Memorias
del
XVIII
Foro Nacional de Estadística

INEGI
www.inegi.gob.mx

Memorias
del
XVIII
Foro Nacional de Estadística

INEGI
www.inegi.gob.mx

Memorias del XVIII Foro Nacional de Estadística

Impreso en México
ISBN 970-13-3653-4

Presentación

El XVIII Foro Nacional de Estadística se llevó a cabo de 13 al 17 de octubre de 2003 en el conjunto Amoxcalli de la Facultad de Ciencias de la Universidad Nacional Autónoma de México, siendo dicha facultad quien organizó el evento.

En estas memorias se presentan algunos de los resúmenes de las contribuciones presentadas en este foro. Todos los resúmenes recibidos fueron incluidos sin un proceso de arbitraje, aunque con una detallada revisión.

La Asociación Mexicana de Estadística agradece a la Facultad de Ciencias de la UNAM su apoyo para la realización de este foro y al Instituto Nacional de Estadística, Geografía e Informática el apoyo para la edición de estas memorias.

El Comité Editorial:

Carlos Díaz Avalos

Antonio González Fragoso

Karim Anaya Izquierdo

Contenido

Intervalos de confianza exactos para el cociente de momios

Rebeca Aguirre Hernández[1]

Federico O'Reilly[2]

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

1. Introducción

Cuando se miden dos variables binarias, V_1 y V_2, a un conjunto de n individuos independientes, los datos suelen presentarse en una tabla de contingencia de 2×2. A menudo interesa determinar si las dos variables son independientes. La prueba exacta de Fisher se usa con esta finalidad si el número total de individuos estudiados no es suficientemente grande como para aplicar métodos asintóticos. El procedimiento consiste en hacer un análisis condicional a las frecuencias marginales de la tabla observada. En este caso, solo una de las frecuencias conjuntas de la tabla, X, puede considerarse como una variable aleatoria; el resto se calculan a partir de X y de las frecuencias marginales como lo muestra la siguiente tabla.

Tabla 1: Tabla de contingencia de 2×2 con frecuencias marginales fijas.

	V_2 Exito	V_2 Fracaso	Total
V_1 Exito	X	$n - X$	n
V_1 Fracaso	$t - X$	$m - t + X$	m
	t	$n + m - t$	$n + m$

Las hipótesis $H_0 : V_1$ y V_2 son independientes *vs.* $H_1 : V_1$ y V_2 no son independientes son equivalentes a $H_0 : \phi = 1$ *vs.* $H_1 : \phi \neq 1$ donde ϕ es la razón de momios. En lugar de probar estas hipótesis se puede construir un intervalo de confianza para ϕ. Si el intervalo contiene al uno se concluye que V_1 y V_2 son independientes.

[1] rebeca@sigma.iimas.unam.mx
[2] federico@sigma.iimas.unam.mx

2. Intervalos de Confianza para la Razón de Momios

Considere la tabla de contingencia mostrada arriba. La variable aleatoria X se distribuye, al condicionar en las frecuencias marginales de la tabla, como una hipergeométrica no central (Fisher (1935)):

$$P(X = j; \phi, n, m, t) = \frac{\binom{n}{j}\binom{m}{t-j}\phi^j}{\sum_{u=x_0}^{x_f}\binom{n}{u}\binom{m}{t-u}\phi^u} \tag{1}$$

donde $x_0 = max\{0, t-m\} \leq j \leq x_f = min\{n, t\}$ y $\phi \geq 0$. A partir de esta distribución, el nivel de significancia correspondiente a la prueba de hipótesis $H_0 : \phi = \phi^*$ *vs.* $H_1 : \phi > \phi^*$ se calcula como:

$$\sum_{x=x_{obs}}^{x_f} P(X = x \mid n, m, t, \phi) \tag{2}$$

donde $P(X = x \mid n, m, t, \phi)$ está dada por la expresión (1) y x_{obs} es el valor observado de X. Análogamente, el nivel de significancia asociado a las hipótesis $H_0 : \phi = \phi^*$ *vs.* $H_1 : \phi < \phi^*$ se calcula como:

$$\sum_{x=x_0}^{x_{obs}} P(X = x \mid n, m, t, \phi) \tag{3}$$

Cornfield (1956) propuso usar las expresiones (2) y (3) para calcular un intervalo exacto al $(1-\alpha) \times 100\,\%$ de confianza para ϕ. El límite inferior, ϕ_{I1}, es aquel valor de ϕ tal que la expresión (2) es igual a $\alpha/2$. El límite superior del intervalo es el valor de ϕ_{S1} que satisface con la restricción de que la ecuación (3) es igual a $\alpha/2$. Es decir, el intervalo exacto de Cornfield es (ϕ_{I1}, ϕ_{S1}) donde ϕ_{I1} y ϕ_{S1} se obtienen por ensayo y error.

El nivel de significancia de la prueba de hipótesis y la probabilidad de cobertura del intervalo de Cornfield son conservadores, es decir, mayores al nivel preestablecido por el investigador. Esto se debe a que X es una variable aleatoria discreta. Para atenuar este problema se ha propuesto (Berry y Armitage (1995)) aplicar una corrección conocida como "mid-P value" que consiste en multiplicar $P(X = x_{obs} \mid n, m, t, \phi)$ por 0.5. Es decir, al momento de calcular el nivel de significancia y los límites del intervalo para ϕ, las expresiones (2) y (3) se reemplazan por (4) y (5) respectivamente:

$$\frac{1}{2}P(X = x_{obs}) + \sum_{x=x_{obs}+1}^{x_f} P(X = x) \tag{4}$$

$$\sum_{x=x_0}^{x_{obs}-1} P(X = x) + \frac{1}{2}P(X = x_{obs}) \tag{5}$$

Existen otros intervalos frecuentistas propuestos para ϕ, ver por ejemplo Agresti y Min (2001).

3. Intervalo Fiducial para la Razón de Momios

R. Fisher fue el precusor de la estadística fiducial pero este enfoque ha tenido poca aceptación. Para el caso de variables aleatorias discretas pertenecientes a la familia exponencial, O'Reilly (2003) propuso definir la función de distribución fiducial como la siguiente combinación lineal convexa:

$$H(\theta \mid x) = \begin{cases} 1 - P(X \leq x) & \text{si } x = x_0 \\ \alpha[1 - P(X \leq x)] + (1 - \alpha)[1 - P(X < x)] & \text{si } x_0 < x < x_f \\ 1 - P(X < x) & \text{si } x = x_f \end{cases}$$

donde $\alpha \in [0, 1]$. Generalmente se usa $\alpha = 1/2$. La derivada de $H(\theta \mid x)$ con respecto a θ da

como resultado la función de densidad fiducial $h(\theta \mid x) = dH(\theta \mid x)/d\theta$. Si X se distribuye como una hipergeométrica no central, $h(\phi \mid x)$ es el cociente de dos polinomios en ϕ. Note que $h(\phi \mid x)$ es una función de ϕ.

Existen dos métodos para construir un intervalo para la razón de momios con base en $h(\phi \mid x)$. Uno de ellos consiste en encontrar los valores ϕ_{I2} y ϕ_{S2} tales que el area de $h(\phi \mid x)$ a la izquierda de ϕ_{I2} y a la derecha de ϕ_{S2} sea igual a $\alpha/2$. Es decir:

$$\int_0^{\phi_{I2}} h(\phi \mid x)d\phi = \int_{\phi_{S2}}^{\infty} h(\phi \mid x)d\phi = \frac{\alpha}{2}$$

Con base en diversos ejemplos se encontró que el intervalo resultante: (ϕ_{I2}, ϕ_{S2}) coincide con el intervalo exacto de Cornfield cuando se aplica la corrección del "mid-P value".

Otro método para construir un intervalo fiducial para ϕ consiste en trazar una línea horizontal que intersecte a $h(\phi \mid x)$. Los puntos en que dicha línea intersecta a $h(\phi \mid x)$ se denotarán como ϕ_{I3} y ϕ_{S3}. Se dirá que (ϕ_{I3}, ϕ_{S3}) es un intervalo fiducial con un nivel de confianza de $(1-\alpha) \times 100\,\%$ si

$$\int_{\phi_{I3}}^{\phi_{S3}} h(\phi \mid x)d\phi = 1-\alpha \qquad \text{donde} \qquad h(\phi_{I3} \mid x) = h(\phi_{S3} \mid x) = y$$

La longitud de los intervalos asi obtenidos es menor o igual a la longitud de los intervalos construidos mediante el método descrito previamente.

