpha_m)$$

y en consecuencia,

$$\begin{aligned} P(\phi_1, \phi_2 \ldots \phi_m|Z) &= \text{Dirichlet}_m(\phi_1, \phi_2 \ldots \phi_m|n_1 + \alpha_1, n_2 + \alpha_2, .., n_m + \alpha_m) \\ &= \text{Dirichlet}_m(\phi_1, \phi_2 \ldots \phi_m|\alpha_1', \alpha_2', .., \alpha_m') \end{aligned}$$

en donde naturalmente, $\alpha_i' = \alpha_i + n_i$; $i = 1, \ldots, m$. Como consecuencia, se obtiene que la distribución marginal de ϕ_j dado Z, se distribuye como una Beta(α_j', β_j') con

$$\beta_j' = \sum_{i \neq j} \alpha_i' \; ; \; 1 = 1, ..m$$

(Bernardo & Smith, 1994) y por lo tanto,

$$P(\pi_j|Z) = P(\delta_j = 1|Z) = \frac{n_j + \alpha_j}{n + \sum_{j=1}^{m} \alpha_j}.$$

Un problema interesante aquí es que los parámetros $\{\alpha_i;\ i = 1, \ldots, m\}$ debieran ser tales que se ajusten al conocimiento inicial sobre las prevalencias $\{p_i;\ i = 1, \ldots, m\}$, es decir, que

el modelo incorpore la información a priori por lo que se tiene una restricción para cada una de las m poblaciones. Una manera natural, en términos predictivos, es que las prevalencias a priori coincidan con los valores esperados de las distribuciones de $\{\alpha_i;\ i = 1, \ldots, m\}$. Esto es, que $p_i = \alpha_i / \sum \alpha_i\ ;\ i = 1, \ldots, m$. Estas restricciones no determinan por completo el valor de los parámetros. Es decir, hay varios conjuntos de α's que las satisfacen.

Ya que en esta situación un valor de α mayor implica una mayor certeza en el conocimiento (menor varianza), se puede pensar que si el conocimiento a priori se pudiera equiparar con cierto tamaño de muestra, es decir, si se puede pensar que la calidad de la información inicial equivale a haber observado una muestra de tamaño k, se puede tomar en cuenta la siguiente expresión:

$$P(\pi_i \mid Z) = \frac{\alpha + k_i}{\alpha + \beta + k},$$

donde $s = k/\alpha + \beta + k$ pondera la información que se obtiene vía la muestra de tamaño k y k_i es el número de individuos en la muestra que pertenecen a la población i. Así, si se le quiere dar el mismo peso a la información inicial que el que tendrá una muestra de tamaño k (i.e. $s = 1 - s$), se obtiene la relación

$$\sum_{i=1}^{m} \alpha_i = k,$$

de donde se obtiene la solución $\alpha_i = k_{pi}$ para $i = 1, \ldots, m$. Si se piensa que la información que se tiene a priori es vaga comparada con la información que la muestra puede proporcionar, una opción es encontrar el punto $(\alpha_1, \alpha_2, \ldots, \alpha_m)$ de menor distancia al punto $(1/2, \ldots, 1/2)$ que, en este caso, corresponde a la distribución de referencia global que se obtiene a partir de la regla de Jeffreys (Box y Tiao, 1973).

Esta solución es simple y se obtiene sin dificultad cuando $m = 2$. Sin embargo, cuando m es un natural arbitrario, la determinación de los parámetros $\alpha_1, \ldots, \alpha_m$ de referencia puede resultar una tarea laboriosa. Por otra parte, cabe hacer notar que cualquier solución que se obtenga de esta manera estar en el cuadrado unitario, ya que $0 < \alpha_j < 1$, para $j = 1, \ldots, m$. De este modo, basta una muestra pequeña (10 o 15 individuos) para que esta información a priori prácticamente no influya en la distribución a posteriori.

Si se adopta esta estrategia puede concluirse que el enfoque bayesiano ofrece un procedimiento para combinar información previa (experimental y/o subjetiva) con atributos de un individuo para llevar a cabo la clasificación. El procedimiento es óptimo, cualesquiera que sean los modelos que describan el comportamiento de los atributos sin importar si están completa o parcialmente especificados. En cualquier caso, se utiliza la distribución diagnóstica correspondiente al nivel de información de acuerdo a ala siguiente regla

	Sin atributos	Con atributos
Sin inf. adicional	$P(\pi_i)$	$P(\pi_i \mid X)$
Con inf. adicional	$P(\pi_i \mid Z)$	$P(\pi_i \mid X, Z)$

Finalmente, para incorporar información adicional que proviene de un muestreo prospectivo es conveniente parametrizar la distribución diagnóstica $P(\pi_i)\ ;\ i = 1, \ldots, m$. Los parámetros $\alpha_j\ ;\ i = 1, \ldots, m$ de la distribución *Dirichlet* correspondiente, aun cuando reconstruyan la distribución diagnóstica a priori no quedan unívocamente determinados. En

este trabajo se han presentado algunas ideas para determinar los valores de referencia α_i para $i = 1, \ldots, m$ que resultan muy simples desde un punto de vista computacional.

Referencias

Bernardo, J.M. (1985) "Diagnóstico automático en Medicina", *Estadística Española*, 108.

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

Box, G.E.P. and Tiao, C.T. (1973) *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Massachusetts.

Filosofía y estadística aplicada

IGNACIO MÉNDEZ RAMÍREZ
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México

Estadística aplicada

Se intenta una caracterización de la estadística aplicada, sin pretender dar una definición. "Es la representación de la realidad con modelos probabilísticos". Otras caracterizaciones de la estadística son:

Uso de la estadística

La estadística se ha convertido en una forma de pensar y una "herramienta" muy poderosa en muchas áreas de actividad humana.

a En investigación científica se ha constituido en una parte muy especial de "el método científico"; cuando es necesario tomar en cuenta la variación por errores de medida o por la falta de uniformidad en los elementos de estudio, o ambos.

b En los procesos productivos y de servicios, su uso es fundamentalmente en la captación de información y optimización de intervenciones.

c En las acciones y decisiones de gobierno de países, estados, instituciones, etc.; se usa fundamentalmente en la captación de información para el diagnóstico y evaluación del estado del país, estado, institución, etc. También para la optimización de acciones y de programas.

Profundizando un poquito más en el papel de la Estadística en la investigación científica podemos proponer tres tipos fundamentales de participación. En el diseño de la investigación, en el proceso de obtener generalizaciones empíricas y en el de conceptualización.

i **Diseño.** La Estadística apoya el diseño de la investigación, por lo menos en: definición de elementos de estudio; características generales; criterios de inclusión y de eliminación; definición, control y vigilancia de la validez interna y externa; selección de la muestra, evitar sesgos (se entiende por sesgo (bias) un error sistemático) de selección; tamaño de muestra; determinar qué, cuándo, cómo y con qué medir, cuántas veces; validez y confiabilidad de las mediciones; eliminación de sesgos durante la conducción; planear el análisis estadístico.

ii **Inferencia.** La inferencia estadística permite concluir de la(s) muestra(s) a la(s) poblaciones. Se requiere un análisis descriptivo exploratorio, previo a la postulación de modelos. La inferencia puede ser para someter a contrastación hipótesis científicas, para estimar parámetros, para buscar relaciones entre variables. Se puede hacer con

diferentes tipos de modelos, y con diferentes supuestos. Siempre debe haber una validación de los modelos, antes de la interpretación. Los métodos estadísticos a usarse dependen del diseño de investigación, del objetivo del estudio y de la teoría que sustenta el fenómeno en cuestión.

iii **Conceptualización.** La Estadística también juega un papel importante en el proceso de conceptualización. Por lo menos para: valorar la validez y confiabilidad de los indicadores de los conceptos; buscar indicadores de conceptos que más se asocian con otros; permitir disminuir la dimensionalidad de un problema; ayudar a formar clasificaciones empíricas o validar las teóricas; permitir inferir conceptos o factores en base a sus correlatos empíricos; poder medir la "forma" de los fenómenos y así tipificarlos; estimular la creatividad para el proceso de conceptualización e invención de teorías.

J. Nelder (1992) considera que la forma en que se enseña y aplica la Estadística minimiza o elimina la importancia del diseño. También G. Box (1993) considera que el objetivo de la Estadística es la mejoría y catálisis de la investigación científica. Sin embargo, también señala que por lo menos en Estados Unidos el reclutamiento, enseñanza y apoyo de los estadísticos parece diseñado en contra del objetivo anterior. El considera que esto surge por categorizar la Estadística como parte de las ciencias matemáticas. Justamente la tesis de este trabajo es que estos aspectos referidos suceden por no conceptualizar adecuadamente el papel de la Estadística en la investigación científica a la luz de la nueva filosofía de la ciencia.

Modelación

De manera general la **matemática** consiste de abstracciones que se refieren a entidades, con ciertas propiedades, relaciones y axiomas de los que se deducen teoremas que constituyen la teoría. Con esto se construyen los modelos matemáticos. En la **realidad** existen entidades (células, personas, familias, barras de hierro, etc.) que tienen propiedades que pueden ser consideradas como debidas a algunos conceptos teóricos (calor, adaptabilidad, pobreza, agresión, represión, etc.) Además se establecen asociaciones que pueden ser debidas a la causalidad de unas propiedades determinando otras. Con esto se construye la teoría sustantiva de la física, química, biología, medicina, ingeniería, etc.

El papel de la Estadística como parte de las matemáticas aplicadas consiste en buscar una buena correspondencia entre los conceptos abstractos de la matemática y las características del fenómeno de la realidad al que se pretende aplicar. Así los modelos pueden ser modificados para ajustarse mejor a la realidad o bien el diseño de investigación, también se puede modificar para que finalmente el modelaje de la realidad mediada por el diseño sea satisfactorio, es decir que produzca errores considerados de poca importancia.

Método científico y estadística

Se considera que no hay un método científico infalible y obligado, sin embargo, hay algunas guías o pasos que conviene seguir. Hay muy variadas versiones de cuales son estos pasos, a continuación presentamos una de ellas, con la aclaración que debe buscarse una realimentación que procure la máxima coherencia entre las diferentes etapas. Las etapas son: Problema, Objetivos, Justificación, Marco Teórico-Empírico, Variables, Hipótesis, Diseño, Conducción, Análisis, Interpretación y Discusión, Conclusiones y Reporte.

Se comentó que la Estadística puede jugar un papel importante en el diseño de la investigación y en la generación y valoración de conceptos. También en el análisis, en donde una propuesta de forma de operar consiste de los siguientes elementos:

a Bases de datos. Especificar con claridad los elementos y las variables asociadas a ellos, procurar evitar errores de transcripción.

b Análisis descriptivo inicial. Es muy conveniente llevar a cabo un análisis numérico y gráfico que describe las características de la información, por el momento sin el planteamiento de modelos estadísticos. Conviene utilizar primero variables en forma aislada y después contemplar la descripción conjunta de dos o tres y raramente cuatro variables en forma simultánea. Esto es para poder tener una idea de las características de la información, patrones de variabilidad, posibles observaciones atípicas o influyentes, así como la presencia de interacciones.

c Propuestas de modelos. Conviene intentar utilizar varios modelos que se eligen a la luz de los objetivos de la investigación, al diseño, las escalas y papel de las variables (pueden ser independientes o causales, dependientes o efectos, factores de confusión o intervinientes). En estudios menos estructurados todas las variables tienen el mismo papel, de tipo exploratorio descriptivo. Es importante para cada modelo explorar los supuestos que ese modelo requiere a partir de la teoría que sustenta el modelo y del diseño de investigación.

d Pruebas de Modelos. Se lleva a cabo el ajuste de los modelos propuestos, para seleccionar de ellos el que tenga resultados más acordes con las consideraciones teóricas y con mayor correspondencia entre los supuestos del modelo y las características de la realidad reflejada en los datos, tomando en cuenta el diseño de investigación.

Interpretación

Con el o los modelos adecuados de acuerdo a las consideraciones anteriores, se procede a una interpretación que puede ser para predecir el comportamiento del fenómeno en estudio, para explicarlo, para construcción de teoría o para la contrastación de hipótesis estadísticas.

Referencias

Box, G. (1993) *The Royal Statistical Society. News and Notes*, **21**, No. 4.

Nelder, J. (1992) *The Royal Statistical Society. News and Notes.* **18**, No. 8.

Uso de un modelo de regresión en dos niveles para estudiar curvas de crecimiento

MARIO MIGUEL OJEDA & J. A. MONTANO-RIVAS
Facultad de Estadística, Universidad Veracruzana

Introducción

En este trabajo, se formula un problema de curvas de crecimiento en genética forestal en términos de un modelo de regresión en dos niveles. Se hace una serie de comentarios sobre resultados de estimación en este contexto, y se obtienen las curvas estimadas.

El análisis de datos de curvas de crecimiento ha recibido gran atención en la literatura dedicada a la biometría. La modelación de datos de este tipo se aborda a través del modelo lineal general de curvas de crecimiento. Para un panorama de la teoría y metodología de este tipo de modelos ver Seber (1984) y referencias que allí se recomiendan.

En esencia el modelo de curvas de crecimiento considera un polinomio de orden k sobre el tiempo para modelar el comportamiento del individuo descrito por una variable respuesta. En el caso de tener varios grupos de individuos esta metodología permite comparar los patrones de crecimiento (curvas promedio). Sin embargo algunas veces más que el estudio de las curvas promedio interesa estudiar la variabilidad de las curvas individuales; esto es, las varianzas y las covarianzas entre los coeficientes de los polinomios sobre el tiempo. Tal es el caso del problema de genética forestal que presentamos a continuación.

Ejemplo. Se desea estudiar comparativamente el comportamiento del crecimiento de 37 familias de Pinus Patula, para lo que se seleccionaron semillas y se pusieron a germinar durante un mes. Posteriormente se tomaron 10 plantas y se transplantaron a un vivero en condiciones controladas donde se tomaron mediciones de la altura de cada planta cada mes. La Figura 1 presenta los comportamientos promedio de las familias sobre el tiempo, para las que además se registró la desviación standard en la última fecha (W_1).

En este trabajo presentamos un modelo de regresión en dos niveles para describir los efectos fijos y aleatorios del crecimiento de las familias. Este modelo es un caso particular del presentado por Ojeda y Juárez-Cerrillo (1994), y permite la modelación de los coeficientes que describen las curvas de crecimiento. Se obtienen estimaciones de los efectos fijos y aleatorios usando el método de mínimos cuadrados generalizados reponderados iterativamente Goldstein (1986).

El modelo y procedimiento de estimación

Sea y_{ij} el promedio de la variable respuesta, altura (Y), para la familia i-ésima en el tiempo t. El patrón de crecimiento será modelado por la ecuación cuadrática

$$y_{ij} = \beta_{oi} + \beta_{1i}\mathbf{t}_{ij}\beta_{21}t_{ij}^2 + e_{ij} \tag{1}$$

donde $e_i^t = (e_{i1}, e_{i2}, \ldots e_{in_i})$ es el vector de errores aleatorios para la familia i-ésima ($i = 1, 2, \ldots, n$), con $E(e_i) = 0$ y $V(e_i) = \sum$. Se asume que estos errores son independientes de

familia a familia.

Para estudiar la variabilidad de los coeficientes de la ecuación cuadrática en (1) proponemos, para el segundo nivel, los siguientes modelos

$$\begin{aligned} \beta_{oi} &= \beta_{oo} + \beta_{o1} w_{2i} + u_{oi} \\ \beta_{1i} &= \beta_{1o} + \beta_{11} w_{2i} + u_{1i} \\ \beta_{2i} &= \beta_{2o} + \beta_{21} w_{22i} + u_{2i} \end{aligned} \tag{2}$$

para los que suponemos una distribución conjunta determinada por el supuesto distribucional sobre los errores aleatorios en el segundo nivel, que establece que $E(u_i) = 0$ y $V(u_i) = \Omega$, donde $u_i^t = (u_{0i}, u_{1i}, u_{2i})$. A Ω se le llama la matriz de varianzas y covarianzas en el segundo nivel, y esencialmente presenta los componentes de la varianza y la covarianza asociados a la variabilidad entre los coeficientes de la ecuación cuadrática.

Sustituyendo las ecuaciones (2) dentro de la ecuación (1) obtenemos una formulación particular del modelo de efectos mixtos. Para la obtención de las estimaciones de los efectos fijos y aleatorios de este modelo Goldstein (1986) propuso e implementó el algoritmo de mínimos cuadrados generalizados reponderados iterativamente, en el cual se basa el sistema *ML3E* Proser et al. (1990). Este método consiste en estimar la estructura de la matriz de varianzas y covarianzas a través de los residuos obtenidos en el ajuste anterior y después implementar el método de mínimos cuadrados generalizados para estimar los efectos fijos del modelo. Este procedimiento se inicia obteniendo los residuos de un ajuste de mínimos cuadrados ordinarios y se detiene hasta que se logra un criterio de convergencia numérica. Para detalles computacionales ver Goldstein y Rasbash (1992).

Resultados

Tabla 1 incluye los resultados obtenidos de realizar el ajuste del modelo que se obtiene de combinar las ecuaciones (1) y (2) presentadas en la sección 2. Se presenta la estimación de los parámetros fijos y de los aleatorios; también sus desviaciones standard. Se ajusta el modelo sin considerar la variable W_1, ya que su contribución al ajuste resultó no significativa. Como puede verse en la Figura 2 las curvas ajustadas observan una mayor variabilidad en la parte del crecimiento final; tal aspecto es claro en la matriz de varianzas y covarianzas del segundo nivel (ver Tabla 1).

	Efectos Fijos	Aleatorios β_{oo}	β_{10}	β_{20}
Constante	4.205	0.353		
(β_{oo})	(0.195)	(0.340)		
Tiempo	−1.444	−0.469	0.447	
(β_{10})	(0.140)	(0.232)	(0.168)	
Tiempo	0.345	0.128	−0.103	0.022
(β_{20})	(0.026)	(0.042)	(0.031)	(0.006)

Tabla 1. Estimaciones de los efectos fijos y aleatorios asociados al modelo que resulta de combinar las ecuaciones (1) y (2). Las desviaciones standard asociadas se presentan entre paréntesis.

Discusión y conclusiones

En los resultados observamos que la variación del término independiente en el polinomio no es significativa, pero si los coeficientes asociados con el término lineal y cuadrático en el tiempo. Por otro lado las covarianzas entre los coeficientes aleatorios resultaron todas significativas.

Al analizar el ajuste de las curvas notamos que el efecto cuadrático produce una caída del crecimiento entre el primero y tercer mes, lo que en general no se corresponde a la forma del crecimiento en este período, por lo que tal vez deberá ensayarse un modelo no lineal o uno susceptible de linealizarse, y realizar comparaciones con los resultados obtenidos aquí.

Agradecimientos

Agradecemos a Lilia Baizabal, del Centro de Genética Forestal de la Universidad Veracruzana, el habernos proporcionado sus datos para este ensayo de metodología estadística. Este trabajo forma parte del proyecto 4066-A9404 registrado ante CONACYT.

Referencias

Goldstein H. (1986) "Multilevel mixed linear model analysis using iterative generalised least squares", *Biometrika*, **73**, 43–56.

Goldstein H. and Rasbash J. (1992) "Efficient computational procedures for the estimation of parameters in multileved model based on iterative generalised least squares", *Comp. Stat. and Data Analysis*, **13**, 63–71.

Ojeda M.M. and Juárez-Cerrillo S.F. (1994) "Biplot display for diagnostic in a two-level regression model for growth curves analysis", paper prested in *The XVIIth International Biometric Conference,* 8–12, Hamilton, Canadá

Prosser R. Rashash J. and Goldstein H. (1990) *ML3: software for three level analysis*, Institute of Education, University of London.

Seber G. A. (1984) *Multivariate observations*, Wiley, New York.

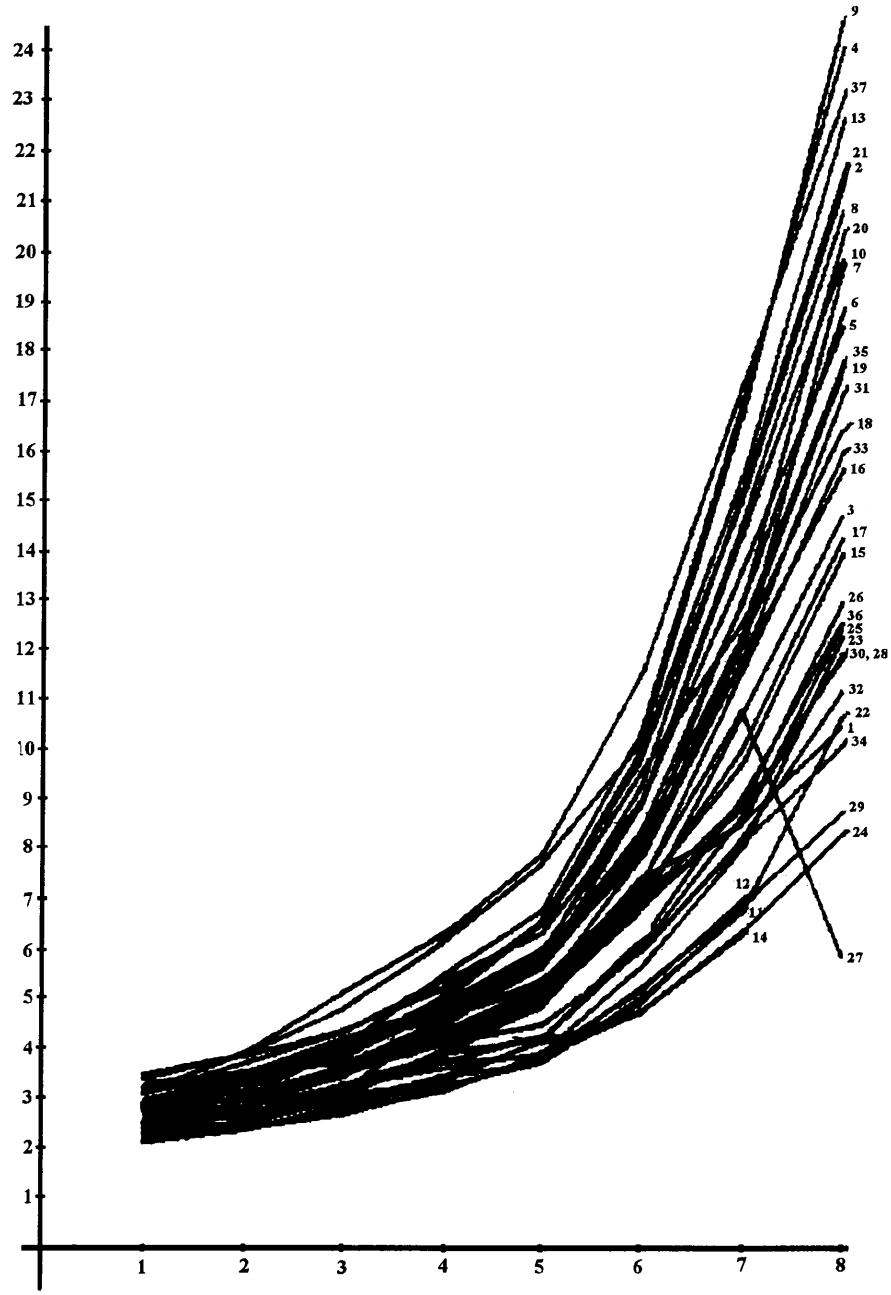


Fig. 1. Perfiles de crecimiento de las 37 familias de Pinus Patula.

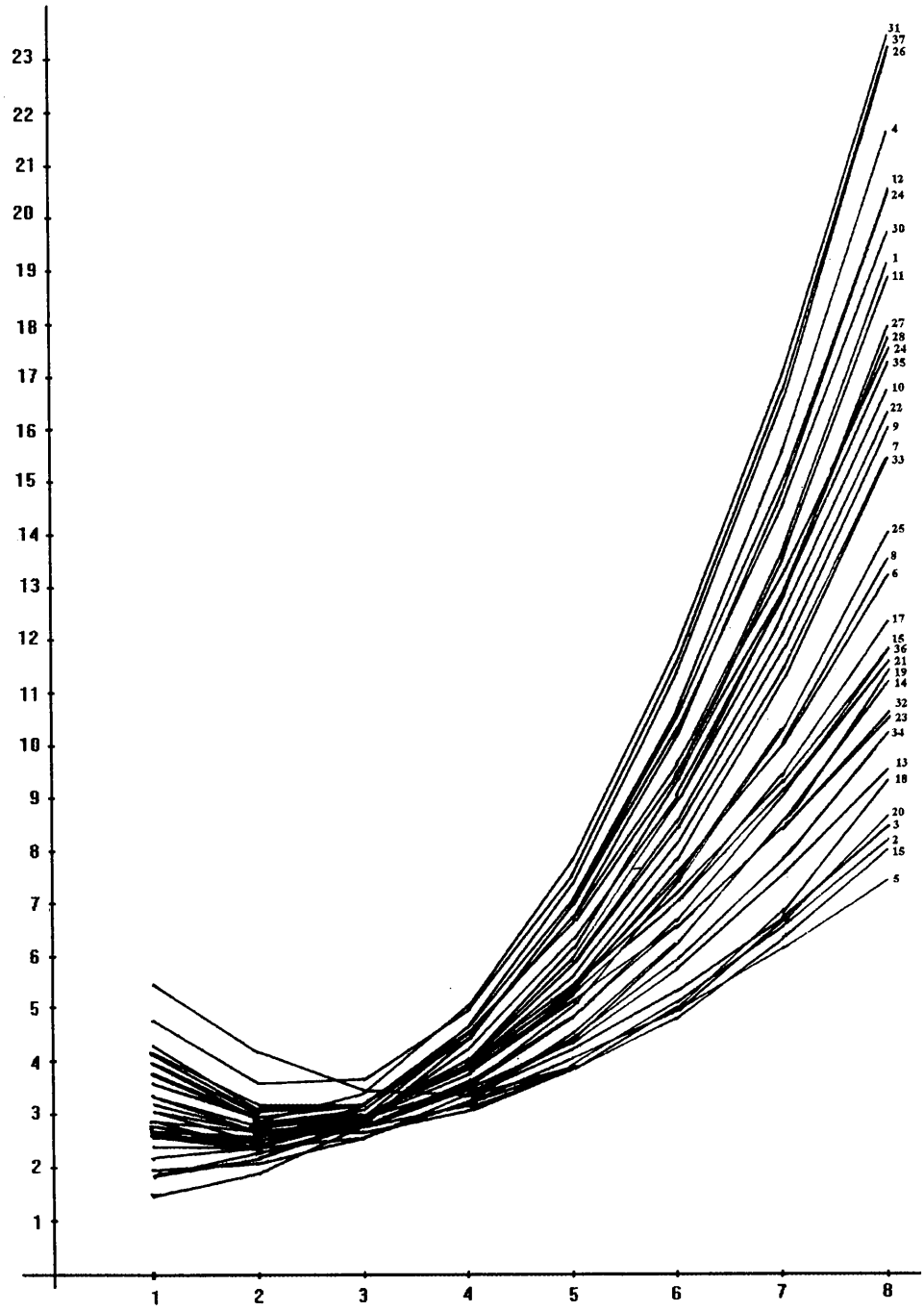


Fig. 2. Perfiles de crecimiento para las 37 familias de Pinus Patula.

Componentes de varianza estimadas de un diseño de bloques al azar con arreglo factorial combinatorio y partición de efectos

EMILIO PADRÓN CORRAL
Centro de Investigación en Matemáticas Aplicadas, Universidad Autónoma de Coahuila

El desarrollo agrícola de un país se basa en la investigación y como el nuestro no es la excepción, se llevó a cabo el siguiente trabajo, donde fue muy interesante investigar el comportamiento de los genotipos de maíz en diferentes localidades o ambientes, pero más aún se checó el verdadero efecto de sus varianzas contando para ello con la herramienta fundamental de las componentes de varianza estimadas, la teoría presentada en este trabajo se aplicó a un experimento de campo, descrito en Soto (1990).

Esta investigación forma parte del programa de mejoramiento genético del Instituto Mexicano del Maíz (Mario E. Castro Gil) de la UAAAN. Consta de dos localidades: Río Bravo, Tamaulipas y Celaya, Guanajuato. En doscientos catorce híbridos, tanto experimentales como testigos.

El modelo utilizado fue.

$$Y_{ijk} = \mu + L_k + R_{j(k)} + G_i + (LG)_{ki} + E_{ijk}$$

donde

$i = \quad 1, 2, 3, \ldots, t$ genotipos
$j = \quad 1, 2, 3, \ldots, r$ repeticiones
$k = \quad 1, 2, 3, \ldots, l$ localidades

Y_{ijk} : Variable aleatoria observable de la k-ésima localidad en la j-ésima repetición del i-ésimo genotipo
μ : Media general
L_k : Efecto de la k-ésima localidad
$R_{j(k)}$: Efecto de la j-ésima repetición dentro de la k-ésima localidad
G_i : Efecto del i-ésimo genotipo
$(LG)_{ki}$: Efecto de la k-ésima localidad y del i-ésimo genotipo
E_{ijk} : Componente aleatoria asociada con la k-ésima localidad en la j-ésima repetición del i-ésimo genotipo

Asumiendo que los efectos son normales, aleatorios e independientes, esto genera el modelo infinito o modelo II de Einsenhart, por lo tanto de acuerdo a dicho modelo, el cuadrado medio de la interacción Localidad-Genotipo es el apropiado cuadrado medio del error para probar genotipos. Además se asume que las esperanzas de efectos son cero, es decir.

$$E[L_k] = E[R_{j(k)}] = E[G_i] = E[(LG)_{ki}] = E[E_{ijk}] = 0.$$

También se asume que las estimaciones de componentes de varianza de productos cruzados de los diferentes efectos son cero, se tiene además que

$$E[L_k^2] = \sigma_L^2;\ E[R/L]^2 = \sigma_{r\subset L}^2;\ E[G_i^2] = \sigma_G^2;\ E[(LG)_{ki}^2] = \sigma_{LG}^2;\ E[E_{ijk}^2] = \sigma_e^2.$$

Del modelo dado anteriormente se encontraron las estimaciones de componentes de varianza de cada uno de sus efectos e interacciones, así como todas sus particiones.

También se supone que la esperanza de los dobles productos son iguales a cero.

Antes de continuar se definirán algunos conceptos agronómicos referentes a la partición de los tratamientos.

Genotipos =	individuos, plantas, animales etc.(En este caso se usaron plantas)
Cruzas =	Serie de nuevas plantas sometidas a ensayo con una característica deseable conocida.
Probadores =	Plantas con características deseables ya conocidas por pruebas estadísticas y agronómicas preliminares
Gen/p_1 =	Número de plantas que se están probando en cruza con la característica germoplásmica uno
g_1/p_1 =	Número de plantas de la población uno que provienen de la característica germoplásmica uno.
g_2/p_1 =	Número de plantas de la población dos que provienen de la característica germoplásmica uno
(g_1/p_1 vs g_2/p_1) =	Contraste o grado de potencialidad entre las plantas de población uno dentro de la característica uno y las plantas de población dos dentro de la característica uno.
Testigos =	Plantas comerciales (Que ya están en el mercado).
Cruza vs Testigo =	Este es un contraste que mide el grado de potencialidad entre cruzas nuevas y las cruzas comerciales.

Todos estos efectos ya mencionados se interactúan con localidad con el objeto de analizar su contribución correspondiente.

A continuación se obtienen las estimaciones de componentes de varianza, para esto se empieza con la estimación de la suma de cuadrados y en ella se generan los grados de libertad de ese efecto y después dividimos la suma de cuadrados entre sus respectivos grados de libertad, formándose la esperanza del cuadrado medio o componente de varianza estimada.

En seguida se resume, presentando solo los valores de las componentes de varianza estimadas de efectos principales e interacciones, para la variable rendimiento de mazorca en ton/ha del experimento de Soto (1990).

F.V.	**C.V.E.**
Localidad	50.0817
Rep/Loc	0.0518598
Genotipos	0.6645
Gen x Loc	0.82
Error	2.108
Total	53.6525298

En este trabajo es interesante no sólo el obtener la magnitud relativa de las varianzas, sino también los porcentajes de la suma de las varianzas de cada uno de los efectos, como a continuación se dan.

σ^2_e	**representa**	$\frac{2.108\times100}{53.6525}$	**3.93%**
σ^2_L	representa	$\frac{50.00817\times100}{53.6525}$	93.21%
$\sigma^2_{r\subset L}$	representa	$\frac{0.0518598\times100}{53.6525}$	0.096%
σ^2_G	representa	$\frac{0.6645\times100}{53.6525}$	1.238%
σ^2_{GL}	representa	$\frac{0.82\times100}{53.6525}$	1.53%

En base a las componentes de varianza estimadas y a futura explotación comercial en diferentes localidades se espera encontrar la mejor combinación que facilite al investigador genetista seleccionar nuevos genotipos con mayor grado de confiabilidad.

También se espera que el desarrollo de esta metodología y sus aplicaciones sean de utilidad a los estudiantes de la maestría en estadística experimental y de otras especialidades, así como a maestros investigadores de la UAAAN que son los más directamente involucrados por su inmediata aplicación en los trabajos de investigación que ellos desarrollan.

Referencias

Einsenhart, CH. (1947) "Las Suposiciones del Análisis de Varianza", *Biometrics*, **3**.

Searle, S.R. (1987) *Linear Models for Unbalanced Data*, John Wiley and Sons, Inc., New York.

Federer, W.T. (1955) *Experimental Design*, The Macmillan Company, New York.

Searle, S.R. (1971) *Linear Models*, John Wiley and Sons, Inc. New York.

Johnson, N.L. and Leone, F.C. (1977) "Statistics and Experimental Design in Engineering and The Physical Sciences", **II**, Second edition, John Wiley and Sons, Inc. New York.

Brownlee, K.A. (1984) *Statistical Theory and Methodology*, In Science and Engineering, Second edition, Robert E. Krieger Publishing Company, Inc.

Soto, S.V. (1990) *Comportamiento de las líneas tropicales "AN_1" y "AN_2" de maíz (Zea mays L.) recobradas por selección gamética en cruza con cuatro probadores de reducida base genética,* Tesis Licenciatura, U.A.A.A.N.

Estimación de funciones de acumulación de especies

FELIPE DE JESÚS PERAZA GARAY
Escuela de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa

Introducción

Supongamos que una población de objetos está dividida en un número desconocido de clases. El objetivo primordial es modelar estadísticamente la relación que guardan el número total de clases identificadas, y el esfuerzo dedicado a su identificación.

En estudios de ecología, una *curva de acumulación de especies* es una gráfica del número de especies encontradas, $S(n)$, en una región particular, como función de alguna medida n del esfuerzo dedicado a encontrarlas.

Soberón & Llorente (1993) consideran que la curva de acumulación puede modelarse como un proceso de nacimientos puros, definiendo tres funciones de transición basadas en tres funciones llamadas *funciones de colecta*, lo que da lugar a los modelos: *Exponencial, logarítmico* y *Clench.*

En este trabajo, proponemos un modelo que toma en cuenta la naturaleza discreta del esfuerzo, y refleja el hecho de que la capturabilidad de las especies puede depender de factores aleatorios. Conjeturamos que la forma de la función de colecta depende de la *probabilidad de captura* de las especies, la cual es función, entre otras cosas, de la abundancia y el método de muestreo. Supondremos además que las ditribución de las probabilidades de captura es $Beta(\alpha, \beta)$.