4. Ejemplo

Considere una tabla de 2×2 en la que todas las frecuencias marginales son iguales a 4 *ie.* $n = m = t = 4$. En este caso, la variable aleatoria X puede tomar los valores: $\{0, 1, 2, 3, 4\}$. La Tabla 2 muestra, para cada posible valor de X, la razón de momios observada, el intervalo exacto de Cornfield, el intervalo corregido de Cornfield y el intervalo fiducial construidos con

un nivel de confianza de 95 %.

De acuerdo con los dos primeros intervalos, existe independencia si $X \in \{1, 2, 3\}$. En cambio, el intervalo fiducial indica que las variables son independientes si $X > 0$. Los intervalos fiduciales son los más angostos. El intervalo corregido de Cornfield y el intervalo fiducial son idénticos cuando $X = 0$. El intervalo exacto de Cornfield y su versión corregida tienen longitud infinita cuando $X = 4$. Además, estos dos intervalos son invariantes ante transfomaciones monótonas de ϕ. Es decir, si (ϕ_I, ϕ_S) es el intervalo para ϕ entonces $(\phi_S^{-1}, \phi_I^{-1})$ es el intervalo para ϕ^{-1}. Esta propiedad queda ejemplificada en la Tabla 2 al comparar los intervalos exactos de Cornfield obtenidos cuando $X = 0$ y $X = 4$ así como los intervalos obtenidos cuando $X = 1$ y $X = 3$. Los intervalos corregidos de Cornfield presentan el mismo comportamiento.

Tabla 2: Intervalos al 95 % de confianza para la razón de momios de una tabla de 2×2 con todas las frecuencias marginales iguales a 4.

	Razón de	**Cornfield**		**Cornfield corregido**		**Fiducial**	
X	Momios	Inferior	Superior	Inferior	Superior	Inferior	Superior
0	0.0	0.0	0.747	0.0	0.499	0.0	0.5
1	0.11	0.0016	4.72	0.00325	3.23	0.0	2.09
2	1	0.033	30.6	0.05	19.9	0.000156	12.238
3	9	0.21	626	0.31	309	0.00374	149.32
4	∞	1.34	∞	2	∞	0.0425	307.5178

Referencias

Agresti, A. y Min, Y. (2001). On Small-Sample Confidence Intervals for Parameters in Discrete Distributions, *Biometrics*, **57**, 963-971.

Berry, G. y Armitage, P. (1995). Mid-P Confidence Intervals: A Brief Review, *The Statistician*, **44**, 417-423.

Cornfield, J. (1956). A Statistical Problem Arising From Retrospective Studies, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 135-148.

Fisher, R.A. (1935). The Logic of Inductive Inference, *J. Roy. Statist. Soc.*, **98**, 39-82.

O'Reilly, F. (2003). Significance Distributions. *Comunicaciones Internas del IIMAS, UNAM.*, Preimpreso No. 117.

Software and applications in multiresponse variables

Guillermo Bali Chávez[1]
Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Ciudad de México
Dariusz Czerski[2]
Akademia Podlaska,
Siedlce, Poland
Andrzej Matuszewski[3]
Institute of Computer Science,
Polish Academy of Sciences
Warsaw, Poland

1. Introduction

The research in multiresponse variables has grown enormously in the last years because its importance for the analysis of different types of questionnaires. Without doubt, this methodology will go gaining practical space for databases analysis of greater complexity. The problem extensively has been referred since the perspective of a categorical variable and one genuine multiresponse variable by Agresti and Liu (1999), Bilder et al. (2000), Decady and Thomas (2000) and Loughin and Sherer (1998). The authors introduce a chi-square statistics to find that dependence exists among the variables. From this point, the search of a significant correlation between two multiple variables has been the objective, among others, of the investigations of Bali and Matuszewski (1998), Matuszewski and Trojanowski (2001) and Bali et al. (2003). In all of them have been discussed with amplitude, the possibility to define what would be understood for "independence" for this type of variables. The statistics used in these papers are referred to the chi-square type too, although there are an important separation between votes and persons. This allowed to divided the algorithms in traditional and democratic schemes. The introduction of various schemes of simulation for votes and persons in the methodology of analysis in the articles referred, prompted the development

[1]gbali@itesm.mx
[2]DAREQ@2com.pl
[3]amat@ipipan.waw.pl

of sophisticated software because it was necessary for our study the implementation of the algorithms in C^{++} and JAVA.

2. Samples schemes

We will present some examples of sampling schemes developed for this research. We will refer to the case when the two questions have three options, but the methodology is implemented in JAVA, for any number of options inside the questions. Besides it is included the algorithms for a traditional (votes) or democratic (persons) counting method, Bali et al. (2003). The statistics *phi* (Yule-Pearson coefficient) is calculated for the real sample in all schemes and is compared with the 95 % and 99 % percentiles , obtained with the bootstrap-type procedures. Monograph oriented for the area of discrete statistics offer different distribution as model of data behavior. Once the data matrix fulfills these requirements under the null hypothesis, the chi square distribution for X^2 can be applied to test this hypothesis.

2.1. Total schema

In this schema there are 2^A probabilities, that correspond to the form of answer of the question 1, where A is the number of options for question 1. This probabilities should sum 1 since collect all the possible combinations of answer. For the question 2 the parameters that are defined are totally analogous to the first question.

2.2. Total modified schema

For the example case, we take two questions with three options, we can find the patterns of responses $(1,1)$, $(1,2)$, $(2,1)$ and $(2,2)$ that are generated in automatic form, where (A,B) is the number of answers that gave a person to each one of the questions. This are the patterns that really influence in the correlation. The case of the persons that answered all options or did not choose anything are excluded. The generation method splits in principle a negative independence. This aspects have been discussed exhaustively for Bali et al. (2003).

2.3. MN2 schema

In this case exist A and B possibilities to choose each one of the answers respectively. Here the probabilities should not sum necessarily one. In most cases this add belongs to the interval $(1, A)$ and $(1, B)$.

2.4. Hypergeometric schema

It is analogous to the total modified schema and be built from the same previous patterns, but with another generation method from the marginal probabilities, the result is in principle a positive independence.

3. Methods to define "equivalent" parameters

We can have a sample of persons that answered a two "pick any" questions. It could be possible that exist correlation between the two answers. From the real sample the parameters of the four schemes can be calculated. Another variant is define the parameters $P_{1,}P_2$ and P_3 for one schema and derive the parameters of the others. There are mathematical transformations but the form are not always trivial. For example in the case of the Hypergeometric Schema, the relation that exists between the parameters from the algorithm and the probabilities of the real sample $Real_1$, $Real_2$ *and* $Real_3$ is not explicit. For example, the transformation formula for the case 3X3 are given by:

$$
\begin{aligned}
Real_1 &= \frac{P_1}{2} + \frac{1}{2}\left(P_2\,\frac{P_1}{1-P_2} + P_3\,\frac{P_1}{1-P_3}\right) \\
Real_1 &= \frac{P_2}{2} + \frac{1}{2}\left(P_1\,\frac{P_2}{1-P_1} + P_3\,\frac{P_2}{1-P_3}\right) \\
Real_1 &= \frac{P_3}{2} + \frac{1}{2}\left(P_1\,\frac{P_3}{1-P_1} + P_2\,\frac{P_3}{1-P_2}\right)
\end{aligned}
$$

In the case of the Total Modified Schema, we obtain:

$$\begin{aligned} Real_1 &= \frac{P_3 + P_2}{2} \\ Real_1 &= \frac{P_1 + P_3}{2} \\ Real_1 &= \frac{P_1 + P_2}{2} \end{aligned}$$

where $P_1 + P_2 + P_3 = 1$. These formulas allow us to connect the two schemes and their outcomes.

4. Simulation results

We realized a research about the use and customs to read between university students located in the Tlalpan delegation. The random sample size was equal to 120 persons and we receive 102 valid questionnaires that was answered by 59 % of females, 41 % of males, 75 % students of Administrative and Humanity careers and 25 % of Engineering careers. We take two multiresponse questions for the questionnaire and we obtain the following results:

For the question, What type of publications are you custom to read?, the answers was 61 % Books, 57 % Magazines and 34 % Internet Webs.

For the question, In which place do you custom to read?, the answers was 79 % in their own room, 31 % in the school and 18 % in the point of purchase

The following table contains the values for the Statistic, 95 % and 99 % percentiles by bootstrap for the Total, Total Modified $[(1,1) = 27, (1,2) = 15, (2,1) = 20, (2,2) = 16]$ and MN2 $[(0{,}55, 0{,}6125, 0{,}3) \times (0{,}8625, 0{,}325, 0{,}2125)]$ Schemes.