Para la estimación de los parámetros utilizamos el método de máxima verosimilitud, así como verosimilitud perfil y bootstrap para conseguir intervalos de confianza. Otro problema que consideramos en este trabajo es el de aquellas listas que no contienen información completa. Finalmente analizaremos algunas listas de datos reales.

Formulación de un modelo de acumulación

Denotemos por k el número total de especies diferentes presentes en la localidad, cuyo valor es fijo pero desconocido. Y, por el momento, adoptemos las siguientes suposiciones: (a) Los procesos de recolección son independientes en el tiempo; (b) En cada tiempo n, cada especie es observada independientemente de las demás, con las mismas probabilidades $p_1, p_2, \ldots, p_k$ y (c) $P_1, P_2, \ldots, P_k$ son independientes e idénticamente distribuidas.

Definamos:

$$X_n = \text{Número de especies nuevas recolectadas en el tiempo } n.$$

Con esta notación, podemos definir con precisión la función de acumulación de especies como (n, N_n), donde $N_0 = 0$ y $N_n = N_{n-1} + X_n$, $n = 1, 2, \ldots$.

El siguiente resultado es la base de la inferencia en este trabajo.

Teorema: Bajo las hipótesis del modelo, la distribución conjunta de $(X_1, X_2, \ldots, X_n)$ es multinomial con parámetros

$$k, E(P_1), E(P_1(1-P_1)), \ldots, E\left(P_1(1-P_1)^{n-1}\right).$$

Inferencia

Supongamos que los datos consisten en $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ donde $(X_1, X_2, \ldots, X_n)$ tiene la distribución

$$MN_n\left(k, \frac{b(\alpha+1,\beta)}{b(\alpha,\beta+n)}, \frac{b(\alpha+1,\beta+1)}{b(\alpha,\beta+n)}, \ldots, \frac{b(\alpha+1,\beta+n-1)}{b(\alpha,\beta+n)}\right). \quad (1)$$

Los parámetros α y β corresponden a la distribución $Beta(\alpha, \beta)$.

Máxima verosimilitud

En vista de que se cuenta directamente con una función de densidad para los datos, es natural estimar los parámetros por máxima verosimilitud. Sin embargo, los aspectos de precisión y construcción de intervalos de confianza para dichas estimaciones, no son tan fáciles de obtener. Un estudio de simulación reveló que en general, las distribuciones marginales de los estimadores de máxima verosimilitud —aún para tamaños de muestra grandes— no siguen una distribución normal. Esto indica que es necesario recurrir a procedimientos alternos para evaluar la precisión de $\left(\hat{k}, \hat{\alpha}, \hat{\beta}\right)$.

Verosimilitud perfil

Si se enfoca la labor de inferencia sobre el *parámetro de interés* k , los parámetros α y β se les denomina *parámetros de estorbo.*

Una de las herramientas que existen para lograr inferencia en estos casos es utilizando la llamada *función de verosimilitud perfil*, que aunque no es una verosimilitud, sus propiedades la hacen funcionar como tal para fines de estimación y construcción de intervalos de confianza. Para una introducción a la metodología basada en verosimilitudes perfil, puede consultarse Barndorff-Nielsen & Cox (1994) y Hinkley *et al.* (1991).

Intervalos de confianza para k, α, β. Bootstrap

Una técnica alternativa, para estimar la matriz de covarianza y otras medidas de precisión de los estimadores es utilizar el método de Bootstrap Efron, (1982). El bootstrap es un método basado en computación intensiva, el cual sustituye un considerable costo de cómputo en lugar de análisis teórico, y permite responder a preguntas que serían muy complicadas usando métodos analíticos.

Observaciones censuradas. Algoritmo *EM*

En esta sección abordamos el problema de las listas que contienen datos faltantes. Dado que no es posible escribir de manera sencilla la distribución para este caso, se recurre a algoritmo *EM* Dempser, Laird, & Rubin, (1977) para estimar los parámetros. Para la implementación de este método al presente trabajo ver Peraza (1995).

Ejemplos

Murciélagos de Chajul

Este ejemplo corresponde a la lista de especies de murciélagos reportadas por Medellín (1986, no publicada) de la Estación Biológica de Chajul en la selva lacandona de México.

Para este problema el resultado estimado del número total de especies es muy semejante al reportado por Soberón y Llorente. Sin embargo, claramente constituye una ventaja el que el modelo desarrollado en este trabajo no se requiere conocimiento a priori sobre las condiciones generales de la zona de trabajo que determinen cuál modelo seleccionar. Esto es importante, pues permite aplicaciones mas generales fuera del contexto de ecología.

Mariposas de Pakitza

Soberón *et al.* (1992) analizan los datos de una colección de 200 horas/hombre (durante septiembre de 1989) en la estación biológica Pakitza, en Perú. Soberón y Llorente comentan que el modelo logarítmico es el mas adecuado para extrapolar argumentando razones de tipo biológico.

Estos datos constituyen un ejemplo de una lista censurada . Los estimadores de los parámetros aplicando el algoritmo *EM* con el modelo desarrollado en el presente trabajo son $\hat{k} = 981$; $\hat{\alpha} = .677887$ y $\hat{\beta} = 74.9913$. Es decir, se estiman 981 especies contra los valores asintóticos de 890 y 611 del modelo de Clench y exponencial, respectivamente. Existe una observación muy interesante en que Robbins (en comunicación personal con Soberón) reporta en fecha muy posterior, la lista en 565 horas/hombre, acumulándose 979 especies. Este último valor es muy cercano al número total estimado por nuestro modelo y excede los estimados con el modelo de Clench y exponencial.

Conclusiones generales

El modelo presentado en este trabajo, constituye una nueva alternativa para resolver el problema de modelar la relación entre esfuerzo y número de especies acumuladas. El hecho de tomar en cuenta la naturaleza discreta de las unidades de esfuerzo y el considerar en la modelación que sólo tienen relevancia directa las probabilidades de captura, lo hacen mas realista y de más fácil interpretación.

Por otra parte, es de importancia el contar con intervalos de confianza para los parámetros.

El contar con una fórmula para el valor esperado condicional a lo observado constituye una ventaja clara sobre los otros métodos. Debe parecer obvio que para efectos de predicción o extrapolación a futuro esta expresión es la más adecuada, además de que se modela la relación entre esfuerzo y captura. Estas cantidades que usualmente no se consideran, son expresiones muy sencillas de calcular bajo el enfoque dado. El obtener además una solución para el caso de las observaciones censuradas fue muy importante, dado que esta situación se presenta de manera frecuente.

Sin embargo, el usuario debe tener conocimientos amplios en programación para implementar los algoritmos requeridos. Esta característica no se presenta en los modelos de acumulación de especies referidos por Soberón & Llorente (1992), en los cuales se ocupan técnicas implementadas en la mayoría de los paquetes estadísticos existentes.

Referencias

Peraza, F. J. (1995) *Un Modelo para la Acumulación de Especies,* Tesis de Maestría en Estadística, Universidad de Guanajuato, Facultad de Matemáticas.

Soberón, J. and Llorente, J. (1993) "The Use of Species Accumulation Functions for the Prediction of Species Richness", *Conservation Biology*, **7**, 3, 480–488.

Estimación de la media poblacional usando información auxiliar desfasada: una comparación de siete procedimientos de estimación

Blanca Rosa Pérez-Salvador
Universidad Autónoma Metropolitana, Unidad Iztapalapa
Ignacio Méndez Ramírez
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México

Introducción

La información auxiliar puede mejorar la precisión de una estimación, especialmente cuando está linealmente asociada con la variable de interés. Por ejemplo, el ingreso familiar promedio puede tener como variable auxiliar, al número de habitantes en la zona de estudio. O sea que si Y representa al ingreso y X al número de habitantes de una zona geográfica, entonces para un cierto rango de X, se espera que

$$Y = \alpha + \beta X + \varepsilon,$$

donde X y Y son contemporáneas; α y β son constantes y $\varepsilon \sim N(0, \sigma^2)$

Bajo este modelo, los valores de X pueden mejorar la precisión de las estimaciones de $\mu = \overline{Y}$. Pero, si los valores de X son de un censo levantado años atrás, entonces podría no ser así.

Para analizar los efectos del uso de información no contemporánea, se calcularon y compararon las precisiones y exactitudes de siete procedimientos de estimación.

En lo que sigue, las letras mayúsculas definen conceptos poblacionales, mientras que las letras minúsculas, conceptos muestrales. (Ejemplo: N representa el tamaño de la población mientras que n representa el tamaño de la muestra).

El modelo

De acuerdo al problema, existen tres variables:

Y, la variable de interés. (Desconocida)

X, la variable auxiliar, contemporánea a Y. (Desconocida)

X', la variable auxiliar, no contemporánea a Y. (Conocida)

tales que: $X = X'(A + Z)$ y $Y = \alpha + \beta X + \varepsilon$; con A, α y β constantes desconocidas, Z y ε v. a. tales que $E(X|X') = X'A$ y $E(Y|X') = \alpha + \beta X'A$.

TABLA I. Procedimientos de estimación

Forma del estimador	Nombre	Método de muestreo
$\hat{\mu}_1 = \overline{y}$	promedio simple	muestreo simple aleatorio, sin reemplazo (mai)
$\hat{\mu}_2 = \overline{X}\,\frac{\overline{y}}{\overline{x}}$	Estimador de razón	mai sin reemplazo
$\hat{\mu}_3 = \overline{y} - \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}\,(\overline{x}-\overline{X})$	Estimador de regresión	mai sin reemplazo
$\hat{\mu}_4 = \overline{y}\left(\frac{\overline{X}}{\overline{x}}\right)^{\delta};\ \delta = \frac{\overline{x}}{\overline{y}}\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$	Primer *estimador* de Swine	mai sin reemplazo
$\hat{\mu}_5 = \overline{y}\,\frac{\overline{X}}{\delta\overline{x}+\overline{X}(1-\delta)}$	Segundo *estimador* de Swine	mai sin reemplazo
$\hat{\mu}_6 = \frac{\overline{X}}{n}\sum\frac{y_i}{x_i}$	muestreo con probabilidad proporcional al tamaño, con reemplazo	
$\hat{\mu}_7 = \overline{X}\frac{\overline{y}}{\overline{x}}$	Estimador de razón insesgado	muestreo con probabilidad proporcional al tamaño agregado

Procedimientos de estimación

- Cada estimador es una función de las variables contemporaneas X y Y, excepto $\hat{\mu}_1$; así, $\hat{\mu}_1$ sirve como un control, dado que no usa la información auxiliar.
- $\hat{\mu}_2$ y $\hat{\mu}_7$ tienen la misma forma pero diferente método de muestreo.
- Sólo $\hat{\mu}_1$, $\hat{\mu}_6$ y $\hat{\mu}_7$ son estimadores insesgados.
- Cuando $|\hat{\beta}_0| \ll |\overline{y}|$, $\delta \approx 1$ y, $\hat{\mu}_4$ y $\hat{\mu}_5 \approx \hat{\mu}_2$, el estimador de razón; y cuando $|\hat{\beta}_1\overline{x}| \ll |\overline{y}|$, entonces $\delta \approx 0$ y, $\hat{\mu}_4$ y $\hat{\mu}_5 \approx \hat{\mu}_1$, el promedio simple. ($\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores de mínimos cuadrados del modelo lineal, $Y = \beta_0 + \beta_1 X + \varepsilon$).

Para adaptar los procedimientos de estimación al problema, se consideró que:

1. En las expresiones de $\hat{\mu}_i$, se remplazan los valores de $\overline{X}$, $\overline{x}$ y x_i por los valores de $\overline{X}'$, $\overline{x}'$ y x_i', respectivamente. Esto es, *"X no se actualiza."*
2. En las expresiones de $\hat{\mu}_i$, se reemplazan los valores de $\overline{X}$ por los valores de $\overline{X}'$, y se miden los valores de x_i. Esto es *" X se actualiza"*. El método resulta más caro, porque se muestrean dos valores, x_i y y_i; además, $\hat{\mu}_6$ y $\hat{\mu}_7$ ya no son insesgados.

La simulación

Para la simulación se ejecutó un programa que:

- genera 200 valores para X y Y con base en 200 valores dados para X'. (En particular los datos de entrada cumplen que $\overline{X}' = 28\,268.56$ y $\sigma_{X'} = 41\,056.95$.)
- selecciona 20 unidades muestrales considerando los diferentes esquemas de muestreo y se estima $\hat{\mu}_i$. Este procedimiento se realiza 100 veces.
- Con los 100 valores del proceso anterior se estima la precisión y exactitud de μ_i.

Para generar X, se consideró:

- Dos valores para A
  1. $A = 1$. Lo que significa que $E(X) = X'$.
  2. $A = e^{0.14}$. El crecimiento poblacional al 2% durante 7 años. $E(X) = X'e^{0.14}$.
- Dos distribuciones para Z:
  1. $Z \sim U(-a, a)$ con $a = 0.05, 0.1, 0.3$ y 0.4.
  2. y $Z \sim N(0, a^2)$ con $a = 0.05, 0.1, 0.3$ y 0.4.

Para generar a Y, se consideró:

- Tres valores para α: 1) $\alpha = 0$. 2) $\alpha = 100$. y 3) $\alpha = 10\,000$.
- Cuatro valores para σ: 10, 100, 1 000 y 10 000.
- Un valor para β igual a 2.

También se consideró actualizar y no actualizar a X.
Los criterios de comparación son:

- Sesgo empírico ..... $\boxed{Sesgo_{\hat{\mu}_i} = \overline{\hat{\mu}}_i - \mu}$
- Desviación estándar empírica ..... $\boxed{s_{\hat{\mu}_i} = \sqrt{\frac{\sum(\hat{\mu}_i - \overline{\hat{\mu}}_i)}{99}}}$
- El error cuadrático medio empírico (ECM) ..... $\boxed{ECM_{\hat{\mu}_i} = \sqrt{\frac{\sum(\hat{\mu}_i - \mu)^2}{100}}}$
- Y el coeficiente de variación empírica (CV) ..... $\boxed{CV = \frac{RMSE_{\hat{\mu}_i}}{\mu}}$

Fig. 1

$X = X'(1 + Z)$

X no se actualizó

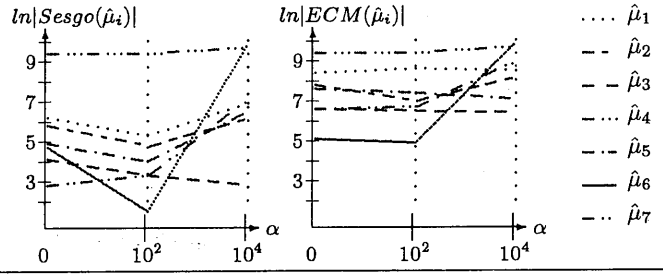


Fig. 2

$X = X'(e^{0.14} + Z)$

X no se actualizó

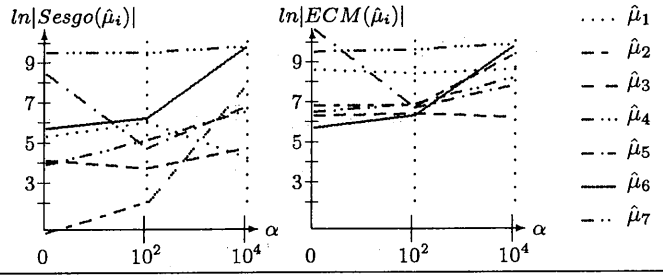


Fig. 3

$X = X'(1 + Z)$

X se actualizó

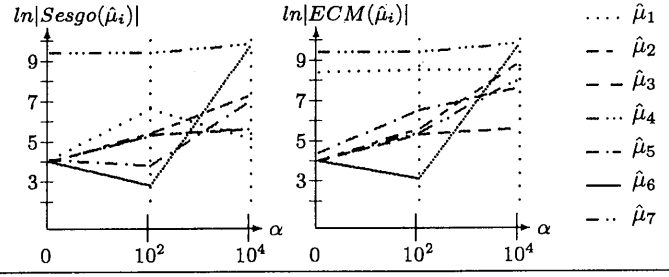


Fig. 4

$X = X'(e^{0.14} + Z)$

X se actualizó

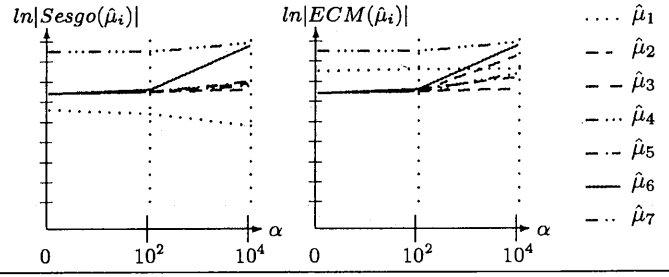


TABLA II

$X = X'(1+Z)$	
X **no se actualiza**	X **se actualiza**
El sesgo, la desviación estandar (DE) y el error cuadrático medio de los estimadores es "grande".	Cuando $\alpha = 0$, $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_5$, $\hat{\mu}_6$ $\hat{\mu}_7$ son "buenos" estimadores; cuando $\alpha = 10\,000$, $\hat{\mu}_3$ es el mejor estimador.
Conclusión: conviene actualizar a X. $\hat{\mu}_6$ es mejor cuando $\alpha \cong 0$ y $\hat{\mu}_3$ es mejor cuando $\alpha \not\cong 0$.	

Resultados

Para cubrir todas las combinaciones, se efectuaron 384 corridas del proceso de simulación previamente ilustrado. Debido a la falta de espacio sólo se muestra la gráfica de unas cuantas corridas. Las funciones de distribución de Z y ε son $Z \sim N(0, 0.1)$ y $\varepsilon \sim N(0, 100)$ para todas las gráficas. Además se utilizó una escala logarítmica.