The number of bootstrap simulation was equal to 60,000 (Democratic Counting).

Schema	Total	Total Modified	MN2
Statistic	5.64	5.64	5.64
95 % Percentil	3.37	3.80	2.49
99 % Percentil	4.75	5.33	3.56

The outcomes of the table show that exist a correlation among the two questions of the questionnaire, because the statistic is always bigger than the percentiles values, both of them. In the real sample can be established a significant dependence between the publications that the persons custom to read and the places that they used to read for Mexican students for the south of the city. Relevant is that the results of the MN2 schema are qualitative different of the other two schemes. This behavior is very interesting and indicate that there are maximal p-value for this correlation problem.

5. Final remarks

The high degree of difficulty that represents to find a p-value,that measure statistical dependence between multiresponse variables has derived in the development of new software to explore the problem since different angles. This paper shows examples of some of the sampling schemes that have been implemented. The use of bootstrap as our form of approximation to the real distribution of the sample was a decisive step to estimate p-values that formalize the independence of the pair of questions in a meaningful way . It is of capital importance to understand that using bootstrap methodology is possible to generate independent samples with marginal probabilities. We would like to add, that the methodology of residues would contribute to additional results and should be to account in our future research.. We must indicate that the aspects that are presented in this research have a strong connection with Data Mining, because both of them allow the analysis of real databases.

References

Agresti A., Liu L-M.(1999). Modelling a categorical variable allowing arbitrary many category choices. *Biometrics*, 55, 936-943.

Bali G., Matuszewski A.(1998). Un modelo de interdependencia entre respuestas múltiples aplicado a la natación de alto rendimiento. Memorias del XIII Foro Nacional de Estadística, Monterrey N.L., pp. 103-108.

Bali G., Matuszewski A., Klopotek M. (2003). Dependence of Two Multiresponse Variables: Importance of The Counting Method. In *Intelligent Information Processing and Web Mining.* pp. 251-260. Springer.

Bilder C., Loughin T., Nettleton D. (2000). Multiple marginal independence testing for pick any/c variables. *Commun. Statist-Simula*, 29, 1285-1316.

Decady Y., Thomas D. (2000). A simple text of association for contingency tables with multiple column responses. *Biometrics*, 56, 893-896.

Loughin T., Scherer P. (1998). Testing association in contingency with multiple column responses. *Biometrics*, 54, 630-637.

Matuszewski A, Trojanowski K. (2001). Models of Multiple Response Independence . *Intelligent Information Systems.* Physica-Verlag (Springer), pp. 209-219.

Optimización multirrespuesta en procesos industriales: una propuesta gráfica

Jorge Domínguez Domínguez[1]
Rosa M. Rocha Morales
Centro de Investigación en Matemáticas, A.C.

1. Introducción

El diseño de experimentos se ha aplicado de manera importante para mejorar la calidad de procesos y productos, adicionalmente para hacer que estos productos sean robustos ante condiciones extremas. Las características de respuesta para evaluar la calidad tienen un enfoque multirrespuesta. Es frecuente encontrar muchas aplicaciones industriales con varias respuestas. La finalidad es alcanzar la calidad global de un producto, por lo que es necesario optimizar de manera simúltanea las respuestas de interés. En esencia, el problema de optimización de varias respuestas involucra la selección de un conjunto de condiciones o variables independientes tales que den como resultado un producto o servicio conveniente. Es decir, se desea seleccionar los niveles de las variables independientes que optimicen todas las respuestas a la vez.

Este trabajo tiene por objetivo mostrar un método gráfico como una alternativa a los métodos analíticos para optimizar de manera global las respuestas que definen la calidad de un producto. Una de las ventajas del método gráfico es que permite generar varias escenarios de posibles soluciones óptimas tales que impacten en la reducción de costos.

2. Planteamiento del problema

Se integró un grupo interdiscipliario con el fin de realizar un proyecto de desarrollo tecnológico para aprovechar los residuos agro industriales. La idea principal del proyecto consiste en

[1] jorge@cimat.mx

construir un extrusor cuya función será elaborar forraje para ganado. Varios factores intervienen en este proceso de extrusión, se tiene interés en investigar mediante un experimento que factores son significativos para el proceso. Con esta información se realizá un segundo experimento para determinar que valores en los diferentes factores dan lugar a un alimento óptimo. Para optimizar el proceso en cada tratamiento experimental se midieron diferentes respuestas.

Varios esquemas experimentales se pueden plantear para este proyecto, tales como, diseños factoriales, diseños factoriales fraccionados, diseño Box - Benhken o diseño central compuesto Box y Draper (1987). Cada una de las respuestas se pueden modelar con los resultados al aplicar alguno de estos diseños. Por lo general estos modelos son lineales y están en función de los factores. Así para n respuestas se tienen n modelos, el i-ésimo modelo para esa respuesta Y_i se escribe como:

$$Y_i = \beta_0 + X^t\beta + X^tBX + \varepsilon \tag{1}$$

donde X^t =$(X_1, \ldots, X_k)$ es un vector de k factores, β_0 la constante, $\beta = (\beta_1, \ldots, \beta_k)$ un vector de parámetros $B = (\beta_{11}, \ldots, \beta_{1k}, \beta_{k1}, \ldots, \beta_{kk})$ matriz de parámetros de segundo orden, y $\varepsilon \sim N(0, \sigma^2)$.

El problema consiste en determinar la mejor combinación de los factores tal que produzcan el óptimo global. Es decir que todas las respuestas den su mejor valor. Esta situación se puede plantear como sigue:

$$\begin{array}{ll} \text{Optimizar} & Y_1 \\ \text{Sujeto a} & Y_2 = L_1 \\ & \vdots \\ & Y_n = L_{n-1} \\ & X \in R\,,\ R : \text{región experimental} \end{array} \tag{2}$$

donde los L_i $(i = 1, \ldots, n-1)$ representan consideraciones importantes o restricciones para las respuestas.

3. Procedimientos de optimización

La expresión (2) describe el planteamiento típico de un problema de programación lineal, la solución es un valor X_0 de X, que genera una respuesta óptima global bajo estas condiciones. Entre las ventajas de este procedimiento es su planteamiento matemático y que puede ser resuelto mediante una hoja de cálculo. Existen otros procedimientos analíticos eficientes, cómo el de la función de deseabilidad (DE) propuesto por Derringer y Suich, R.(1980) analizado y aplicado por Derringer(1994). El método citado permite crear varios esenarios para posibles soluciones. El procedimiento denominado función distancia (DI) fue propuesto por Khuri y Colon (1981), con éste también se obtienen una solución óptima puntual. Varios investigadores han producido trabajos interesantes en esta dirección, ver Del Castillo et.al (1996). Un efonque que considera a la función de pérdida (PE) es propuesto por Ames et.al. (1997).

No obstante que el método gráfico (MG) es un procedimiento descritivo, éste contiene o ilustra a todos los resultados que se obtienen con los métodos anteriormente citados. Éste funciona relativamente fácil ante situaciones cuando existen dos o tres factores, se complica un poco con cuatro. Se considera como una buena práctica que en la estrategias de optimización primero se realizan experimentos para eliminar factores que aportan poca infomación a la respuesta de interés. Así que en la estrategia experimental para obtener un óptimo se trabaja con un número reducido de factores.

4. El método gráfico

Para ilustrar el método gráfico se utilizará un problema descrito por Box-Draper (1987). Éste hace referencia a un diseño experimental del tipo factorial completo 3^3, el cual se aplicó con el propósito de estudiar la capacidad de una imprenta para imprimir tinta de color en unas etiquetas. Se considera que tres factores tienen efecto en la impresión de la tinta. Los factores

y sus niveles:

Factores	Niveles		
	-1	0	1
x_1 : velocidad	30	45	60
x_2 : presión	90	110	130
x_3 : distancia	12	20	28

Por cuestiones de espacio remitimos a lector a consultar los datos en el libro de los autores citados. Con los resultados reportados al aplicar el diseño se pueden obtener los modelos para la media y la desviación estándar. El siguiente paso es la optimización conjunta de estas respuestas. Es importante hacer notar que la optimización conjunta de la media y la varianza es relevante en muchos procesos. Ya que una solución óptima permitirá mejorar la característica del proceso (óptimo de la media) y con la mínima varianza (óptimo de la varianza).