Los resultados más importantes se resumen en las siguientes tablas.

TABLA III

$X = X'(e^{0.14} + Z)$	
X **no se actualiza**	X **se actualiza**
la desviación estandar (DE) el sesgo y el ECM son "pequeños".	La DE es "pequeña", pero el sesgo y el ECM es "grande".
Conclusión: no conviene actualizar a X, a menos que el sesgo se corrija.	

En esta última tabla, se observa que cuando en promedio X cambia "mucho" respecto a X', es mejor no actualizarla. En este caso, el mejor estimador parece depender de α y de σ_Y, como se muestra en la tabla IV.

TABLA IV El mejor proceso de estimación, de acuerdo a los valores de α y σ_Y en el caso que X se actualice y $X = X'(e^{0.14} + Z)$.

	$\alpha = 0$ y 100 ($\leq 0.1\%$ of σ_Y)	$\alpha = 10\,000$ ($\cong 10\%$ of σ_Y)
$\sigma = 10$, 100 1 000 ($\leq 1\%$ of σ_Y)	El mejor procedimiento es 6	El mejor procedimiento es 3
$\sigma = 10\,000$ ($\cong 1\%$ of σ_Y)	Los mejores procedimientos son 3, 6 y 7.	El mejor procedimiento es 3

Finalmente, la siguiente tabla orienta al investigador cómo elegir el procedimiento de estimación con base en los datos de una muestra piloto.

TABLA V Elección de acuerdo a los valores de A y wα

		$H_0: A = 1$ *vs* $H_a: A \neq 1$	
		si no es significativa, actualice a X y	si es significativa, no actualice a X y
$H_0: \alpha = 0$ *vs* $H_a: \alpha \neq 0$	si no es significativa	elija $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_6$, y $\hat{\mu}_7$	elija $\hat{\mu}_6$
	si es significativa	elija $\hat{\mu}_3$	elija $\hat{\mu}_3$

Calculo de la tasa de riesgo

FRANCISCO PABLO RAMÍREZ GARCÍA & MARISA MIRANDA TIRADO
Gerencia de Ciencias Básicas e Investigación Prospectiva, SGIDTTI, Instituto Mexicano del Petróleo

Introducción

En todo proyecto de inversión se debe contemplar un riesgo. Es por ello que los inversionistas consideran invertir en varios proyectos a la vez, de tal forma que se minimice el riesgo de la pérdida total del capital de financiamiento, es decir, dividir los riesgos a que el capital está sujeto. Esto es equivalente al dicho "los huevos no se deben poner en una sola canasta".

Los procedimientos usados para evaluar financieramente los riesgo inherentes al capital por lo general no contemplan indicadores matemáticos que involucren a todo los proyectos en el portafolio de inversiones. Sin embargo, en el análisis financiero se examinan diversas metodologías como el Valor Presente Neto (VPN), la Tasa Interna de Retorno (TIR), la Relación Costo-Beneficio, el Modelo de Asesoramiento del Precio de Capital, Modelo de Evaluación de Dividendos y el Modelo de Gordon (Powell, 1986), las cuales se aplican para responder si un proyecto que será apoyado financieramente dará un beneficio positivo.

Para analizar carteras de inversión, ha sido necesario generar diversos estudios que conformen escenarios más amplios en los que la inversión de capital global dé un beneficio positivo; sin importar que algunos de los proyectos en la cartera se cancelen, no reporten dividendos o no sea factible su implementación a nivel industrial, a lo que se denominará proyectos nulos o rojos. A continuación se ofrece una metodología que permite dar solución al problema, para ello se utiliza el método de evaluación del VPN, debido principalmente a las ventajas que tiene sobre otros métodos, tales como sencillez, unicidad, aditividad, su amplio uso para evaluar proyectos de inversión y por contemplar la recuperación de capital independiente del concepto de depreciación de bienes de capital.

El VPN permite determinar el beneficio en capital, a valor presente, sobre la inversión en base al flujo de efectivo futuro que éste generará y compararlo con los desembolsos requeridos (Coss, 1993), matemáticamente se representa como:

$$VPN_{i,n} = \sum_{j=1}^{n} \frac{FE_j - I_j}{(1+i)^j} \tag{1}$$

donde:

- I_j Inversión requerida en el año j,
- i Tasa de interés,
- n Duración de la cartera de proyectos,
- FE_j Flujo de Efectivo del año j.

Cuando el VPN es menor o igual a cero se rechaza el proyecto; de lo contrario, la inversión proporcionará una ganancia.

Planteamiento

Dado un portafolio con m proyectos, de los cuales sólo s llegan a su fin, es decir, los números "negros" en el portafolio, ¿Cuál debe ser la tasa de recuperación en los proyectos negros (s), de tal forma que se recupere la inversión en los proyectos rojos ($m-s$) y dé un beneficio en la inversión total del portafolio? La recuperación de capital se debe contemplar independiente del concepto de depreciación de bienes de capital.

En base al VPN del portafolio de m proyectos se desea encontrar la tasa de inversión que contempla que los proyectos terminales recuperen la inversión global, a la cual se denominará una tasa de riesgo (r), que es mayor a i la Tasa Mínima Atractiva de Rendimiento:

$$VPN_{p=i,n}^{\kappa=1} = VPN_{p=r,n}^{\kappa=k}, \tag{2}$$

donde: $\kappa = \frac{s}{m}$ proporción de proyectos que llegan a su fin $(0 \leq k \leq 1)$;

$$VPN_{p,n}^{\kappa} = \sum_{l=1}^{m}\sum_{j=0}^{n} \frac{\kappa FE_{lj} - I_{lj}}{(1+\rho)^j}.$$

En el caso de hacer la inversión solo una vez, al inicio de los proyectos, entonces $I_0 = \sum I_{ol}$ y

$$VPN_{i,n}^{\kappa} = \kappa \sum_{l=1}^{m}\sum_{j=0}^{n} \frac{FE_{lj}}{(1+r)^j} - I_0.$$

Empleando 2 se tiene

$$VPN_{r,n}^{\kappa=k} = k\left[VPN_{i,n}^{\kappa=1} + I_o\right] - I_o. \tag{3}$$

Ejemplos

Para mostrar el procedimiento se ilustra a continuación con dos ejemplos el cálculo de la tasa de riesgo, cada uno con un paquete de 10 proyectos:

Cuando los flujos de efectivo sean iguales para todos los años, es decir, $FEj = FE_l$ para toda j (Ejemplo 1) o cuando los flujos de efectivo sean iguales para todos los proyectos, es decir, $FE_{lj} = FE_j$ para toda l (Ejemplo 2) y $FE = \sum\sum FE_{lj}$. En ambos casos el VPN se reduce a:

$$VPN_{i,n} = FE\frac{1-(1+i)^{-n}}{i} - I_0\ ; \tag{4}$$

$$VPN_{i,n}^{k} = kFE\frac{1-(1+i)^{-n}}{i} I_0. \tag{5}$$

Utilizando la ecuación (3), la tasa de riesgo (r) es tal que:

$$\frac{1-(1+r)^{-n}}{r} - k\frac{1-(1+i)^{-n}}{i} = 0 \tag{6}$$

cuya solución se encuentra por métodos numéricos.

De estos ejemplos se puede observar que:

Ejemplo 1

	Inv. Inicial					
AÑO	1994	1995	1996	1997	1998	1999
Flujo Efec.	($1,000)	$325	$325	$325	$325	$325
i=	15.00%	TIR=	18.72%			
VPN=	$77.78					
Si k=	0.7	r=	31.97%			
Si k=	0.9	r=	19.60%			

Ejemplo 2

	Inv. Inicial					
AÑO	1994	1995	1996	1997	1998	1999
Flujo Efec.	($1,000)	$300	$375	$450	$525	$600
i=	15.00%	TIR=	30.13%			
VPN=	$381.55					
Si k=	0.7	r=	31.97%			
Si k=	0.9	r=	19.60%			

- A mayor proporción de proyectos que llegan a su fin, menor tasa de riesgo.
- Si $k = 0$ entonces r tiende a infinito.
- Si $k = 1$ entonces $r = i$.
- A mayor VPN, mayor TIR (En este caso es posible calcular el TIR, dado que sólo se realiza una inversión al inicio de los proyectos).

Modelo de actualización bayesiana

Dentro de una cartera de inversión, la realización de uno de los proyectos no depende de lo que suceda con los demás, es decir, se considera que los proyectos son estadísticamente independientes. Es posible agrupar los proyectos en carteras de inversión, de tal forma que cada paquete contenga exclusivamente proyectos con la misma viabilidad, es decir, que cumpla con el supuesto de estacionalidad. De esta forma, el éxito o fracaso de una cartera de inversión se puede contemplar como un proceso Bernoulli.

Para obtener la distribución de la fracción de proyectos que llegan a su fin (K), de una cartera dada, se empleó un procedimiento Bayesiano. La distribución apriori de K se supuso como uniforme y como normal con parámetros (μ, σ) y se aplica al ejemplo 1 desarrollado en el inciso anterior, suponiendo que solo 7 proyectos llegan a su fin, o son negros.

Distribución uniforme

Donde X es $[0, 1]$ y la muestra obtenida es de tamaño m, en la cual resultaron s éxitos $(X = s)$; entonces, haciendo uso de los métodos Bayesianos, la distribución aposteriori de K es una distribución $\beta(s + 1, m - s + 1)$:

$$P(K \setminus X = s) = \frac{(m+1)! =!}{s!(m-s)!} k^s (1-k)^{m-s} \tag{7}$$

con

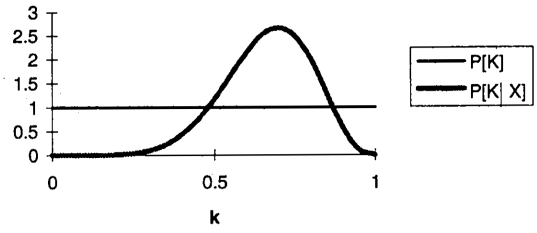
$$\overline{k} = \frac{(s+1)}{(m+2)} \qquad \text{y} \qquad \sigma_k^2 = \frac{(m-s+1)(s+1)}{(m+3)(m+2)^2}$$

Al aplicar estos resultados al ejemplo 1, con $k = 0.7$ ($m = 10$ y $s = 7$), se obtienen las distribuciones a prior (uniforme) y aposteriori $\left(\beta\,(8,4)\,,\ \text{con } \overline{k} = 0.67 \text{ y } \sigma^2 = 0.36\right)$ de la figura 1.

Distribución de la tasa de riesgo

Para obtener la distribución de la tasa de riesgo ($\mathbf{R}$) basta con aplicar la transformación dada por la ecuación (6) y utilizar métodos numéricos para obtener la distribución aposteriori, considerando dos aspectos:

Figura 1: Distribuciones de k
7 éxitos de 10



- La función de probabilidad de una transformación monótona continua deja a la función de densidad inalterada salvo una constante de normalización.
- El cambio inherente a la transformación monótona, conlleva un cambio de escala, el que en el análisis matemático se representa por el Jacobiano de la transformación.

Al realizar la transformación correspondiente a la fracción de proyectos que llegan a su fin respecto a la tasa de riesgo, la función de distribución aposteriori (7) se modifica como se muestra en la figura 2. La gráfica sólo corresponde al intervalo $[0, 1]$, dado que la mayor parte de la información se encuentra en este intervalo. No obstante que la variable aleatoria $\mathbf{R}$ tiene dominio $[0.15, \infty]$, el que corresponde al dominio de $\mathbf{K}[1, 0]$.

Conclusiones

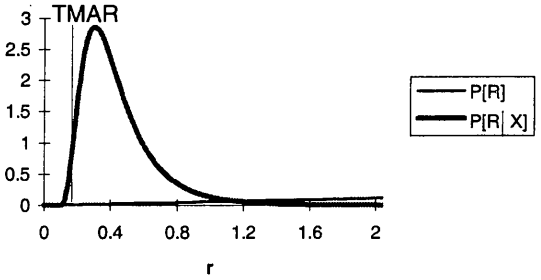
Es posible cuantificar el riesgo de un paquete de inversión en dos puntos:

- La información personal acerca del proyecto de inversión (grados de creencia; Lindley, 1969).
- Experiencia en proyectos de inversión con riesgos similares.

Los métodos Bayesianos permiten agregar la información obtenida del segundo punto a la información personal (primer punto) que se expresa como la distribución apriori. Además es posible conocer la probabilidad de que el VPN sea positivo, utilizando métodos de sensibilidad (Montecarlo).

Una ventaja del método Bayesiano es que permite generar criterios en base a probabilidades de éxito en la cartera seleccionada; estos criterios pueden ser:

Figura 2: Distribuciones de la tasa r
7 éxitos de 10



- Si $P[R \geq TMAR] \geq 0.90$ entonces se acepta la cartera, al implicar que se corre un riesgo pequeño de no alcanzar el rendimiento interno de retorno del paquete de proyectos.
- Si $P[R \leq TMAR] \geq 0.20$ entonces existe un gran riesgo de perder la inversión inicial en el paquete, por lo que se debe rechazar.

De la figura 2 se obtiene que $P[R \geq TMAR] = 0.9887$, indicando que una tasa mayor a la $TMAR$ en el 98.87% de los casos

Agradecimientos

Se reconoce a la Srita. Marisa Miranda Tirado el realizar muchos de los cálculos necesarios para este trabajo y el presentarlo en el Foro Nacional de Estadística, realizado en Saltillo, Coah., 1994.

Referencias

Powell, T. E. (1986) *A review of recent developments* in project evaluation, Chemical Engineering.

Coss Bu, R. (1993) *Analisis y Evaluación de proyectos de inversión*, Limusa.

Lindley,D. V. (1969) *Introduction to Probability and Statistics from a Bayesian viewpoint*, Cambridge University Press.

Estimación no paramétrica de la función de supervivencia bivariada

María del Refugio Rivera Rendón
UACPyP-Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México y A.C. Nielsen

Introducción

[1]En el análisis de supervivencia multivariado se miden k diferentes tiempos de falla en el mismo individuo o un tiempo de falla en k individuos diferentes pero que estén asociados, formando para cada observación un vector de k entradas, y cada una de las coordenadas está sujeta a censura con la misma estructura del caso univariado.

El desarrollo del análisis de supervivencia multivariado es aún incipiente. La parte más compleja es la construcción de estimadores en el caso bivariado ya que la generalización a más dimensiones es inmediata.

En el presente trabajo se esbozarán dos de las propuestas de estimación no paramétrica de la función de supervivencia bivariada para datos censurados, basadas en el estimador Kaplan Meier de la función de supervivencia univariada.

Campbell (1981)

Supone censura aleatoria para las componentes (T_1, T_2) y que las censuras ocurren independientemente de los tiempos de falla. Propone dos estimadores , el más interesante está basado en los desarrollos de Efron (1967) acerca de la función de autoconsistencia.

En el caso bivariado tenemos 4 posibles resultados para cada pareja: que los 2 sujetos fallen, que el primero falle y el segundo sea censurado, que el primero presente censura y el segundo falle, y por último, que los 2 individuos presenten censura.

Para construir la función de verosimilitud, se definen:

δ_{ij} = Número de parejas en que los 2 sujetos mueren, el primero t_{1i} al tiempo y el segundo al tiempo t_{2j}

α_{ij} = Número de parejas en las que el primer sujeto es censurado al tiempo t_{1i} y el segundo presenta falla al tiempo t_{2j}

β_{ij} = Número de parejas en las que el primer sujeto presenta falla al tiempo t_{1i} y el segundo es censurado al tiempo t_{2j}.

λ_{ij} = Número de parejas en las que los 2 sujetos presentan censura, el primero al tiempo t_{1i} y el segundo al tiempo t_{2j}.

[1]Parte de la tesis de Maestría en Estadística e Investigación de Operaciones de oa UACPyP del CCH e IIMAS-UNAM.

La función de verosimilitud propuesta es:

$$L = \prod_{i=1}^{I} \prod_{j=1}^{J} \Delta_{ij}^{\delta_{ij}} S_{ij}^{\lambda_{ij}} Q_{ij}^{\alpha_{ij}} R_{ij}^{\beta_{ij}}$$

donde:

S_{ij} =Es la probabilidad de que el primer individuo sobreviva más allá del tiempo T_{1i} y el segundo individuo sobreviva más allá del tiempo t_{2j}.

$\Delta_{ij} = S_{ij} + S_{i-1,j-1} - S_{i,j-1} - S_{i-1,j}$ es la probabilidad de muerte en el rectángulo $(t_{1_{i-1}}, t_{1i}] \times (t_{2_{j-1}}, t_{2_j}]$

$\hat{Q}_{ij} = \hat{S}_{i,j-1} - \hat{S}_{ij}$ es la probabilidad de muerte en el rectángulo $(t_{1_i}, \infty] \times (t_{2_{j-1}}, t_{2_j}]$

$R_{ij} = S_{i-1,j} - S_{ij}$ es la probabilidad de muerte en el rectángulo $(t_{1_{i-1}}, t_{1_i}] \times (t_{2_j}, \infty]$

Para maximizar L y estimar S_{ij} por máxima verosimilitud, para valores fijos de $\delta_{ij}, \lambda_{ij}$, α_{ij} y β_{ij} se deriva el $\log(L)$ con respecto a S_{ij} y se iguala a cero, obteniéndose una ecuación que al resolverse no se tiene una expresión cerrada para $\hat{S}_{ij}$, se propone resolverlo por un método iterativo, por ejemplo el algoritmo EM.

La idea de autoconsistencia de Efron en este caso se refiere a que el estimador de S_{ij} cumpla con la siguiente ecuación:

$$n\hat{S}_{ij} = N_{ij} + \sum_{\substack{l>j \\ k<i}} \alpha_{kl} \frac{\hat{Q}_{il}}{\hat{Q}_{kl}} + \sum_{\substack{k>i \\ l<j}} \beta_{kl} \frac{\hat{R}_{kj}}{\hat{R}_{kl}} + \sum_{\substack{k<i \\ l<j}} \lambda_{kl} \frac{S \max(i,k), \max(j.l)}{S_{kl}} \qquad (1)$$

donde:

$$\hat{Q}_{il} = \hat{S}_{i,j-1} - \hat{S}_{ij}, \quad \hat{R}_{ij} = \hat{S}_{i-1,j} - \hat{S}_{ij} \text{ y } N_{ij} = \sum_{\substack{k>i \\ l<j}} \delta_{kl} + \sum_{\substack{k\geq i \\ l>j}} \alpha_{kl} + \sum_{\substack{k>i \\ l\geq j}} \beta_{kl} + \sum_{\substack{k\geq i \\ l\geq j}} \lambda_{kl}.$$

Este estimador tiene la propiedad que en el caso de no censura se reduce a la función de supervivencia empírica. Tiene saltos únicamente en los puntos de doble falla o en el último valor censurado en cualquier dimensión. Es generalizable a más dimensiones. Una desventaja que se le atribuía era la complejidad del cálculo por no tener una expresión cerrada y resolverse por medios iterativos, en la actualidad ya no se considera un problema.

Dabrowska (1988)

Propone un estimador para la función de supervivencia bivariada basado en las ideas de Aalen y Johasen (1978) y de Gill y Johansen (1987) quienes sugieren expresar a la función de supervivencia univariada como producto integral de la función acumulada de riesgo. Se busca una representación similar para el caso bivariado construyendo un vector con entradas que representan la función de riesgo acumulado correspondiente a falla simple y doble.

Se expresa la función de supervivencia bivariada en términos de la función acumulada de riesgo bivariada y se usa el principio de substitución, es decir, se reemplaza el estimador Kaplan-Meier en su forma de producto integral.

Sean $T = (T_1, T_2)$ un par de variables aleatorias no negativas y continuas definidas en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ y sea $S(t_1, t_2) = P[T_1 \geq t_1, T_2 \geq t_2]$ la correspondiente función de supervivencia conjunta.

Por una función de riesgo bivariada entenderemos un vector de funciones dado por $\lambda(t_1, t_2) = \lambda_{10}(t_1, t_2)$, $\lambda_{01}(t_1, t_2)$, $\lambda_{11}(t_1 t_2)$ donde λ_{10} representa el riesgo de que la primera componente falle en t_1 y la segunda sobreviva a t_2, λ_{01} de que la primera sobreviva a t_1 y la segunda componente falle en t_2 y λ_{11} de que se presente falla simultánea al tiempo t_1 en la primera componente y al tiempo t_2 en la segunda.

Las correspondientes funciones de riesgo acumulados se definen a partir de las anteriores. La función de supervivencia bivariada $S(t_1, t_2)$ se expresa en términos del vector de riesgo acumulado

$$\Lambda(t_1, t_2) = (\Lambda_{10}(t_1, t_2), \Lambda_{01}(t_1, t_2), \Lambda_{11}(t_1, t_2)),$$

como sigue:

$$S(t_1, t_2) = S(t_1, 0) S(0, t_2) \exp\left\{ \int_0^{t_1} \int_0^{t_2} \lambda_{11}(u, v) - \lambda_{01}(u, v) \lambda_{10}(u, v) \, du dv \right\}, \qquad (2)$$

En donde $S(t_1, 0)$ representa la función marginal de supervivencia correspondiente a la primera componente y $S(0, t_2)$ a la segunda componente.

Bajo un esquema de censura aleatoria, un candidato natural para un estimador de la función de supervivencia bivariada es:

$$\hat{S}(t_1, t_2) = \hat{S}(t_1, 0) \hat{S}(0, t_2) \prod_{\substack{0 < u \leq t_1 \\ 0 < v \leq t_2}} \left(1 - \hat{L}((u_{i-1}, u_i] \times (v_{j-1}, v_j])\right)$$

donde $\hat{S}(t_1, 0)$ y $\hat{S}(0, t_2)$ son los estimadores $K - M$ usuales:

$$\hat{S}(t_1, 0) = \prod_{u \leq t_1} \left(1 - \hat{\Lambda}_{10}((u_{i-1}, u_i], 0)\right),$$

$$\hat{S}(0, t_2) = \prod_{v \leq t_2} \left(1 - \hat{\Lambda}_{01}(0, (v_{j-1}, v_j])\right)$$

y

$$\begin{aligned} & \hat{L}((u_{i-1}, u_i] * (v_{j-1}, v_j]) \\ = \; & \hat{\Lambda}_{10}((u_{i-1}, u_i], v_j) \hat{\Lambda}_{01}(u_i, (v_{j-1}, v_j]) - \hat{\Lambda}_{11}((u_{i-1}, u_i], (v_{j-1}, v_j]). \end{aligned}$$

Notar que las marginales de $\hat{S}(t_1, t_2)$son los estimadores $K - M$ univariados. El estimador $\hat{S}(t_1, t_2)$ no depende del orden que se la haya asignado a las componentes. En ausencia de censura se reduce a la función de supervivencia empírica. Es un estimador consistente. La extensión al caso general con k componentes, se hace de forma inductiva.

Este es uno de los mejores estimadores que se tienen hasta ahora pero Pruitt (1988) analiza los casos en los que se asigna probabilidades negativas.

Referencias

Campbell, G. (1982) "Nonparametric bivariate estimation with randomly censored data", *Biometrika*, **68,** 417–422.

Dabrowska, D.M. (1988) "Kaplan-Meier estimate on the plane", *The Annals of Statistics*, **16,** 1475–1489.

Peterson, A.V. (1977) "Expresing the Kaplan-Meier estimator as a function of empirical subsurvival functions", *J. Amer. Statist. Assoc.*, **72,** 854–858.

Pruitt,R. (1988) *The distribution of the Bivariate Kaplan-Meier Estimate. Technical*, Report No. 517, School of Statistics University of Minnesota.

Detección de observaciones discrepantes en observaciones Poisson

SILVIA RUIZ VELASCO ACOSTA
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México

Introducción

En estudios epidemiológicos es común buscar observaciones discrepantes grandes, es decir en un solo lado, por ejemplo: tenemos datos de cáncer en diferentes sitios y sólo nos interesan los sitios en que existen más casos que los esperados, no los sitios en que existen menos casos que los esperados. En particular si tenemos una serie de observaciones Poisson con denominadores conocidos, es decir tenemos $Y_1, \ldots, Y_n$ variables independientes Poisson cuya media es $\lambda_{d_1}, \ldots, \lambda_{d_n}$, con λ desconocido y d_i conocidas.

Si denotamos como S la variable aleatoria que representa a la suma de las observaciones, entonces condicionalmente en $S = \sum Y_j$ los conteos $Y_1, \ldots, Y_n$ tienen una distribución multinomial con índice $s = \sum y_j$, y con probabilidades $(d_1/\sum d_j, \ldots, d_n/\sum d_j)$.

Escribiendo $a_i = d_i/\sum d_j$, entonces marginal en las otras celdas, Y_i tiene una distribución binomial con media sa_i y varianza $sa_i(1-a_i)$. Esto sugiere definir un residual estandarizado como

$$R_i = \frac{Y_i - sa_i}{sa_i(1-a_i)},$$

que corresponde a un residual de Pearson, o en su caso como

$$R_i' = 2\left(Y_i \log\left(\frac{Y_i}{sa_i}\right) + (s - Y_i)\log\left(\frac{s - Y_i}{s(1-a_i)}\right)\right),$$

que corresponde al residual de la devianza.

Podríamos entonces definir como estadística de prueba $T = \max R_i$ o $V = \max R_i'$. El nivel de significancia exacto puede encontrarse usando la distribución multinomial, o bien, un nivel de significancia aproximado para T (o V) puede encontrarse como

$$\mathrm{P}(T > t) = (\mathrm{F}(-t))^n.$$

Alternativamente siguiendo el trabajo de David y Borton (1962) podríamos encontrar una aproximación asintótica mejorada en el caso en que las a_i sean iguales. Si definimos

$$I_i = \begin{cases} 1 \text{ si } Y_i > m \\ 0 \text{ en otro caso} \end{cases},$$

entonces $M = \sum I_i$ representa el número de celdas que contienen al menos $m+1$ objetos.

Para algún λ fijo es posible probar que cuando m y n tienden a infinito para cualquier t, $\mathrm{E}(M^{(t)}) = \lambda^t(1 + O(1/m))$ por lo tanto la función de distribución de m puede aproximarse como

$$P(m) = P(M = 0) = e^{-\lambda} = e^{-\mu},$$

donde $\mu = \mathrm{E}(M)$.

Es posible generalizar este argumento al caso en que las probabilidades son diferentes. Definiendo

$$I_i = \begin{cases} 1 \text{ si } R_i > t \\ 0 \text{ en otro caso} \end{cases}$$

y

$$M(t) = \sum I_i(t),$$

entonces

$$\mathrm{E}(M(t)) = \sum \mathrm{E}(I_i(t)),$$

donde

$$\mathrm{E}(I_i(t)) = \sum_u sC_u(a_i)^u(1-a_i)^{s-u};\ u = [t], .., s.$$

Si t es grande la distribución de $M(t)$ esta cerca de la distribución Poisson, es decir, definiendo $\mathrm{E}(M(t)) = \mu(a,t)$, tenemos

$$P(M(t) = 0) = P(T < t) = e^{-\mu(a,t)},$$

por lo tanto el nivel de significancia es aproximadamente

$$1 - e^{-\mu(a,t)}.$$

La distribución de $M(t)$ puede mostrar subdispersión relativa a la distribución Poisson. Esto implica que la aproximación de $P(M(T) = 0)$ puede ser grande y por lo tanto el nivel de significancia se exageraría.

Por tal razón se utilizó una expansión de Charlier (tipo B) para ajustarlo. En este caso la distribución que aproxima a la Poisson está dada por

$$f(x) = \sum \frac{c_j V^j p\,(x,\lambda)\,(-1)^j}{j!},$$

donde V^j es el operador diferencia aplicado a la distribución Poisson, los coeficientes c_j se obtiene igualando los momentos de la variable aleatoria Poisson con los momentos de $f(x)$. En este caso sólo se corrigió utilizando dos términos.

Para evaluar qué tan bien funcionan las aproximaciones, en el caso de igualdad de los d_i se evaluaron las tres aproximaciones con el valor real obtenido a través de la función multinomial y en el caso de diferentes d_i, con los resultados de 5000 simulaciones. Las simulaciones se realizaron para diferentes números de celdas $(5, 20)$, para diferentes índices de la distribución multinomial $(4, 6, \ldots, 20)$, para diferentes valores de $t(1, ..., 4)$ y para diferentes distribuciones de las d_i. En términos generales la aproximación Charlier siempre es la mejor, aunque cuando t es igual a cuatro, la aproximación Poisson coincide con la Charlier. Por otra parte la aproximación normal funciona adecuadamente cuando el valor de t es igual a uno y a medida que el índice de la distribución multinomial se incrementa.

Posibles generalidades

Caso de dos o mas factores

En este caso suponemos que Y_{ij} son variables independientes Poisson con medias $d_{ij}\lambda_i\mu_j$, por ejemplo datos de grupo de ocupación y tipos de cáncer y lo que nos interesa es también residuales grandes. Una manera de atacar el problema sería considerar las poblaciones por separado. Otra manera sería considerar la distribución de las celdas dados los marginales, lo cual es bastante más complicado que en el caso de un solo factor.

Sobredispersión en los datos

Es común que los datos presenten sobredispersión, es decir una varianza mayor a la esperada bajo la suposición de distribución Poisson. Una forma en que se puede atacar este problema es suponiendo un componente aleatorio, η_i, tal que

$$\begin{aligned} \mathrm{E}(Y_i|\eta_i) &= \lambda d_i \eta_i, \\ \mathrm{V}(Y_i|\eta_i) &= \lambda d_i \eta_i, \\ \mathrm{E}(Y_i) &= \lambda d_i, \\ \mathrm{V}(Y_i) &= \lambda d_i + (\lambda d_i)^2 V(\eta_i). \end{aligned}$$

Si el objetivo es estimar el parámetro de sobredispersión, existen diversas formas de hacerlo, las mas comunes son por medio de componentes de varianza, o bien componentes de varianza en la transformación $\log(Y_i+1/2)$. Un enfoque paramétrico a este problema, es suponer una distribución para el componente aleatorio, η_i, con media uno. En particular si suponemos una gamma con media uno la distribución de Y_i es una binomial negativa.

Cuando el interés es encontrar observaciones discrepantes de este tipo de distribuciones, el enfoque paramétrico permite repetir el análisis anterior definiendo

$$R_i = \left(\frac{Y_i - \mathrm{E}(Y_i|n, \sum Y_i = s)}{\mathrm{V}(Y_i|n, \sum Y_i = s)} \right)^{1/2}$$

Referencias

David F. N. and Barton, D. E. (1962) *Combinatorial Chance*, Griffin.

Análisis y diagnóstico bayesiano de regresión
ANA TURRENT Y MANUEL MENDOZA
Departamento de Estadística y Actuaría, Instituto Tecnológico Autónomo de México

El modelo de regresión lineal

La popularidad del modelo de regresión lineal es atribuible a que es un modelo sencillo que intuitivamente parece adecuado en muchas situaciones, lo mismo que a la facilidad de su análisis sin importar el enfoque, frecuentista o bayesiano, que se adopte.

Como es bien conocido, a grandes rasgos, un modelo de regresión lineal múltiple consiste en la siguiente estructura. Sea $X_1, X_2, .., X_{p-1}$ una colección de variables cuyo valor se cree tiene influencia sobre el valor que toma otra variable Y. Una vez fijos los valores de $X_1, X_2, .., X_{p-1}$, la relación entre estas variables (explicativas) y la variable Y, (de respuesta) se describe, al menos en forma aproximada, con la expresión

$$Y = \theta_1 + \theta_2 X_1 + \theta_3 X_2 + \cdots + \theta_p X_{p-1} + \varepsilon$$

donde,

i) $\theta_1, \theta_2, \ldots, \theta_p$ son los parámetros de regresión desconocidos.

ii) ε es un error aleatorio con media cero y varianza desconocida.

Tradicionalmente se supone que el error se distribuye como una Normal y que los errores para distintas observaciones no están correlacionados lo que, aunado al supuesto de normalidad, implica que los errores son independientes. Si se cuenta con una muestra de tamaño n de las variables $Y, X_1, X_2, \ldots, X_{p-1}$, esta muestra puede expresarse como

$$\mathbf{Y} = \mathbf{X}\theta + \varepsilon$$

donde,

$\mathbf{Y}_{n\times 1} = (y_1, y_2, \ldots, y_n)'$ un vector de observaciones de la variable Y, dimensión $n \times 1$.

$$\mathbf{X}_{n\times p} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & \cdots & x_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{bmatrix}$$

una matriz de dimensiones $n \times p$ y rango completo (p), que incluye una columna de unos cuando está presente el término independiente en el modelo.

$\theta_{p\times 1} = (\theta_1, \theta_2, .., \theta_p)'$ un vector de parámetros desconocidos de dimensión $px1$.

$E_{n\times 1} = (E_1, E_2, ...E_n)'$ un vector de errores aleatorios $E_i \sim N(0, \tau^{-1})$ independientes.

De los supuestos anteriores se sigue que la distribución de probabilidad o verosimilitud asociada a las observaciones en $\mathbf{Y}$, dados los parámetros τ y θ y las observaciones de las variables explicativas contenidas en $\mathbf{X}$, es una Normal n-variada con vector de medias $\mathbf{X}\theta$ y matriz de covarianzas $\tau^{-1}\mathbf{I}_{n\times n}$. Es decir, el modelo de regresión lineal múltiple se reduce a la estructura

$$\mathbf{Y}|\mathbf{X},\theta,\tau \sim \mathbf{N}_n(\mathbf{X}\boldsymbol{\theta},\tau^{-1}\mathbf{I}_{n\times n}).$$

Análisis de regresión bayesiano

Para proceder al análisis del modelo desde una perspectiva bayesiana, es necesario reflejar las creencias iniciales sobre θ y τ en una distribución de probabilidad a priori: $p(\theta,\tau)$. A partir de esa distribución y mediante la aplicación del Teorema de Bayes se puede obtener la distribución posterior de $\boldsymbol{\theta}$ y τ :

$$P(\theta,\tau|\mathbf{X},\mathbf{Y}) = \frac{P(\mathbf{Y}|\mathbf{X},\theta,\tau)P(\theta,\tau)}{P(\mathbf{Y}|\mathbf{X})} \propto P(\mathbf{Y}|\mathbf{X},\theta,\tau)P(\theta,\tau).$$

En el modelo $P(\boldsymbol{\theta},\tau|\mathbf{Y},\mathbf{X})$ está contenida toda la información disponible sobre los parámetros, tanto la proveniente de la muestra observada como la que describe las creencias a priori. Desde el punto de vista bayesiano, cualquier inferencia que se produzca sobre θ y τ se basa en esta distribución y por lo tanto reflejará necesariamente la información que ésta contiene.

Los modelos de regresión lineal se utilizan fundamentalmente con al menos uno de los siguientes dos propósitos:

i) Analizar la relación de la variable de respuesta con las variables explicativas,

ii) Predecir el valor de la variable de respuesta para un conjunto de valores de las variables explicativas.