A partir de la expresión (1), los modelos para la media y varianza son: $\widehat{Y}_1 = \widehat{\mu}(x)$ y $\widehat{Y}_2 = \widehat{\sigma}(x)$ respectivamente. Los modelos que se obtienen al ajustar la media y la desviación estándar se describen en las dos siguientes exrpesiones:

$$\begin{aligned}\widehat{Y}_1 = \widehat{\mu}(x) = \quad 327.6 \quad & +177.0\,x_1 \quad +109.4\,x_2 \quad +131.5\,x_3 \\ & +32.0\,x_1^2 \quad \text{-}22.4\,x_2^2 \quad \text{-}29.1\,x_3^2 \\ & +66.0\,x_1\,x_2 \quad +75.5\,x_1\,x_3 \quad +43.6\,x_2\,x_3\end{aligned}$$

y

$$\begin{aligned}\widehat{Y}_2 = \widehat{\sigma}(x) = \quad 34.9 \quad & +11.5\,x_1 \quad +15.3\,x_2 \quad +29.2\,x_3 \\ & +4.2\,x_1^2 \quad \text{-}1.3\,x_2^2 \quad +16.8\,x_3^2 \\ & +7.7\,x_1\,x_2 \quad +5.1\,x_1\,x_3 \quad +14.1\,x_2\,x_3.\end{aligned}$$

Se grafican ambos modelos, las curvas de nivel que se generan de cada respuesta se sobreponen. Como se tienen tres variables de entrada, lo que se hace para ver posibles soluciones, es considerar el plano de un par de factores y dejar fijo el otro factor en un nivel. En la gráfica a la izquierda de la Figura 1 se muestra el plano (X_1,X_3) y el valor $X_2 = 0$ para el factor 2. Un conjunto de soluciones factibles se encuentran en la región de intersección. Esta representación se hace en el espacio como se ilustra en la gráfica a la derecha de la Figura 1.

Se puede notar que al ir cambiando el valor del factor 2 se generará una región en el espacio de posibles soluciones óptimas. Esta situación se observa en la Figura 2.

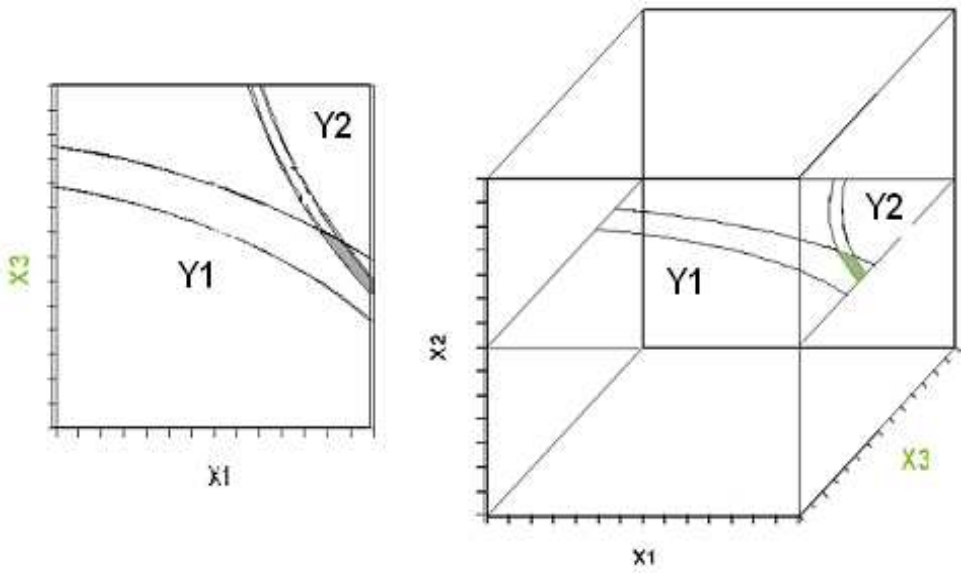


Figura 1: Solución en el plano (X_1, X_3) y $X_2 = 0$; corte en el espacio con $X_2 = 0$.

En la Figura 2 se muestra el potencial del método gráfico, ya que la región que se obtiene a partir de este método da lugar a un esenario de soluciones factibles para optimizar las dos respuestas a la vez. Además se ilustra las soluciones que generan algunos de los métodos analíticos citados.

En la Tabla 1 se describen los resultados óptimos que se obtienen a partir de diferentes métodos, (soluciones óptimas $X_{i0}, i = 1, 2, 3$, $Y_{j0}, j = 1, 2$). La mejor solución que se plantea para este proceso es que la media tenga un valor en 500 con la menor varianza. Tanto en la Figura 2 como en la Tabla 1 se puede apreciar que la solución que presentan los métodos la función de distancia (DI) y la función de pérdida (FP) caen fuera de la región experimental. La solución generada por la función de deseabilidad esta en el espacio de soluciones propuestas por el procedimiento gráfico. A nivel general se puede concluir que en este caso el método gráfico tiene una menor varianza. Además se puede explorar en la región experimental valores apropiados de los factores donde se consiga reducir costos de operación, cuestión que no es fácil en los otros métodos.

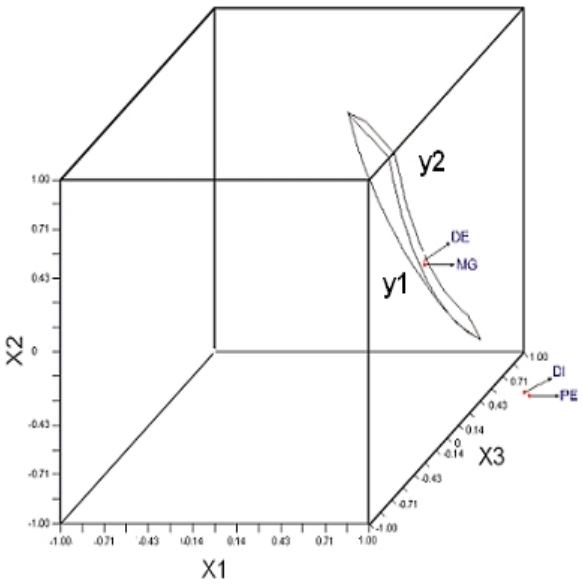


Figura 2: Región de soluciones factibles: método gráfico; óptimos generados por métodos analíticos.

Tabla 1: Soluciones óptimas generadas al aplicar los diferentes procedimientos

Métodos	X_{10}	X_{20}	X_{30}	Y_{10}	Y_{20}
MG método gráfico	1.00	0.14	-0.29	500.0	44.7
DE función deseabilidad	1.00	0.17	-0.30	500.0	45.1
PE función de pérdida	1.57	-0.72	-0.09	500.0	45.1
DI función distancia	1.60	-0.62	-0.19	501.9	45.5

5. Discusión

A nivel de conclusiones resaltaremos algunas ventajas que se consideran importantes del método gráfico sobre los otros procedimientos. En referencia al método gráfico se puede decir que, es visual, práctico y amigable. Por otro lado, presenta soluciones competitivas con respecto a los métodos analíticos. Resulta muy útil porque puede aplicarse en procesos cotidianos. Evidentemente es una guía para los responsables de los procesos con el fin de disminuir costos, seleccionando niveles de los factores que den lugar a una combinación de tratamientos más económica. Por su carácter gráfico es un auxiliar para detectar la presencia de varias regiones factibles. Ayuda a generar mayor conocimiento del proceso bajo estudio. Finalmente, la integración del método gráfico con los métodos analíticos proporcionan una alternativa potencial para dar solución a los problemas multirrespuesta.

Referencias

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

Box, G. E. P. and Draper, N. R. (1987). Empirical Model Building and Response Surfaces. John Wiley & Sons, New York, NY.

Del Castillo, E., Montgomery, D. C. and McCarville, D. R. (1996). "Modified Desirability Functions for Multiple Response Optimization". *Journal of Quality Technology* 28, pp. 337-345.

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

Derringer, G. (1994). "A Balancing Act: Optimizing a Product's Properties". *Quality Progress* pp. 51-58.

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

Inferencia paramétrica para una clase de cadenas de Markov

Gabriel Escarela [1]
Universidad Autónoma Metropolitana,
Unidad Iztapalapa

1. Introducción

Las cadenas de Markov pueden ser aplicadas a varios problemas de ciencia actuarial y de control estocástico. Estas cadenas pueden usarse para modelar sucesiones de datos donde los miembros de la sucesión dependen de su vecindario. Algunos ejemplos incluyen el del modelado del monto correspondiente al total de indemnizaciones pagadas por una compañía de seguros en un periodo determinado (e.g. Beard *et. al.*, 1990), y el de inventarios estocásticos donde la demanda de un producto en específico en un periodo determinado depende altamente de lo demandado en el periodo precedente (e.g. Porteous, 2002).

La mayoría de la literatura disponible suele partir de la suposición de que los elementos de la sucesión de variables aleatorias son independientes. Esta suposición puede dar una serie de inferencias erróneas cuando existe cierto grado de dependencia entre las variables aleatorias. De hecho, muchas veces un investigador desea llevar a cabo pruebas de hipótesis sobre esta independencia tratando de usar tanta información de la sucesión como sea posible. La meta de este estudio es la de proponer un procedimiento para modelar este tipo de secuencias en un marco completamente paramétrico.

2. Construcción de cadenas de Markov

Sean Y_i, $i = 0, 1, \ldots$, los valores de la cadena de Markov. Una familia de cadenas de Markov puede definirse por las probabilidades de transición (e.g Rolski *et. al.*, 2000)

$$P_i(Y_i < y \,|\, Y_{i-1} = z).$$

[1] escarela@yahoo.com

En este caso, la distribución inicial $P_0(y) = \Pr\{Y_0 < y\}$ completa la definición de una cadena en particular. Cuando se desea modelar cadenas de Markov estacionarias, es necesario construir probabilidades de transición las cuales no dependen del índice i. Una cadena de Markov puede entonces definirse usando una función de distribución bidimensional $F(x, y)$, la cual cuenta con una derivada parcial finita con respecto a y. Así, la probabilidad de transición se puede calcular con

$$P_i(Y_i < x|Y_{i-1} = z) = \frac{\partial F(x,z)}{\partial z} \Big/ \frac{\partial F(\infty,z)}{\partial z},$$

donde $F(\infty, z)$ denota la distribución marginal de la segunda variable.

En este estudio nos concentramos a una clase de modelos de cadenas de Markov generada por la distribución bivariada del tipo Morgenstern, la cual está definida por (e.g. d'Este, 1981)

$$F(x, z) = F(x)F(z)\{1 + a[1 - F(x)][1 - F(z)]\},$$

donde a es el parámetro que controla el nivel de dependencia entre las marginales, el cual satisface

$$|a| \leq 1,$$

y $F(x)$ denota la forma que toma cada distribución marginal; aquí suponemos que $F(x)$ es absolutamente continua.

Bajo el modelo de Morgenstern es posible demostrar que la probabilidad de transición es

$$\begin{aligned} P(x|z) &= P_i(Y_i < x|Y_{i-1} = z) \\ &= F(x)\{1 + a[1 - F(x)][1 - 2F(z)]\}. \end{aligned}$$

y la función de densidad correspondiente es

$$p(x|z) = f(x)\{1 + a[1 - 2F(x)][1 - 2F(z)]\}.$$

La distribución estacionaria de la Cadena de Markov es efectivamente $F(x)$, con la restricción de que $|a| < 1$. Esto se puede confirmar al notar que la distribución estacionaria $Q(x)$ debe cumplir

$$Q(x) = \int_{-\infty}^{\infty} P(x|z)\, Q(z)\, \mathrm{d}z,$$

pero en este caso solo $F(x)$ lo puede cumplir.

La esperanza condicional de Y_i se puede calcular en términos de su predecesor Y_{i-1}, el parámetro de dependencia a, y la distribución estacionaria $F(x)$ como se muestra a continuación:

$$\begin{aligned} \mathrm{E}[Y_i|Y_{i-1}=z] &= \int_{-\infty}^{\infty} x\, f(x)\,\{1+a[1-2F(x)][1-2F(z)]\}\,\mathrm{d}x \\ &= m_X + a\,[1-2F(z)]\,\alpha, \end{aligned}$$

donde m_X denota la media de la distribución estacionaria $F(x)$ y

$$\alpha = \int_{-\infty}^{\infty} x\,[1-2F(x)]\,f(x)\,\mathrm{d}x.$$

La varianza condicional correspondiente es

$$\begin{aligned} \mathrm{Var}[Y_i|Y_{i-1}=z] &= \sigma_X^2 + a[1-2F(z)]\int_{-\infty}^{\infty} x^2\,[1-2F(x)]\,f(x)\,\mathrm{d}x - \\ &\quad -\{a[1-2F(z)]\alpha\}^2 - 2\,m_X\,a\,[1-2F(z)]\,\alpha, \end{aligned}$$

donde σ_X^2 es la varianza de $F(x)$. Es posible demostrar también que la covarianza del n-ésimo orden es

$$\mathrm{Cov}^{(n)} = \mathrm{Cov}[Y_i, Y_{i+n}] = \frac{a^n}{3^{n-1}}\alpha^2,$$

y entonces el coeficiente de correlación correspondiente es

$$\mathrm{Corr}^{(n)} = \frac{a^n\alpha^2}{3^{n-1}\sigma_x^2}.$$

Las densidades de transición para el paso n son

$$\begin{aligned} p^{(n)}(x|z) &= \int_0^{\infty} p^{(n-1)}(x|y)\,p(y|z)\,\mathrm{d}y \\ &= f(x)\left\{1+\frac{a^n}{3^{n-1}}[1-2F(x)][1-2F(z)]\right\}. \end{aligned}$$

3. Inferencia

El problema de estimación de los parámetros consiste en dos partes: la primera es encontrar un procedimiento para estimar los parámetros de la distribución estacionaria $F(x)$, y la segunda consiste en estimar el parámetro de dependencia a. Como los estimadores de máxima verosimilitud son asintóticamente eficientes dadas ciertas condiciones de regularidad, es conveniente usar el método de máxima verosimilitud para estimar los parámetros en cuestión.

Supóngase que la cadena de Markov ha observado desde el valor inicial X_0 hasta el n-ésimo X_n. Entonces la función log-verosimilitud es

$$\log L = \sum_{i=1}^{n} \log p(X_i|X_{i-1}) + \log a_0(X_0).$$

donde el último término de la suma no depende del parámetro a y entonces es irrelevante para el proceso estimación. El problema no es lineal y por ello es necesario usar métodos numéricos. En la actualidad la mayoría de los paquetes estadísticos incluyen rutionas que optimizan funciones multivariadas. En el paquete S-PLUS, por ejemplo, la función `nlminb` del paquete estadístico S-PLUS puede encontrar el punto que minimiza $-\log L$ bajo restricciones como la de $|a| < 1$; cuando se trata de obtener el Hessiano, es posible usar el paquete matemático MAPLE y así aproximar la matriz de covarianzas de los estimadores.

Para encontrar una prueba sobre la independencia de la sucesión de las variables aleatorias, i.e. cuando se desea encontrar una prueba para $H_0 : a = 0$ contra $H_1 : a \neq 0$, es posible basarse en la aproximación $(\hat{a} - a)/\hat{\text{se}}_{\hat{a}} \sim \text{N}(0, 1)$, donde $\hat{a}$ representa el estimador de máxima verosimilitud de a y $\hat{a}$ es el estimador del error estándar de $\hat{a}$. Este procedimiento puede compararse con la prueba basada en el estadístico de prueba de correlación definido por

$$D = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - 1/2)(Y_{i-1} - 1/2)$$

Evidentemente la "prueba Morgenstern" es un poco más potente que la prueba de correlación pues hace mayor uso de la información disponoble, la cual incluye la forma de $F(x)$.

4. Discusión

El método propuesto forma un esquema completamente parámetrico bastante conveniente y conciso para obtener inferencias sobre una sucesión de variables aleatorias. Desde luego, otras funciones bivariadas pueden sugerirse para defenir nuevas familias de cadenas de Markov. Una forma conveniente de crear estas familias es a través del uso de *cópulas* (ver Joe, 1997, capítulo 8), las cuales son clases de distribuciones bivariadas especificadas en términos de distribuciones marginales y una función cópula $C_a(\cdot,\cdot)$, la cual es una función de distribución bivariada en el cuadro unitario $[0,1]^2$. De hecho, el modelo de Morgenstern presentado en este estudio está determinado por la cópula (Joe, 1997, p.35)

$$C_a(u,v) = uv[1 + a(1-u)(1-v)], \quad |a| \leq 1.$$

La elección de un cópula adecuada es aún un problema sin resolver. Un aspecto que debe jugar un papel preponderante para esta selección es la estructura de dependencia. La estructura presentada en este estudio, por ejemplo, no cubre un rango de dependencia muy amplio. Esto se puede ilustarar cuando se seleccciona una distribución Weibull para $F(x)$:

$$F(x) = 1 - \exp\left\{-\left(\frac{x}{\lambda}\right)^{\beta}\right\}, \qquad \lambda, \beta > 0.$$

En este caso, después de calcular varios valores de β, que juega el papel primordial en σ_X^2, se puede determinar que que $\text{Corr}^{(1)} \leq 0{,}33$, que es un coeficiente bastante limitado. Por lo tanto, otras familias de cópulas deben ser consideradas.