En particular respecto al segundo apartado, el problema de pronóstico se puede plantear de la siguiente manera. Sea y^* el valor de la variable respuesta que corresponde al vector fijo y conocido $\mathbf{z}$ de variables explicativas, $(\mathbf{z}_{p\times 1} = (1, z_1, z_2, .., z_{p-1})')$. Entonces,

$$P(y^*|\mathbf{Y},\mathbf{X}) = \int P(y^*|\mathbf{z},\theta,\tau)\, P(\theta,\tau|\mathbf{Y},\mathbf{X})\, d\tau d\theta$$

donde $p(y^*|\mathbf{z},\boldsymbol{\theta},\tau)$ es la distribución de probabilidad condicional asociada a y^*, que bajo el supuesto de normalidad y no correlación de los errores es una Normal univariada con media $\mathbf{z}'\boldsymbol{\theta}$ y varianza τ^{-1}. Para obtener un pronóstico puntual $\hat{y}^*$ del valor que tomará la variable respuesta dado el vector $\mathbf{z}$ se procede como habitualmente en un problema de estimación puntual bayesiano.

Diagnóstico del modelo de regresión lineal

El procedimiento bayesiano para obtener la distribución posterior de los parámetros y producir inferencias es un procedimiento óptimo y por lo tanto la validez de los resultados

depende exclusivamente de que el modelo y los supuestos en los que se basó éste sean correctos. Más en general, en Estadística y bajo cualquier enfoque que se aplique, es absolutamente necesario realizar diagnósticos sobre la validez de los supuestos involucrados al ajustar a una serie de datos un modelo particular.

Típicamente, para diagnosticar un modelo sólo se cuenta con el conjunto de observaciones con las que se ajustó el modelo, bajo el supuesto de que todas provienen del mismo modelo desconocido, y con el modelo ajustado, que, bajo el enfoque bayesiano, consiste básicamente en la distribución posterior tanto de los parámetros de regresión como de la precisión del error: Es decir, se cuenta con $p(\boldsymbol{\theta}, \tau|\mathbf{Y}, \mathbf{X})$.

Para realizar el diagnóstico de un modelo de regresión se pueden adoptar dos diferentes vías.

i) Probar la validez de cada uno de los supuestos por separado: Normalidad, Independencia de los errores, etc.

ii) Confrontar el modelo ajustado con las observaciones y verificar que confirme, es decir, que sea capaz de reconstruir bajo ciertos límites, lo que se observó en la realidad.

Cuando se elige la segunda estrategia, aparecen, de manera natural, dos tipos de observaciones: las *atípicas* y las *influyentes*. Una observación que pertenece a cualquiera de estos dos tipo constituye una señal de que algo puede andar mal con el modelo ajustado. Sobra decir que al detectar observaciones *atípicas* y/o *influyentes* el procedimiento a seguir no consiste en eliminarlas del conjunto de observaciones con las que se ajustó el modelo sino examinar las posibles que les dieron origen.

Detección de observaciones atípicas

El procedimiento sugerido para detectar observaciones atípicas está basado en el supuesto de que las n observaciones provienen del mismo modelo no conocido. Si este supuesto es correcto, sera razonable esperar que si se omite una observación, el modelo ajustado con el resto de las observaciones sea capaz de predecir, dentro de ciertos límites de tolerancia, el valor de la variable respuesta de la observación que se omitió. Así el procedimiento consiste en:

Para cada observación $(y_i, \mathbf{X}_i)$; $i = 1, 2, \ldots, n$.

i) Omitir la i-ésima observación y ajustar el modelo con las observaciones restantes, es decir, obtener la distribución posterior $P(\boldsymbol{\theta}, \tau|(\mathbf{Y}, \mathbf{X})_{-i})$.

ii) Obtener la distribución predictiva $P(y^*|\mathbf{X}_i, (\mathbf{Y}, \mathbf{X})_{-i})$ y un pronóstico del valor de la variable respuesta $\hat{y}^*_{-i}$, dados el vector $\mathbf{X}_i$ y la muestra $(\mathbf{Y}, \mathbf{X})_{-i}$.

iii) Comparar la distribución predictiva $P(y^*|\mathbf{X}_i, (\mathbf{Y}, \mathbf{X})_{-i})$ con el valor observado de la variable respuesta para la observación $i : y_i$.

Por facilidad, y utilizando el hecho de que si se utiliza una distribución a priori Normal-Gamma o la inicial de Jeffreys la distribución predictiva es una t de Student, la comparación indicada en el punto (iii) puede llevarse a cabo evaluando la probabilidad

$$\pi = P\left[|y^*|X_i, (Y, X)_{-i} - \hat{y}^*_{-i}| \geq |y_i - \hat{y}^*_{-i}|\right].$$

Claramente, π es la probabilidad de que la distribución predictiva sin la observación i produzca pronósticos tan extremos o aún más extremos que el valor observado y_i. Si π es pequeña ($\pi < \alpha$; con $\alpha = 0.05$, por ejemplo), se concluye que dada la información contenida en la muestra $(\mathbf{Y}, \mathbf{X})_{-i}$ es improbable que la variable respuesta tome el valor y_i. En otras palabras, la i-ésima observación es atípica (el resto de las observaciones no son capaces de pronosticarla). Como un comentario marginal, vale la pena indicar que desde una perspectiva bayesiana, la determinación del tamaño de α puede considerarse un problema de decisión y resolverse con los métodos disponibles.

Detección de observaciones influyentes

Toda la información sobre los parámetros de regresión y la precisión del error esté contenida en la distribución posterior $P(\boldsymbol{\theta}, \tau|\mathbf{Y}, \mathbf{X})$, y por lo tanto desde el punto de vista bayesiano cualquier medida de la influencia de la i-ésima observación en el modelo ajustado deberá estar basada en una medida de la diferencia entre las distribuciones $P(\boldsymbol{\theta}, \tau|\mathbf{Y}, \mathbf{X})$ y $Q(\boldsymbol{\theta}, \tau|(\mathbf{Y}, \mathbf{X})_{-i})$.

Por supuesto, existen muchas maneras de medir esta diferencia, aquí se propone calcular la divergencia o discrepancia logarítmica entre las distribuciones finales correspondientes como una medida de la influencia de una observación o grupo de observaciones

$$D(P(\theta, \tau|\mathbf{Y}, \mathbf{X}),\ Q(\theta, \tau|(\mathbf{Y}, \mathbf{X})_{-i})) = \int P(\theta, \tau|\mathbf{Y}, \mathbf{X}) \ln \left[\frac{P(\theta, \tau|\mathbf{Y}, \mathbf{X})}{Q(\theta, \tau|(\mathbf{Y}, \mathbf{X})_{-i})} \right] d\tau d\theta.$$

Para facilitar el cálculo de las divergencias se puede trabajar por separado con las distribuciones marginales posteriores:

$$P(\tau|\mathbf{Y}, \mathbf{X}) \text{ y } Q(\tau|(\mathbf{Y}, \mathbf{X})_{-i})\ i = 1, 2, \ldots, n$$

$$P(\theta j|\mathbf{Y}, \mathbf{X}) \text{ y } Q(\theta_j|(\mathbf{Y}, \mathbf{X})_{-i})\ j = 1, 2, \ldots, p;\ i = 1, 2, \ldots, n.$$

Al trabajar por separado con las distribuciones marginales, se mide la influencia de las observaciones para θ_j y τ de manera separada. Una propuesta para medir la influencia global de una observación es utilizar las distribuciones predictivas, $P(y^*|\mathbf{X}_i, \mathbf{Y}, \mathbf{X})$ y $Q(y^*|\mathbf{X}_i, (\mathbf{Y}, \mathbf{X})_{-i})$.

Entonces, una medida de la influencia global de la i-ésima observación es

$$D(P(y^*|\mathbf{X}_i, \mathbf{Y}, \mathbf{X}), Q(y^*|\mathbf{X}i, (\mathbf{Y}, \mathbf{X})_{-i})).$$

Si se calculan los valores de las divergencias correspondientes a cada una de las observaciones ($i = 1, 2, .., n$), se puede establecer por medio de esta medida un orden entre las observaciones basado en su grado de influencia Girón, Martínez y Morcillo (1992). Puesto que la cuestión de interés es detectar las observaciones cuya influencia es significativa para el modelo ajustado, aparece de manera natural la idea de aplicar a las divergencias algunas técnicas usuales del Análisis Exploratorio de Datos para detectar estas observaciones. En particular, se explora la realización de un Diagrama de Caja y Brazos (*Box-Plot)* con los valores de las divergencias y la idea de considerar influyentes a las observaciones cuya divergencia sea extrema por la derecha en el *Box-Plot.*

Conclusiones

Realizar diagnósticos sobre la validez de los supuestos involucrados al ajustar a una serie de datos un modelo particular es inevitable en los procedimientos estadísticos. Recientemente, algunas técnicas habituales de diagnóstico del Modelo de Regresión Lineal han sido exploradas desde una perspectiva bayesiana. La contribución principal de este trabajo es incorporar ciertas ideas del Análisis Exploratorio de Datos al problema de detección de las observaciones que tienen impacto significativo en las conclusiones que se desprenden del modelo ajustado. Estas ideas se han aplicado a datos no sólo provenientes de simulaciones, sino también a datos reales, y funcionan razonablemente bien, aunque en ocasiones pueden conducir a detectar observaciones que no necesariamente son atípicas o influyentes. En cualquier caso se requiere de investigación adicional, tanto para analizar los efectos combinados de las observaciones que son simultáneamente atípicas e influyentes, como para establecer, en forma clara, un procedimiento secuencial de diagnostico.

Referencias

Girón, F.J., Martínez, L. and Morcillo, C. (1992) *A Bayesian Justification for the Analysis of Residuals and Influence Measures. In : Bayesian Statistics 4*, (Bernardo, J.M. et al. eds), Oxford University Press, Oxford.

Modelando heterogeneidad en análisis de supervivencia: una aplicación en genética animal

BELEM TREJO VALDIVIA
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México
FELIPE RUIZ LÓPEZ
Centro Nacional de Investigaciones en Fisiología y Mejoramiento Animal, INIFAP-SAGDR

Introducción

En ganado productor de leche, la duración de vida productiva (tiempo desde que la vaca pare por primera vez hasta que es desechada del establo; DVP) ha sido relacionada con la rentabilidad total del establo. Cambios en la DVP pueden afectar la productividad a través de su influencia sobre la edad promedio del hato y de su efecto directo sobre la tasa de reemplazo. Las vacas Holstein alcanzan su madurez productiva (y con esto su producción más elevada) aproximadamente a las 3 lactancias, sin embargo la edad promedio de vacas Holstein en México se estima alrededor de 2.06 lactancias (Ruiz et al., 1994), por lo que se hace necesario identificar a los factores que intervienen en la determinación de la DVP en México. El problema radica entonces en cómo modelar el comportamiento de la DVP de vacas Holstein en México que nos permita evaluar la magnitud de las diferentes fuentes de variación y la posibilidad de incrementar la DVP a través de programas de mejoramiento genético.

Estudios anteriores (Abubakar et al., 1987; Ruiz 1991) han relacionado la supervivencia del ganado productor de leche en México con el valor genético para leche del semental de la vaca. Estos estudios concluyeron que hijas de sementales con valores genéticos promedio para leche alcanzaron mayores longevidades que las hijas de sementales con valores genéticos altos o bajos. Además, se ha relacionado la DVP con el país de origen del semental, el nivel de producción estandarizado y con efectos medio ambientales temporales representados por los efectos de año-hato.

Uno de los problemas en el modelaje de DVP es la existencia de datos censurados, ya sea porque no se cuenta con la información completa de la vaca (como cuando cambia de establo o cuando el establo se sale del programa que recopila información) o porque la vaca se encuentra en producción al momento del estudio, siendo éste último caso el más importante ya que son estas vacas las que nos interesan más en los programas de mejoramiento genético.

Recientemente se han propuesto procedimientos estadísticos para analizar la posible heterogeneidad (no observable) de la población bajo estudio (Clayton y Cuzick, 1986; Trejo, 1991, entre otros). Dichos procedimientos permiten evaluar la necesidad de incluir términos aleatorios extra en los modelos para la función de riesgo en datos de supervivencia.

El objetivo de este trabajo fue el de determinar si es necesario incluir más términos en el modelo que explica la DVP de vacas Holstein en México.

Material y métodos

Se utilizaron registros de producción y longevidad de 36477 vacas Holstein, proporcionados por la Asociación Holstein de México A.C., que parieron entre los años de 1970 y 1992 y que contaran cuando menos con 10 medias hermanas paternas cuyas DVP hubiesen sido observadas (no-censuradas). Se ajustó un modelo Weibull estandarizado de riesgos convergentes de Clayton-Cuzick (Clayton y Cuzick, 1986) que incluyó a los efectos de año-hato (variables *dummy* que indican el año y hato en el que se presentó el primer parto; en total 771 años-hato), nivel de producción estandarizado (se consideran 9 categorías equiprobables, desde una muy mala productora hasta una muy buena productora) y país de origen del semental (Estados Unidos de América, Canadá o México). Con el objeto de simplificar el análisis se transformaron datos para eliminar el parámetro de escala de la distribución Weibull, de manera que el modelo fuera el siguiente:

$$\lambda\left(t \mid \underline{z}, \xi\right) = \rho t^{\rho-1} \exp\left[\beta' \underline{z}\right] \xi, \tag{1}$$

donde:

$$E\left(\xi\right) = 1, \ \mathrm{Var}\left(\xi\right) = \theta.$$

Después de ajustar los efectos anteriores, se aplicó la prueba de heterogeneidad y se examinó la posibilidad de representar la variación aún no explicada por medio del semental ajustando un segundo modelo que, además de los efectos del modelo anterior, incluyó al padre de la vaca (variables dummy para identificar 589 sementales). La prueba de heterogeneidad consiste en probar $H_0 : \theta = 0$ vs. $H_a : \theta > 0$.

La estadística de prueba es:

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} Y_i^2 - \delta_i Y_i \right).$$

Bajo H_o,

$$\sqrt{n}\tau \rightarrow N\left(0, \nu\right),$$

con

$$\nu = E\left(\delta Y^2\right).$$

Resultados

El nivel de censura encontrado en la información fue del 15%, lo que es inferior a lo reportado anteriormente por Ruiz et al. (1994) por haber utilizado una definición diferente de censura. Los resultados del modelo (1) se encuentran en el cuadro 1.

Las hijas de toros Estadounidenses tendieron a permanecer en el hato 1 mes más que las de toros Canadiense o Mexicanos. Se encontró una relación cuadrática entre nivel de producción y DVP. La prueba de heterogeneidad tuvo un valor de 3.461, lo que determinó la necesidad de incluir un término extra en el modelo ($P < .01$) y se concluyó que el semental puede representar a dicho término ($P < .01$) ya que el valor de la estadística de prueba calculada fue de 0.781. El siguiente paso será ajustar al semental como un efecto aleatorio y determinar la proporción heredable de esta característica bajo este tipo de modelos.

Cuadro 1: Estimadores y coeficientes de regresión para el modelo de riesgos proporcionales sin incluir al semental.

	Modelo sin semental [1]		Modelo con semental [2]	
Parámetro	Estimador	Error estándar	Estimador	Error estándar
ρ	1.684	0.452 E-4	1.670	0.567 E-4
País origen del semental				
EUA	-0.070	0.225 E-3	-0.072	0.307 E-3
Canadá	-0.003	0.331 E-3	-0.004	0.466 E-3
Nivel de prod.				
1	-0.347	0.520 E-3	-0.368	0.668 E-3
2	-0.611	0.496 E-3	-0.635	0.637 E-3
3	-0.671	0.479 E-3	-0.678	0.615 E-3
4	-0.694	0.471 E-3	-0.714	0.602 E-3
5	-0.688	0.472 E-3	-0.700	0.605 E-3
6	-0.647	0.472 E-3	-0.664	0.604 E-3
7	-0.597	0.469 E-3	-0.609	0.600 E-3
8	-0.468	0.483 E-3	-0.471	0.624 E-3
Año-hato prom.	0.307	0.038	0.297	0.044
Efecto semental promedio				
EUA			0.028	
Canadá			0.037	

Referencias

Abubakar, B.Y., McDowell, R.E. and VanVleck, L.D. (1987) "Interaction of genotype and environment for breeding efficiency and milk production of Holsteins in Mexico and Colombia", *Trop. Agric.*, **64**, 1.

Clayton, D.G., and Cuzick, J. (1986) *The semi-parametric Pareto model for regression analysis of survival times,* Papers on Semiparametric Models at the ISI Centenary Session Amsterdam, Gill, R.D. and Voors, M.N., Center for Mathematics and Computer Science, Amsterdam.

Cox, D.R. (1972) "Regression models and life-tables (with Discussion)", *J.R. Statist. Soc., B,* **34**, 187.

Ducrocq, V. (1994) "Statistical analysis of length of productive life for dairy cows of the Normande breed", *J. Dairy Sci.,* 77–855.

Ducrocq, V., Quaas, R.L., Pollak, E.J. and Casella, G. (1988) "Length of productive life of dairy cows. 1. Justification of a Weibull model". *J. Dairy Sci.*, 71–3061.

Ruiz, F.J. (1991) *Relationships among length of productive life, milk yield and profitability of US, Canadian and Mexican Holstein sires in Mexico,* Ph.D. Dissertation, Cornell University, Ithaca, New York, U.S.A.

Ruiz, F.J., Oltenacu, P.A. y Blake, R.W. (1994) "Efecto del nivel de producción de leche sobre la duración de vida productiva de ganado Holstein en México", Tec. Pec. en México, Técnica Pecuaria en México, 32–33.

Trejo-Valdivia, G.M.B. (1991) *Proportional Hazards Models with Heterogeneous Treatment Effects,* Ph. D. Tesis no publicada, University of London.