Referencias

Beard, R.E.; Pentikäinen, T. y Pesonen, E. (1990). *Risk Theory: The Stochastic Basis of Insurance*. Nueva York: Chapman & Hall.

d'Este G.M. (1981). A Morgenstern type bivariate Gamma-Distribution. *Biometrika*, **68**, 339-340.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Boca Raton: Chapman & Hall.

Porteous, E.L. (2002). *Foundations of Stochastic Inventory Theory*. California: Stanford University Press.

Rolski, T.; Schmidli, H.; Schmidt, V. y Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*. Chichester: Wiley.

Muestreo por seguimiento de nominaciones: Intervalos de confianza bootstrap del tamaño de una población de difícil detección[1]

Martín H. Félix Medina[2]
Escuela de Ciencias Físico-Matemáticas de la Universidad Autónoma de Sinaloa
Pedro E. Monjardin[3]
Escuela de Ciencias Físico-Matemáticas de la Universidad Autónoma de Sinaloa

1. Introducción

El Muestreo por seguimiento de nominaciones (denominado en Inglés como Link-tracing sampling o Snowball sampling) es un método que se ha propuesto para muestrear poblaciones de difícil detección, tales como poblaciones de drogadictos, niños de la calle, trabajadoras sexuales, etc. En este método se selecciona una muestra inicial de miembros de la población de interés, y se les pide a las personas seleccionadas que nominen a otros miembros de la población objetivo. A las personas que fueron nominadas se les puede pedir que nominen a otras personas, y el proceso de nominación puede continuar de esta manera hasta que se satisfaga alguna regla de terminación del muestreo.

Félix Medina y Thompson (2003) desarrollaron una variante de este tipo de muestreo en la cual, el difícilmente justificable supuesto de una muestra inicial Bernoulli se substituye por una muestra aleatoria simple de sitios, la cual se selecciona de un marco muestral que cubre sólo una parte de la población de interés. Posteriormente, Félix Medina y Monjardin (2003) consideraron esta variante y propusieron estimadores del tamaño poblacional derivados bajo el enfoque Bayesiano. Mediante un estudio de simulación, estos autores mostraron que sus estimadores se desempeñaban mejor que los estimadores máximo verosímiles propuestos por

[1]Trabajo realizado con apoyos parciales de CGIP-UAS y PROMEP UASIN-EXB-01-01
[2]mhfelix@uas.uasnet.mx
[3]pemo@uas.uasnet.mx

Félix Medina y Thompson (2003). Cabe señalar que Félix Medina y Monjardin (2003) usaron el enfoque Bayesiano sólo para construir estimadores del tamaño poblacional, pero usaron el enfoque frecuentista para realizar inferencias. Así, mediante el método Delta (aproximaciones lineales de Taylor), obtuvieron estimadores de varianza basados en diseño y propusieron intervalos de confianza que construyeron bajo el supuesto de normalidad de los estimadores propuestos.

En este trabajo se utiliza la metodología Bootstrap para obtener intervalos de confianza del tamaño poblacional a partir de los estimadores propuestos por Félix Medina y Monjardin (2003). Por tanto, estos intervalos no requieren del supuesto de normalidad, y además, por el procedimiento que se usa para construirlos, son basados en diseño y, consecuentemente, robustos a la especificación errónea de los modelos supuestos.

2. Diseño muestral y notación

El diseño muestral que se considera en este trabajo es el propuesto por Félix Medina y Thompson (2003). Así, se supone que una parte U_1 de la población de interés U es cubierta por un marco muestral de N sitios $A_1, \ldots, A_N$, tales como parques, hospitales o cruceros de calles. De este marco se selecciona una muestra aleatoria simple sin reemplazo $S_0 = \{A_1, \ldots, A_n\}$ de n sitios, y a las personas de la población de interés que pertenecen a los sitios en S_0 se les pide que nominen a otros miembros de la población. Como convención, se dice que una persona es nominada por el sitio A_i, si cualquiera de las personas de ese sitio la nominan.

Se denotará por τ el tamaño de U, por τ_1 el de U_1, por $\tau_2 = \tau - \tau_1$ el de $U_2 = U - U_1$, y por m_i el número de personas en A_i. Finalmente, se usarán los conjuntos de variables $\{x_{ij}^{(1)}\}$ y $\{x_{ij}^{(2)}\}$ para indicar el proceso de nominación. Así, $x_{ij}^{(1)} = 1$ si la persona $u_j \in U_1 - A_i$ es nominada por el sitio A_i, y $x_{ij}^{(1)} = 0$ en otro caso. Similarmente, $x_{ij}^{(2)} = 1$ si la persona $u_j \in U_2$ es nominada por el sitio A_i, y $x_{ij}^{(2)} = 0$ en otro caso.

3. Estimadores del tamaño poblacional basados en las modas de las distribuciones posteriores

En este trabajo se consideran los estimadores propuestos por Félix Medina y Monjardin (2003). Así, se supone que los m_i's son realizaciones de variables aleatorias independientes Poisson, y que las $x_{ij}^{(k)}$'s son realizaciones de variables aleatorias independientes Bernoulli con medias $p_i^{(k)}$'s, $k = 1, 2$, e $i = 1, \ldots, n$. Con respecto a las distribuciones iniciales de los parámetros, se consideran tres casos para las distribuciones de τ_1 y τ_2. En un caso se supone que los τ's tienen distribuciones Poisson, en otro que tienen distribuciones uniformes no informativas, y en el otro que tienen distribuciones de Jeffreys. Las probabilidades de nominación $p_i^{(k)}$'s se transforman mediante la transformación logit $\alpha_i^{(k)} = \ln[p_i^{(k)}/(1 - p_i^{(k)})]$, y se supone que los $\alpha_i^{(k)}$'s tienen distribuciones normales. A partir de estas distribuciones, se obtienen las distribuciones posteriores conjuntas de τ_1, τ_2, $\alpha_i^{(1)}$ y $\alpha_i^{(2)}$, $i = 1, \ldots, n$, y se usan como estimadores de estos parámetros las modas de las distribuciones posteriores.

Las inferencias acerca de los τ's se realizan bajo el enfoque frecuentista. En particular, intervalos de confianza de los τ's se construyen bajo el supuesto de normalidad de los estimadores y usando la forma: $\left(\hat{\tau} - z_{1-\alpha/2}\sqrt{\hat{\mathbf{V}}(\hat{\tau})}, \hat{\tau} + z_{1-\alpha/2}\sqrt{\hat{\mathbf{V}}(\hat{\tau})}\right)$, donde $\hat{\tau}$ es un estimador de τ, $\hat{\mathbf{V}}(\hat{\tau})$ es un estimador, basado en diseño, de la varianza de $\hat{\tau}$, y $z_{1-\alpha/2}$ es el $1 - \alpha/2$ cuantil de la distribución normal estándar.

4. Intervalos de confianza Bootstrap

En este trabajo, los intervalos de confianza Bootstrap de los τ's se obtuvieron mediante una combinación de las variantes del Bootstrap conocidas como paramétrica y no paramétrica (para una descripción de estas variantes ver Efron y Tibshirani, 1993). Los intervalos de confianza se obtuvieron como sigue. (i) De los tamaños m_i, $i = 1, \ldots, n$, de los n sitios en la muestra inicial S_0 se selecciona una muestra aleatoria simple con reemplazo de tamaño n. Sean $i = i_1, \ldots, i_n$ los subíndices de los m_i que son seleccionados. (ii) Para cada $i = i_1, \ldots, i_n$,

se generan muestras de tamaños $\hat{\tau}_1 - m_i$ y $\hat{\tau}_2$ de distribuciones Bernoulli con medias $\hat{p}_i^{(1)}$ y $\hat{p}_i^{(2)}$, respectivamente, donde $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{p}_i^{(1)}$ y $\hat{p}_i^{(2)}$ son estimaciones de τ_1, τ_2, $p_i^{(1)}$ y $p_i^{(2)}$ respectivamente que se calculan a partir de la muestra originalmente observada. Estas muestras simulan los valores de los conjuntos de variables $\{X_{ij}^{(1)}\}$ y $\{X_{ij}^{(2)}\}$. (iii) Las estimaciones para τ_1, τ_2 y $\tau = \tau_1 + \tau_2$ se calculan a partir de las muestras generadas en los pasos (i) y (ii), y usando el mismo procedimiento que el que se usa para calcular las estimaciones $\hat{\tau}_1$ y $\hat{\tau}_2$. (iv) Las distribuciones Bootstrap de $\hat{\tau}_1$, $\hat{\tau}_2$ y $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$ se obtienen repitiendo los pasos (i)-(iii) un número grande de veces. (v) Intervalos de confianza Bootstrap del $100(1-\alpha)\,\%$ de τ_1, τ_2 y τ se obtienen mediante el método percentil, esto es, los extremos inferior y superior de los intervalos están dados por los cuantiles $\alpha/2$ y $1-\alpha/2$ de las distribuciones Bootstrap de $\hat{\tau}_1$, $\hat{\tau}_2$ y $\hat{\tau}$.

5. Estudio Monte Carlo

Para realizar este estudio se generó una población finita de $N = 250$ valores m_i a partir de la distribución Poisson con media 7.2. Con los valores generados se obtuvo $\tau_1 = \sum_1^N m_i = 1897$. El valor de τ_2 se fijó en 700, por lo que $\tau = 2597$. Las probabilidades de nominación se generaron mediante el modelo $p_i^{(k)} = 1 - \exp(-\beta_k m_i)$, donde los valores de β_k fueron tales que se tuvieron dos casos. Caso 1: $\left(\mathbf{E}(p_i^{(1)}), \mathbf{E}(p_i^{(2)})\right) = (0{,}05, 0{,}03)$ y Caso 2: $\left(\mathbf{E}(p_i^{(1)}), \mathbf{E}(p_i^{(2)})\right) = (0{,}01, 0{,}006)$. Los valores de los parámetros de las distribuciones iniciales se fijaron como en Félix-Medina y Monjardin (2003).

En la Tabla 1 se presentan los resultados del estudio. Se puede observar que cuando las probabilidades de nominación son grandes (Caso 1), los desempeños de los intervalos Bootstrap y los desempeños de los intervalos normales son bastante aceptables y similares entre sí. Sin embargo, los intervalos normales tienen desempeños ligeramente superiores a los de los Bootstrap. Finalmente, no existe diferencia significativa entre los desempeños de los intervalos que se construyeron a partir de los distintos estimadores de los tamaños poblacionales. Por otro lado, cuando las probabilidades de nominación son pequeñas (Caso 2), los desempeños de ambos tipos de intervalos no son tan buenos como los del caso anterior. Específicamente, aunque los porcentajes de cobertura sólo disminuyen ligeramente, las longitudes sí se incre-

Tabla 1: Intervalos de Confianza al nivel 95 %

	$\left(\mathbf{E}(p_i^{(1)}), \mathbf{E}(p_i^{(2)})\right) = (0.05,0.03)$				$\left(\mathbf{E}(p_i^{(1)}), \mathbf{E}(p_i^{(2)})\right) = (0.01,0.006)$			
	Bootstrap		Normal		Bootstrap		Normal	
	Cob	Long	Cob	Long	Cob	Long	Cob	Long
$\hat{\tau}_1^{(P)}$	0.933	124.8	0.958	123.0	0.861	386.3	0.906	378.4
$\hat{\tau}_2^{(P)}$	0.958	162.3	0.947	159.2	0.999	480.9	0.939	501.1
$\hat{\tau}^{(P)}$	0.944	204.8	0.945	201.5	0.964	623.1	0.927	629.2
$\hat{\tau}_1^{(U)}$	0.951	125.7	0.962	124.2	0.916	400.6	0.945	393.4
$\hat{\tau}_2^{(U)}$	0.952	167.4	0.946	163.4	0.955	1828.6	0.923	1164.9
$\hat{\tau}^{(U)}$	0.947	209.5	0.950	205.5	0.952	1914.3	0.948	1251.2
$\hat{\tau}_1^{(J)}$	0.934	125.7	0.939	124.2	0.925	399.6	0.943	392.2
$\hat{\tau}_2^{(J)}$	0.958	165.5	0.950	162.3	0.925	880.1	0.876	908.5
$\hat{\tau}^{(J)}$	0.944	208.4	0.951	204.8	0.929	1002.3	0.909	1006.9

Resultados basados en 2000 réplicas. Cob=Cobertura y Long=Longitud. Estimadores Bootstrap basados en 1000 muestras Bootstrap. $\hat{\tau}_k^{(P)}$, estimadores basados en distribución inicial Poisson. $\hat{\tau}_k^{(U)}$, estimadores basados en distribución inicial no informativa uniforme. $\hat{\tau}_k^{(J)}$, estimadores basados en distribución inicial de Jeffreys.

mentan significativamente. No es muy claro cual de los dos tipos de intervalos tiene mejor desempeño. Sin embargo, es claro que los intervalos que se construyen a partir de los estimadores $\hat{\tau}_1^{(U)}$, $\hat{\tau}_2^{(U)}$ y $\hat{\tau}^{(U)}$ tienen malos desempeños, y que los intervalos que se construyen a partir de los estimadores $\hat{\tau}_1^{(P)}$, $\hat{\tau}_2^{(P)}$ y $\hat{\tau}^{(P)}$ son los de mejores desempeños.

6. Conclusiones

En este trabajo se usa la metodología Bootstrap para construir intervalos de confianza del tamaño poblacional a partir de los estimadores propuestos por Félix-Medina y Monjardin (2003). A partir de los resultados del limitado estudio de simulación que se realizó en esta investigación, parece ser que los intervalos basados en el supuesto de normalidad tienen

desempeños ligeramente superiores que los de los intervalos Bootstrap. Sin embargo, se requieren estudios de simulación más extensos que el presente para determinar con mayor certeza cual de los dos tipos de intervalos tiene mejor desempeño.

Referencias

Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Félix-Medina, M.H., and Monjardin, P.E. (2003). Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population: a Bayesian approach. A publicarse en *Proc. of the Section on Survey Research Methods of the American Stat. Assoc.*

Félix-Medina, M.H., and Thompson, S.K. (2003). Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. A publicarse en *Journal of Official Statistics*.

Rutinas y formularios en MS-Access para hacer muestreo con reemplazo y sin reemplazo de un grupo de datos y un ejemplo de aplicación

Alejandro Liedo Galindo[1]
Instituto Nacional de la Pesca

1. Introducción

La idea de hacer rutinas o procedimientos y formularios en MS-Access para hacer muestreo con o sin reemplazo, surgió después de la lectura de las notas de un proyecto de remuestreo en el que usaban el programa STATS que tiene su propio lenguaje, estas notas abarcan el contenido de lo que podría ser un primer curso de estadística a nivel universitario. Utilizan el remuestreo para resolver todo tipo de problemas mediante un enfoque experimental frecuentista. La idea era hacer una aplicación que permitiera plantear y resolver problemas de ese tipo, pero que quedara abierto, es decir, que se pudiera revisar incluso muestra por muestra con el MS-Access u otra aplicación. En las tres secciones siguientes explico los procedimientos para hacer el muestreo, los formularios que sirven como interfase para el usuario y finalmente el ejemplo.

2. Las rutinas

El procedimiento `GeneraMuestrasCR(n As Long, tam As Integer)` genera muestras sin reemplazo y tiene dos parámetros, que pasan como argumentos al llamar al procedimiento, `n` es el número de muestras y `tam` el tamaño de cada muestra. El muestreo con reemplazo es más fácil, simplemente se toma elemento por elemento de una tabla y lo escribe en otra hasta completar la muestra y lo repite hasta completar el número de muestras. Para hacerlo se utilizan algunos objetos de acceso a datos (DAO) los cuales se declaran al principio del

[1] `aliedo@prodigy.net.mx`

procedimiento, los más importantes son los objetos *recordset* que van a servir para obtener las muestras `MiRSDatos`, y para escribirlas `MiRsMuestras` a estos conjuntos de registros se les asignan los de las tablas "datos1" y muestras respectivamente mediante el siguiente código:

```
Set MiRSDatos  = dbs.OpenRecordset('datos1')
Set MiRsMuestras = dbs.OpenRecordset('muestras')
```

Se asigna a la variable `ndatos` el número de registros en `MiRSDatos` y se verifica que al menos haya un registro para proceder. Se activa el índice principal para el conjunto de registros `MiRSDatos`, en este caso el índice existe previamente en la tabla "datos1" y es el campo elemento el que se usa para construir el índice. Con el método *MoveFirst* se ubica el primer registro de la colección y se asigna el valor del campo elemento a la variable auxiliar `nAux`. Mediante ciclos `For` anidados se hace para $i = 1$ hasta `n` (número de muestras) y desde $j = 1$ hasta `tam`, elegir mediante la expresión `nsel = Int(Rnd() * ndatos) + nAux` un número entre `nAux` y `nAux + ndatos` que sirve para localizar el registro seleccionado con la instrucción `MiRSDatos.Seek '=', nsel`, para copiar los valores de los campos de `MiRSDatos` a `MiRSMuestras` hasta completar el tamaño de muestra y se repite hasta completar en número de muestras.