Un procedimiento para la selección de variables por componentes principales

José Vences Rivera
Intituto Nacional de Estadística Geografía e Informática, Aguascalientes

Introducción

En el contexto del análisis estadístico multivariado, una de las técnicas más utilizadas es la de *componentes principales*, que tiene como propósito fundamental describir la dispersión de un conjunto de datos en un espacio p-dimensional, desde otro espacio de menor dimensión, mediante nuevas variables, es decir, componentes principales, formadas por combinaciones lineales de las variables originales sujetas a restricciones. No obstante que se reduce la dimensión, cuando se toman solamente unas cuantas componentes principales, la interpretación es con base en el conjunto total de variables originales, lo cual constituye una limitante, sobre todo cuando se dificulta la interpretación de las combinaciones.

En muchos problemas prácticos el interés no solamente es reducir la dimensionalidad espacial, sino el número de variables a ser consideradas en estudios posteriores. En la actualidad existen varios procedimientos para seleccionar o descartar variables, como es el mismo análisis de componentes principales bajo ciertas consideraciones (Jolliffe, 1972) y el análisis de factores, entre otros, pero no han sido del todo satisfactorios.

La finalidad de este trabajo es presentar un procedimiento para la selección de variables mediante la aplicación de la técnica de componentes principales, utilizando los criterios de la *varianza generalizada* y de la *traza*. Se presentan también los supuestos básicos para el desarrollo de dos programas de cómputo de fácil operación (mismos que se encuentran disponibles en la versión original del artículo), a través de los cuales se puede obtener un subconjunto de variables representativas.

Marco teórico

Sean los vectores aleatorios $\underset{\sim}{X}_{p\times 1}$ y $\underset{\sim}{\mathbf{Y}}_{k\times 1}$ $(k < p)$. Este último dado mediante la transformación

$$\underset{\sim}{\mathbf{Y}}_{k\times 1} = A^t_{k\times p}\, \underset{\sim}{X}_{p\times 1}$$

Para el caso de la selección de variables, se tiene que la matriz A debe ser de la forma

$$A = [I, 0]^t,$$

donde la matriz identidad I es de orden $(k \times k)$ y la matriz de ceros 0 es de orden $k \times (p-k)$; o también se puede obtener una matriz permutando los renglones de A, que es equivalente a permutar los elementos de $\underset{\sim}{X}$, lo cual resulta más sencillo. Así, considérense todas las

posibles particiones de $\underset{\sim}{X}$ de la forma

$$\underset{\sim}{X} = \begin{bmatrix} \underset{\sim}{X}_1 \\ \underset{\sim}{X}_2 \end{bmatrix},$$

donde $\underset{\sim}{X}_1$ es el vector de las k variables seleccionadas y $\underset{\sim}{X}_2$ es el vector de las $p-k$ variables descartadas.

Con esto, la matriz de varianzas y covarianzas se particiona como

$$\Sigma_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Por lo tanto, el problema de seleccionar k variables es equivalente a seleccionar la matriz Σ_{11} de entre las $\binom{p}{k}$ combinaciones posibles. Para esto, dos de los criterios de optimización más aceptados tanto por los resultados obtenidos como por la sencillez del cálculo, son el del *determinante* y el de la *traza*, es decir,

$$1.\ \max |\Sigma_{11}| = \prod_{i=1}^{K} \lambda_i = \min |\Sigma_{22.1}|$$

$$2.\ \max \operatorname{tr}(\Sigma_{11}) = \sum_{i=1}^{K} \lambda_i = \min \operatorname{tr}(\Sigma_{22.1})$$

donde los λ_i's son los primeros k eigenvalores ordenados (de mayor a menor) de Σ_X, correspondientes a las primeras k componentes principales. En el primer caso se expresa la varianza desde el punto de vista multivariado. Por otra parte, $\Sigma_{22.1}$ es la matriz de varianzas y covarianzas de las variables no seleccionadas dadas las seleccionadas, que viene dada como

$$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

En Seber (1994) puede encontrarse un tratamiento detallado de estas expresiones.

Finalmente, la bondad del subjconjunto de variables seleccionadas puede ser evaluada a través del porcentaje de la variación explicada (McCabe, 1984):

$$P(\%) = \left(1 - \frac{\sum_{i=1}^{p-k} \theta_i}{\sum_{i=1}^{p} \lambda_i} \right) \times 100,$$

donde los θ_i's son los eigenvalores asociados a $\Sigma_{22.1}$.

Sistema de cómputo

Se desarrollaron dos programas de cómputo utilizando los criterios del *determinante* y de la *traza*, donde se aprovecha la relación entre el número k de variables a seleccionar y el

número de componentes principales que explican un determinado porcentaje de la varianza total inherente al conjunto original de variables.

En general, cuando el porcentaje de varianza requerido es suficiente grande, el número de variables seleccionadas que explican aproximadamente el mismo porcentaje de varianza obtenida por componentes principales, es igual o ligeramente mayor al número de éstas.

Así, dado el porcentaje requerido de varianza explicada se compara con el obtenido por las primeras componentes principales, si las cantidades son similares entonces k es igual al número correspondiente de componentes principales, de lo contrario, k es igual a este número más uno o dos.

Después de obtener k, se procede a calcular $|\Sigma_{11}|$ (también se puede utilizar $|R_{11}|$) de entre todas las $\binom{p}{k}$ alternativas posibles, seleccionando aquellas variables asociadas a $\max|\Sigma_{11}|$; o bien, para el caso del criterio de la *traza*, se procede a calcular $\text{tr}(\Sigma_{22.1})$ para todas las posibilidades, y se seleccionan aquellas variables asociadas a $\min \text{tr}(\Sigma_{22.1})$.

Finalmente, se calcula el porcentaje de la varianza total explicada por las variables seleccionadas, el cual en la mayoría de los casos prácticos resulta igual o mayor al requerido.

Algunas consideraciones para el cálculo eficiente en la selección de variables con características similares a las presentadas en este trabajo, pueden verse en Furnival (1971) y McCabe, Jr. (1975).

Conclusiones

La selección de variables por los criterios del *determinante* y de la *traza*, constituye una herramienta útil en la reducción de dimensionalidad, sobre todo cuando el número de variables consideradas es grande, donde algunas de ellas presentan cierto grado de correlación, o bien, no aportan de manera importante a la varianza total subyacente al sistema.

La selección de variables por los dos criterios mencionados, puede implementarse en una computadora de manera relativamente fácil. El tiempo de CPU depende fundamentalmente del número de variables consideradas originalmente, así como del porcentaje requerido de varianza explicada y del número de observaciones (aunque éste no tiene importancia en comparación de los dos primeros).

Se recomienda utilizar el criterio del *determinante* cuando la meta principal es la retención de la varianza; en tanto que el criterio de la *traza* es conveniente para propósitos de predicción.

Para una selección final de variables, se recomienda correr varias veces los programas de cómputo como los que se mencionan en este trabajo, aumentando o disminuyendo el porcentaje de varianza requerido; así como la utilización de otra técnica como *análisis de factores*. Esto en combinación con elementos de juicio: la interpretación y el costo de medición de alguna variable en aplicaciones posteriores, entre otros.

Referencias

Furnival, G.M. (1971) "All possible regressions with less computation", *Technometrics*, **13**, 2, 403–408.

Jolliffe, I.T. (1972) "Discarding variables in a principal components analysis", I: Artificial data, Applied statistics, *Journal of the Royal Statistical Society*, Ser. C., 160–173.

Seber, G.A.F. (1984) *Multivariate observations*, Ed. John Wiley & Sons, New York.

McCabe, G.P. (1984) "Principal variables", *Technometrics*, **26**, 2, 137–144.

McCabe, G.P. Jr. (1975) "Computations for variable selection in discriminant analysis, *Technometrics*, **17**, 1, 103–109.

El hombre más alto del mundo

José Aurelio Villaseñor Alba
Colegio de Postgraduados
Barry C. Arnold
University of California, Riverside, CA, USA

Introducción

Si cada individuo en un proceso de ramificación tiene un atributo medible asociado (por ejemplo su altura), es de interés estudiar el comportamiento probabilístico del valor máximo del atributo en una generación dada y en todo el proceso. En este trabajo se identifican las distribuciones asintóticas asociadas y se muestra que satisfacen la ecuación funcional de Schroder.

Considérese un proceso ramificado de Galton-Watson $\{X_n\}_{n=1}^{\infty}$ en donde se supone que los tamaños de las familias son variables aleatorias independientes e idénticamente distribuidas (iid) cuya distribución común tiene función generatriz de probabilidades (fgp) $P_z(s)$. Entonces

$$X_n = \sum_{i=1}^{X_{n-1}} Z_{i,n-1}, \tag{1}$$

en donde las $Z_{i,n-1}$ son variables aleatorias iid con función generatriz de probabilidades $P_z(s) = E(s^2)$.

Además supóngase que cada miembro de la población tiene una característica asociada la cual es medible, digamos Y_i, que pudiera ser por ejemplo la altura individual. La altura del hombre más alto del mundo en la n-ésima generación, estará dada por

$$M_n = \max_{i \leq X_n} Y_i, \tag{2}$$

en donde se supone que las Y_i son iid con función de distribución común dada por F, la cual se supone que tiene soporte en $[0, \infty)$.

En este trabajo se pone especial atención a la distribución asintótica de M_n definida por (2), es decir la distribución asintótica de la altura del hombre más alto del mundo cuando n tiende a infinito. En este contexto, nuestro objetivo es entonces identificar las condiciones apropiadas para la distribución de los tamaños de las familias (con fgp $P_z(s)$) y para la distribución de la altura de los individuos $F(y)$ que permita la derivación de la distribución (no degenerada) límite para M_n, convenientemente estandarizada. También se espera conocer la clase de distribuciones límites resultantes.

Algunos resultados útiles de procesos ramificados

Por conveniencia, se supone que $P(X_o = 1) = 1$, es decir el proceso inicia con un solo ancestro. De aquí se sigue que la fgp de X_n es igual a la n-ésima iteración de la fgp del

tamaño de las familias. Esto es

$$\mathrm{P}_{X_n}(s) = \mathrm{P}_Z^{(n)}(s). \tag{3}$$

Para evitar la extinción de la población se supone que $\mathrm{E}(Z) = \mu > 1$. En estas condiciones se tiene que

$$X_n/\mu^n \xrightarrow{\text{c.s.}} X_\infty, \tag{4}$$

en donde X_∞ es una variable aleatoria tal que su función generatriz de momentos (fgm) $\varphi_\infty(t) = \mathrm{E}(e^{-tX_\infty})$ satisface la ecuación funcional

$$\varphi_\infty(\mu t) = \mathrm{P}_Z(\varphi_\infty(t)). \tag{5}$$

La ecuación (5) es conocida como la ecuación de Schroder (ver Kuczma, 1968). La forma general de la solución de (5) es dada por

$$\varphi_\infty(t) = \lim_{n\to\infty} \mathrm{P}_Z^{(n)}\left(e^{-t/\mu^n}\right). \tag{6}$$

Desafortunadamente no es fácil evaluar iteraciones de la fgp en general, por lo que la expresión (6) no puede ser usada para obtener expresiones cerradas para $\varphi_\infty(t)$, aunque nos permitirá obtener ciertas características de la distribución límite.

El caso excepcional se tiene cuando fgp es bilineal, es decir,

$$\mathrm{P}_z(s) = (1 - \gamma + (\gamma - q)s)/(1 - qs), \tag{7}$$

en donde $0 < 1 - q < \gamma < 1$ (para asegurar que $\mu > 1$). En este caso se puede verificar que

$$\varphi_\infty(t) = \left(1 + \frac{(1-\gamma)\,t}{\gamma + q - 1}\right) / \left(1 + \frac{qt}{\gamma + q - 1}\right). \tag{8}$$

La expresión (8) puede ser reconocida como la fgp de una variable aleatoria X_∞ la cual es igual a cero con probabilidad $(1-\gamma)/q$ (que es la probabilidad de extinción a largo plazo) y es igual a una variable aleatoria exponencial con media $q/(\gamma + q - 1)$ con probabilidad $(\gamma + q - 1)/q$ (ver Harris, 1948).

El hombre más alto del mundo

La altura del individuo más alto en la n-ésima generación es dada por

$$M_n = \max_{i \le X_n} Y_i, \tag{9}$$

en donde las Y_i son variables aleatorias iid con función de distribución común F y X_n es el tamaño de la población.

Si condicionamos respecto X_n, entonces la función de distribución de M_n es dada por

$$\begin{aligned} \mathrm{F}_{M_n}(x) &= \mathrm{E}(\mathrm{P}(M_n \le x \mid X_n)) = \mathrm{E}(\mathrm{F}(x)^{X_n}) \\ &= \mathrm{P}_{X_n}(\mathrm{F}(x)) = \mathrm{P}_Z^{(n)}(\mathrm{F}(x)). \end{aligned} \tag{10}$$

Es natural esperar que la distribución límite de M_n, estandarizada apropiadamente, dependerá del comportamiento de la cola del lado derecho de F y de la distribución límite de X_n.

Así, supóngase que F pertenece al dominio de atracción maximal de una distribución extrema G (uno de los tres tipos posibles). Por lo tanto existen funciones $a(t) > 0$ y $b(t)$ tales que

$$\lim_{t\to\infty} [\mathrm{F}(a(t)x + b(t))]^t = \mathrm{G}(x). \tag{11}$$

Sea $x_n = a(\mu^n)x + b(\mu^n)$ en donde $\mu = \mathrm{E}(Z) > 1$. Por las ecuaciones (10), (11) y la convergencia uniforme en (6) se tiene que

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{M_n - b(\mu^n)}{a(\mu^n)} \leq x\right) = \lim_{n\to\infty} \left(\mathrm{P}_z^{(n)} \exp\left(-\frac{(-\log \mathrm{F}(x_n)^{\mu^n})}{\mu^n}\right)\right) \tag{12}$$

$$= \varphi_\infty(-\log \mathrm{G}(x))$$

en donde φ_∞ satisface (6). Si se denota por F_∞ a la función de distribución de X_∞ entonces la distribución límite de (12) se puede expresar como

$$\int_0^\infty (\mathrm{G}(x))^\theta \, d\mathrm{F}_\infty(\theta).$$

En el caso cuando el tamaño de la familia es una variable aleatoria geométrica, es decir, cuando (7) se cumple con $\gamma = 1$, se tiene que $\varphi_\infty(t) = (1+t)^{-1}$ la cual corresponde a una variable exponencial estándar. Las posibles leyes límites correspondientes para M_n son $\mathrm{H}_1(x) = (1 + X^{-\alpha})^{-1}$, $x > 0$; $\mathrm{H}_2(x) = (1 + (-x)^\alpha)^{-1}$, $x < 0$ y $\mathrm{H}_3(x) = (1 + e^{-x})^{-1}$, $-\infty < x < \infty$. Nótese que la última corresponde a la distribución logística estándar.

En Galambos (1992) se encuentra una revisión de bibliografía muy completa sobre distribuciones límites y caracterizaciones basadas en máximos geométricos de variables iid.

Referencias

Galambos, J. (1992) *Characterizations.* In *Handbook of the Logistic Distribution,* N. Balakrishnan, ed., Decker, New York, 169–188.

Harris, T.E. (1948) "Branching processes", *Annals of Mathematical Statistics,* **19**, 474–494.

Kuczma, M. (1968) *Functional equations in a single variable,* PWN-Polish Scientific Publications, Warsaw.

Bondad de ajuste para muestras censuradas

José Salvador Zamora Muñoz
UACPyP del CCH e Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México

Introducción

[1]El estudio de los métodos para probar supuestos distribucionales de una población mediante procesos estadísticos, recibe el nombre de Bondad de Ajuste. En el caso de que estas pruebas se utilicen para muestras completas se han estudiado de manera extensiva por muchos autores. Existen extensiones a estas pruebas cuando se trabaja con una muestra censurada, especialmente con censura controlada por el investigador. Sin embargo, son pocas las extensiones anteriores para el caso en que se tiene una muestra con censura aleatoria, que es la más común en poblaciones de seres vivos. El objetivo principal de este trabajo es dar un panorama de los trabajos que se han realizado en lo referente a muestras censuradas.

De las diferentes pruebas de Bondad de Ajuste que existen, estaremos interesados en las pruebas basadas en la función de distribución empírica, en particular en las estadísticas del tipo Cramér-von Mises, que son de la forma

$$Q = n \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 \, \Psi(x) \, dF(x),$$

con $F_n(x)$ la distribución empírica de la muestra, $F(x)$ la distribución hipotética y $\Psi(x)$ una función que asigna pesos a las diferencias cuadráticas $\{F_n(x) - F(x)\}^2$. Cuando $\Psi(x) = 1$ tenemos la estadística Cramér-von Mises, W^2. Cuando $\Psi(x) = \{F(x)(1 - F(x))\}^{-1}$ tenemos la estadística de Anderson-Darling, A^2. Una modificación a la estadística W^2, es la estadística de Watson definida como:

$$U^2 = n \int_{-\infty}^{\infty} \left\{ F_n(x) - F(x) - \int_{-\infty}^{\infty} (F_n(t) - F(t)) \, dF(t) \right\}^2 dF(x).$$

Hay un uso generalizado de estas estadísticas cuando la muestra es completa, por lo tanto no incluiremos más sobre ellas. La parte de interés para nosotros será la metodología que se usa para calcular la distribución asintótica de estas estadísticas.

Distribución asintótica de las estadísticas

La metodología que se usa para obtener la distribución asintótica de cada estadística es similar a la que se usa en estadística para realizar la descomposición de un vector aleatorio en sus componentes principales, en este caso es el proceso estocástico, asociado con cada estadística, el que se descompone en componentes principales, de la siguiente manera.

[1]Parte de la tesis de Maestría en Estadística e Investigación de Operaciones, UACPyP del CCH e IIMAS-UNAM.

Se puede descomponer la función de covarianza $K(s,t)$ como

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j f_j(s) f_j(t),$$

con λ_j el eigenvalor de la función de covarianza y f_j su correspondiente eigenfunción asociada, ambas son solución de la ecuación integral

$$\lambda f(t) = \int_0^1 K(s,t)\ f(s)\, ds.$$

Se pueden construir las correspondientes componentes principales

$$Z_j = \int_0^1 \mathbf{Y}(t)\, f_j(t)\, dt$$

de donde se obtiene finalmente que

$$\int_0^1 \mathbf{Y}^2(t)\, dt = \sum_{j=1}^{\infty} \lambda_j Z_j^2$$

con Z_j^2 variables aleatorias i.i.d. $\chi^2_{(1)}$.

Esta descomposición se usa para encontrar la función característica de cada estadística, la cual se puede invertir para encontrar la correspondiente función de distribución. En D'Agostino y Stephens (1986), se puede encontrar tablas de cuantiles para cada una de estas estadísticas.

Hasta aquí se ha considerado solamente muestras completas. Ahora veamos las extensiones de esta metodología para los casos donde se tiene una muestra censurada. Introduciremos estas ideas a partir de considerar los tipos más comunes de censura que ocurren en el área estadística conocida como Análisis de Supervivencia.

Censura tipo I y tipo II

Estos tipos de censura se catalogan como censura controlada por el investigador. En el primer caso se determina un tiempo límite de observación, por lo que las observaciones que no presentaron el evento de interés se consideran como censuras. En el segundo, el investigador observa hasta que ocurren r fallas de n posibles, nuevamente el resto de las observaciones resultan censuradas.

Para estos dos tipos de censuras Pettitt y Stephens (1976) proponen las siguientes modificaciones a las estadísticas anteriores:

$$_{q,p}W_n^2 = \int_q^p Y_n^2(t)\, dt,$$

$$_pU_n^2 = \int_0^p \{Y_n(t) - 1/p \int_0^p Y_n(s)\, ds\}^2\, dt,$$

$$_{q,p}A_n^2 = \int_q^p \frac{Y_n^2(t)}{t(1-t)}\, dt,$$

con $Y_n(t) = \sqrt{n}\{F_n(t) - F(t)\}$, $0 < q$, $p < 1$ y $q \leq t \leq p$.

La metodología que se presentó anteriormente, para encontrar la distribución asintótica de las estadísticas, puede extenderse a este caso, considerando ahora el proceso estocástico definido en $[q, p]$ en lugar de $[0, 1]$ como en los casos anteriores.

Censura aleatoria

Este tipo de censura ocurre independientemente de la voluntad del investigador y representa los casos en que el individuo abandona el estudio, muere por una causa que no es de interés, permanece aun vivo al finalizar el estudio, etc.

La única propuesta que existe para realizar pruebas de bondad de ajuste usando una estadística del tipo Cramér-von Mises con este tipo de censura es la de Koziol-Green (K-G), que proponen para la distribución de la censura el modelo

$$(1 - H) = (1 - F^0)^\beta.$$

El parámetro β puede interpretarse como un parámetro de censura; $\beta = 0$ corresponde a no censura.

La estadística de prueba propuesta por K-G, es

$$\Psi_n^2 = \int_0^1 Y_n^2(t)dt = n \int_0^1 \{\hat{F}_n^0(t) - F(t)\}^2 \, dt.$$

$\hat{F}_n^2(t)$ es el estimador Kaplan-Meier. Este estimador es el equivalente a la función de distribución empírica cuando se tiene una muestra censurada.

Nuevamente, podemos realizar la descomposición en componentes principales de Ψ^2, y a partir de esta descomposición encontrar la función característica de esta estadística e invertirla numéricamente por el método de Imhof (1961). Koziol-Green (1976) muestran los valores asintóticos de esta estadística para distintos valores del parámetro de censura β.

Referencias

D'Agostino, R.B. and Stephens M.A. (1986) *Goodness of fit Tecniques*, Marcel Dekker, Inc., New York.

Imhof, J.P. "Computing the distribution of cuadratic forms in normal variables", *Biometrica*, **48**, 419–426.

Koziol, J.A. and Green, S.B. (1976) "A cramér-von Mises for randomly censored data", *Biometrica* , **63**, 465–474.

Pattitt, A.N. and Stephens, M.A. (1976) "Modified Cramér-Von Mises statistics for censored data", *Biometrika* , **63**, 291–298.

Shorack, G.R. and Wellner, J.A. (1986) *Empirical Processes with Applications to Statistics*, John Wiley, New York.

Una prueba de bondad de ajuste para un proceso puntual de Poisson no homogéneo

MARTÍN ZAVALA LEÓN
Universidad Autónoma de Sinaloa
VÍCTOR MANUEL PÉREZ-ABREU CARRIÓN
Centro de Investigación en Matemáticas, A.C.

Introducción

En este trabajo se construye una prueba de bondad de ajuste para probar si un proceso puntual es Poisson no homogéneo basado en la funcional generatriz de probabilidades para el proceso, esto es, para el proceso de Poisson, el logaritmo de su funcional generatriz de probabilidades es lineal con respecto a una variable real ρ, por lo que la segunda derivada es cero en todo ρ.

Los procesos estocásticos puntuales modelan el número de puntos que caen en cualquier región no importando el orden de ocurrencia. La región de importancia puede estar en la recta real, un plano o el espacio. Entre los procesos estocásticos puntuales, el proceso de Poisson no homogéneo es de gran importancia. De aquí la necesidad de contar con una prueba de bondad de ajuste para decidir cuándo un proceso puntual es Poisson.

Es importante mencionar que en este trabajo suponemos un esquema de muestreo de repeticiones de un proceso puntual en un espacio fijo. Así, los resultados asintóticos de nuestra prueba son respecto al número de repeticiones del proceso.

Proceso de Poisson

En la teoría de los procesos puntales, el proceso de Poisson es considerado como el más importante, debido a que es un buen modelo para diversos fenómenos que ocurren en la naturaleza, ingeniería y ciencias sociales.

Definición. Un proceso puntual en $\mathcal{X}$ es un *Proceso de Poisson* si existe una medida $\Lambda(\cdot)$ sobre $\mathcal{B}_{\mathcal{X}}$, finita sobre cada conjunto acotado, tal que para cada subfamilia finita de $\mathcal{B}_{\mathcal{X}}$ $\{A_i, i = 1, \ldots, k\}$ de conjuntos acotados y ajenos se tiene que

$$\Pr\{N(A_i) = n_i, i = 1, \ldots, k\} = \prod_{i=1}^{k} \frac{\Lambda(A_i)^{n_i}}{n_i!} \exp\{-\Lambda(A_i)\}. \tag{1}$$

A la función de conjuntos $\Lambda(\cdot)$ la llamamos *medida parámetro* o *medida de intensidad del proceso.* La ecuación (1) incluye dos casos especiales cuando $\mathcal{X} = \mathbb{R}^d$: para un *proceso de Poisson homogéneo* $\Lambda(A) = \lambda \ell(A)$, donde ℓ es la medida de Lebesgue, y para el *proceso de Poisson no homogéneo*, $\Lambda(A) = \int_A \lambda(x)dx$, para el caso en que $\Lambda(\cdot)$ tiene densidad $\lambda(\cdot)$.

La definición anterior implica las siguientes características de un proceso de Poisson:

(a) El número de puntos en cada conjunto A_i tiene distribución de Poisson con parámetro $\Lambda(A_i)$; y

(b) Las variables aleatorias $N(A_i)$ y $N(A_j)$ son independientes si $A_i \cap A_j = \phi$ para $i \neq j$.

Para el caso especial en que $\mathcal{X}$ es la recta real, existe una propiedad que especifica a un proceso de Poisson Homogéneo, llamada especificación por intervalo la cual establece que:

(c) Iniciando en un punto origen, los tiempos entre ocurrencias $X_1, X_2, \ldots$ de puntos sucesivos son independientes e idénticamente distribuidos con distribución exponencial con parámetro λ.

La funcional generatriz de probabilidades

Definición. Sea $\mathcal{U}$ la clase de las funciones h: $\mathcal{X} \to C$ Borel medibles que satisfacen la condición $|h(x)| \leq 1$. Dado un proceso puntual finito, definimos a su *funcional generatriz de probabilidades (fl.g.p.)* en $h \in \mathcal{U}$ por

$$G[h] = E\left(\prod_{i=1}^{N} h(x_i)\right), \tag{2}$$

donde el producto es uno si $N = 0$, y es cero si $N > 0$ y $h(x_i) = 0$ para algún i.

La fl.g.p. para el proceso de Poisson, que es nuestro caso de interés, la podemos expresar por argumentos estándar como

$$G[h] = \exp\left\{-\int_{\mathcal{X}} (1 - h(x))\Lambda(dx)\right\}.$$

Si además $0 < \rho < 1$, y $1 - h \in \mathcal{V}(\mathcal{X})$, entonces $1 - \rho h \in \mathcal{V}(\mathcal{X})$ y

$$\begin{aligned} G[1 - \rho h] = &\exp\left\{-\int_{\mathcal{X}} (1 - (1 - \rho h(x)))\Lambda(dx)\right\} \\ = &\exp\left\{-\rho \int_{\mathcal{X}} h(x)\Lambda(dx)\right\}. \end{aligned} \tag{3}$$

Prueba propuesta

Definición. *Sean $N_1, \ldots, N_n$ copias independientes de un proceso puntual arbitrario N con localizaciones de puntos correspondientes $\{x_{ij} : j = 1, \ldots, n;\ i = 1, \ldots, N_j(\mathcal{X})\}$, la* Funcional Generatriz de Probabilidades Empírica (fl.g.p.e.) $\hat{G}_n[h]$ *está dada por*

$$\hat{G}_n[h] = \frac{1}{n} \sum_{j=1}^{n} \prod_{i=1}^{N_j(\mathcal{X})} h(x_{ij}), \qquad h \in \mathcal{V}(\mathcal{X}). \tag{4}$$

La base de la prueba es la forma de la fl.g.p. Recordemos que el logaritmo de la fl.g.p. para un proceso puntual de Poisson tiene la forma

$$Y[h] = \log \mathrm{G}[1-h] = -\int_{\mathcal{X}} h(x)\Lambda(dx), \qquad h \in \mathcal{V}(\mathcal{X}), \tag{5}$$

donde Λ es la medida de intensidad del proceso.

Retomamos la idea de Castro y Pérez-Abreu (1994) al introducir una variable real, $0 \leq \rho \leq 1$, y para $h \in \mathcal{V}(\mathcal{X})$ tomar las funciones de la forma $\rho h \in \mathcal{V}(\mathcal{X})$, con lo que (5) queda expresada como

$$Y_\rho[h] = \log \mathrm{G}[1-\rho h] = -\rho \int_{\mathcal{X}} h(x)\Lambda(dx), \qquad h \in \mathcal{V}(\mathcal{X}).$$

La idea es ver a $Y_\rho[h]$ como una función de ρ para h fija, lo cual es una línea recta. Vista $Y_\rho[h]$ de este modo y con h fija, tenemos que una manera de caracterizar a una línea recta es que su segunda derivada con respecto a ρ es idénticamente 0. Esta es la caracterización en la que nos basamos para construir la prueba.

Trabajamos entonces con el estimador $\hat{Y}_{n,\rho}[h]$ de $Y[h]$, y bajo la hipótesis de Poisson la segunda derivada respecto de ρ de $\hat{Y}_{n,\rho}[h]$ debe estar cerca de cero para todo $0 \leq \rho \leq 1$. Tenemos

$$\hat{Y}_{n,\rho}[h] = \log \hat{G}_n[1-\rho h].$$

Debido a la forma tan complicada de la segunda derivada respecto a ρ nos restringiremos a evaluarla en $\rho = 0$. Aunque esto parece una restricción importante, los resultados de simulación muestran que no lo es. Así,

$$\begin{aligned} Y''_{n,0}[h] \;=\; & \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{\Big(\int h(x)N_j(dx)\Big)^2 - \int h^2(x)N_j(dx) - \\ & \int h(x)N_j(dx)\int h(x)N_i(dx)\}, \end{aligned} \tag{6}$$

entonces, escribiendo

$$Z_j[h] = \int h(x)N_j(dx),$$

obtenemos

$$Y''_n[h] = \frac{1}{n}\sum_{j=1}^{n}\Big\{(Z_j[h]-\Lambda[h])^2 - Z_j[h^2]\Big\} - \Big(\overline{Z}[h]-\Lambda[h]\Big)^2, \tag{7}$$

donde $\Lambda[h] = EZ[h]$.

Con lo anterior, estamos probando si $\dfrac{\partial^2}{\partial\rho^2}Y_\rho[h] = 0$ en $\rho = 0$ para toda función $h \in \mathcal{V}(\mathcal{X})$, por lo que podemos elegir h's y rechazar la hipótesis de que el proceso es Poisson para valores

grandes de $\frac{\partial^2}{\partial\rho^2}\hat{Y}_{n,\rho}[h]$ valuada en $\rho = 0$. Sin embargo, tenemos una mejor opción con una prueba multivariada si construimos el vector aleatorio

$$Y_n''[h_1,\ldots,h_k] = (Y_n''[h_1],\ldots,Y_n''[h_k])^T, \tag{8}$$

que toma en cuenta varias funciones $h's$ simultáneamente. Entonces obtenemos la estadística de prueba la partir de

$$Z_n[h_1,\ldots,h_k] = \left(\sqrt{n}Y_n''[h_1,\ldots,h_k]\right)^T \Sigma^{-1}_{Y''[h_1,\ldots,h_k]} \left(\sqrt{n}Y_n''[h_1,\ldots,h_k]\right), \tag{9}$$

donde $\Sigma_{Y''[h_1,\ldots,h_k]}$ es la matriz Covarianza de

$$\begin{aligned} Y''[h_1,\ldots,h_k] = &\ (Y''[h_1],\ldots,Y''[h_k])^t \\ = &\ \left((Z[h_1]-\Lambda[h_1])^2 - Z[h_1^2],\ldots,(Z[h_k]-\Lambda[h_k])^2 - Z[h_k^2]\right)^T, \end{aligned} \tag{10}$$

sustituyendo esta matriz por un estimador de ella. Así, tenemos el siguiente resultado asintótico:

Teorema. *Bajo la hipótesis de que el proceso puntual es Poisson, la estadística dada por*

$$D_n[h_1,\ldots,h_k] = \left(\sqrt{n}Y_n''[h_1,\ldots,h_k]\right)^T \hat{\Sigma}^{-1}_{Y''[h_1,\ldots,h_k]} \left(\sqrt{n}Y_n''[h_1,\ldots,h_k]\right) \tag{11}$$

tiende en distribución a una Ji-cuadrada con k grados de libertad, χ^2_k y los elementos de $\hat{\Sigma}^{-1}_{Y''[h_1,\ldots,h_k]}$ están dados por $\widehat{Cov}(Y_1''[h_i], Y_1''[h_j]) = 2\hat{\Lambda}^2[h_ih_j]$.

En conclusión, la prueba que proponemos depende de la elección de funciones $h's$, del número de éstas y de la estadística $D_n = D_n[h_1,\ldots,h_k]$ dada por (11). Es decir, bajo la hipótesis de un proceso Poisson, D_n debe estar cercana a 0 cuando n es grande, luego primero elegimos k funciones $h's$ adecuadas para la hipótesis alternativa y rechazamos para valores grandes de D_n.

Resultados y conclusiones

Para estimar la potencia de la prueba se realizaron simulaciones sobre procesos de renovación y procesos de cúmulos y de Gauss-Poisson en $\mathbb{R}$ y $\mathbb{R}^2$. La prueba tiene buena potencia para todos los procesos mencionados en $\mathbb{R}^2$, incluso mejor que la potencia de la prueba de McDonald. Sin embargo, para los procesos en $\mathbb{R}$ tiene malas potencias con las funciones $h's$ con las que se intentó.

Por las razones anteriores concluimos que la prueba se puede usar cuando el fenómeno que se está modelando está en $\mathbb{R}^2$ y si el fenómeno se encuentra en $\mathbb{R}$ lo primero que recomendamos hacer es una prueba de Lilliefors que tiene muy buena potencia aunque la hipótesis a probar es que el proceso es Poisson homogéneo.

Referencias

Cox, D.R. and Isham V. (1980) *Point Processes.* Monographs on Applied Probability and Statistics, Chapman and Hall, New York.

Daley, D.J. and Vere-Jones, D.(1988) *An Introduction to the Theory of Point Processes,* Springer Series in Statistics, Springer-Verlag, New York.

Kingman, J.F.C. (1993) *Poisson Processes,* Oxford Science Publications, Oxford University Press, Oxford.

Ripley, B.D. (1981) *Spatial Statistics,* Wiley, New York.

Castro, J.I. and Pérez-Abreu, V. (1994) "Some Statistical Analyses of the Cluster Point Processes of Hurricanes on the Coast of México". *Statistics for the Environment 2: Water Related Issues*, John Wiley & Sons, 95–108.

Chouinard, A. and McDonald, D. (1985) "A Characterization of Nonhomogeneous Poisson processes", *Stochastics*, **15**, 113–119.

Davies, R.B. (1977) "Testing the Hypothesis that a Point Process is Poisson", *Advances in Applied Probability*, **9**, 724–746

McDonald, D. (1989) "On Nonhomogeneous, Spatial Poisson Processes", *The Canadian Journal of Statistics*, **17**, 2, 183–195.

Nakamura, M. and Pérez-Abreu, V. (1993a) "Empirical Probability Generating Function. An Overview", *Insurance: Matematics and Economics* **12**, 287–295.

Nakamura, M. and Pérez-Abreu, V. (1993b) "Exploratory Data Analysis for Counts using the Empirical Probability Generating Function", *Communications in Statistics - Theory and Methods*, **22**, 3, 827–842.

Nakamura, M. and Pérez-Abreu, V. (1993c) "Use of an Empirical Probability Generating Function for Testing a Poisson Model", *The Canadian Journal of Statistics*, **21**, 2, 149-156.