Este algoritmo está suponiendo que el campo elemento cumple con que son valores consecutivos, y el valor menor no necesariamente es uno. Si los valores no fueran consecutivos, habría valores que producirían error cuando `nsel` no tenga un valor correspondiente en el campo elemento.

El procedimiento `GeneraMuestrasSR` tiene por objetivo generar muestras sin reemplazo, lo que hace este procedimiento es elegir un registro de la tabla "datos1" y verifica que no esté en la muestra (en la tabla "muestrasaux"), para esto se utiliza una variable lógica llamada `ya`, que al iniciar el procedimiento está en *False*, cuando se encuentra un valor que no está en la tabla auxiliar, esto es cuando el método `MirsMuestrasAux.NoMatch` regresa un *true*, se procede entonces a darle valor *true* a la variable `ya`.

```
While Not ya
nsel = Int(Rnd() * ndatos) + nAux
MiRSDatos.Seek '=', nsel
MirsMuestrasAux.Seek '=', nsel
If MirsMuestrasAux.NoMatch Then
ya = True
End If
Wend
```

El registro o mejor dicho los campos del registro seleccionado se copian a un registro nuevo en el conjunto de registros `MirsMuestrasAux` mediante el siguente código.

```
With MirsMuestrasAux
.AddNew
.Fields('valor') = MiRSDatos('valor')
.Fields('valor1') = MiRSDatos('valor1')
.Fields('muestra') = I
.Fields('elemento') = MiRSDatos('elemento')
.Update
End With With
```

Después de completar la muestra, esto es cuando el conjunto de registros `MirsMuestrasAux` tiene igual número de registros que tamaño de muestra, se procede a anexar todos los registros del conjunto `MirsMuestrasAux` al conjunto `MiRSMuestras` , para esto se construye un código SQL que se almacena en la variable SQL y se crea una nueva consulta mediante la instrucción `Set consultaAux = dbs.CreateQueryDef('', SQL)` y se ejecuta mediante la instrucción `consultaAux.Execute`.

Código SQL que se almacena en la variable SQL:

```
INSERT INTO Muestras ( muestra, valor, valor1, elemento )
SELECT MuestrasAux.muestra, MuestrasAux.valor, MuestrasAux.valor1,
MuestrasAux.elemento
FROM MuestrasAux;
```

De forma similar, para borrar el contenido de la tabla "muestrasaux" se utiliza el código:

```
DELETE MuestrasAux.muestra, MuestrasAux.valor, MuestrasAux.valor1,
MuestrasAux.elemento
FROM MuestrasAux;
```

Y esto se repite hasta completart el número de muestras.

3. Los formularios

El formulario **añadir datos** tiene dos cuadros de texto, uno para escribir el valor que se desea añadir y otro para el número de veces que se desea añadir. El botón **añadir** llama a un procedimiento que añade el valor a la tabla de datos tantas veces como se haya especificado en el cuadro de texto Número de veces.

Figura 1: Formulario **añadir datos**

El botón **añadir 1000** llama a un procedimiento que llama a la función `rnd()` que regresa un número aleatorio de una distribución uniforme entre cero y uno. Y lo escribe en la tabla "datos1", esto lo repite mil veces.

El Botón **Borrar datos actuales**, llama a un procedimiento que borra todos los registros de la tabla "datos1".

El formulario **muestreo** tiene tres cuadros de texto y tres botones, los primeros dos cuadros de texto se usan para indicar el número de muestras y el tamaño de las mismas respectivamente y el tercero es para mostrar el grado de avance en el proceso del muestreo ya que dependiendo de la computadora el proceso de obtener muestras puede llegar a tardarse y en ese campo se puede ver que muestra se está obteniendo. El botón **Generar Muestras con reemplazo** llama a la rutina que genera las muestras con reemplazo `GeneraMuestrasCR(n As Long, tam As Integer)`, el cuál tiene dos argumentos el primero para el número de muestras y el segundo para el tamaño de las mismas. Mientras que el botón **GeneraMuestrasSR** llama a la rutina `GeneraMuestrasSR(n As Long, tam As Integer)`, con los mismos argumentos para el número de muestras y el tamaño.

Figura 2: Formulario **muestreo**

4. El ejemplo

Se va a usar el muestreo sin reemplazo para estimar la probabilidad de diferentes resultados al seleccionar 5 cartas de un mazo de 52. El problema se puede dividir en tres partes: la primera es la obtención de las muestras, la segunda contar los diferentes resultados y la tercera es comparar los resultados obtenidos con los esperados mediante una prueba de bondad de ajuste χ^2.

Se simula un mazo de 52 cartas mediante una tabla de 52 registros 4 registros con cada número del 1 al 13, esto hace con el formulario **añadir datos** poniendo cada número en el cuadro de texto **Valor a añadir** y el cuatro en el cuadro de texto **número de veces** y haciendo clic en el botón **Añadir**. Se toma un número "grande" de muestras (30,000) de tamaño 5 sin reemplazo, es decir, no puede aparecer la misma carta en una muestra más de una vez. Esto se hace mediante el formulario **muestras**. Una vez que hemos obtenido las muestras se debe contar cuantas muestras dieron como resultado 5 cartas con números diferentes, en cuantas 1 par y tres diferentes, dos pares, par y tercia y finalmente 4 iguales, para contestar esto hay que revisar las muestras y contar. Para lograrlo se hicieron las consultas que a continuación se describen.

1. La primera consulta se basa en la tabla muestras, agrupa por muestra y valor, y cuenta. El resultado de esta consulta consta de tres campos que son la muestra, el número y cuántas veces apareció en la muestra.

   El código SQL:

   ```
   SELECT Muestras.muestra, Muestras.valor,
   Count(Muestras.valor) AS CuentaDevalor
   FROM Muestras
   GROUP BY Muestras.muestra, Muestras.valor;
   ```

muestra	valor	Cuenta de valor
1	6	2
1	8	1
1	11	1
1	12	1

2. La segunda consulta se basa en la primera, agrupa por muestra y cuenta de valor valor y cuenta. Esta consulta nos dice para cada muestra cuántas cartas aparecieron 1 vez, 2 y hasta cuatro veces.

   El código SQL:

```
SELECT [f18-primera].muestra, [f18-primera].CuentaDevalor,
Count([f18-FROM [f18-primera]
GROUP BY [f18-primera].muestra, [f18-primera].CuentaDevalor;
```

muestra	Cuenta de valor	Cuenta de cuenta de valor
1	1	3
1	2	1

3. La tercera consulta se basa en la segunda, es de referencias cruzadas agrupa por muestra (filas) y cuenta de valor (columnas) y cuenta de cuenta (valor). Para cada muestra nos dice cuantas veces aparecieron 1, 2, 3 o 4 cartas en la muestra.

   El código SQL:

```
TRANSFORM Sum([f18-segunda].CuentaDeCuentaDevalor)
AS SumaDeCuentaDeCuentaDevalor
SELECT [f18-segunda].muestra
FROM [f18-segunda]
GROUP BY [f18-segunda].muestra
PIVOT [f18-segunda].CuentaDevalor;
```

muestra	1	2	3	4
1	3	1	0	0
2	5	0	0	0
3	5	0	0	0
4	5	0	0	0
5	5	0	0	0

4. La cuarta consulta se basa en la tercera, y cuenta el número de veces que ocurrió cada evento en el total de las muestras. Una corrida de 30,000 muestras dio como resultado los valores en la columna observados y la columna esperados muestra los valores del producto de la probabilidad de cada evento por el 30,000 (las probabilidades se calcularon mediante el conteo de puntos muestrales):

El código SQL:

```
SELECT [f18-tercera].[1], [f18-tercera].[2], [f18-tercera].[3],
[f18-tercera].[4], Count([f18-tercera].muestra)
AS CuentaDemuestra
FROM [f18-tercera]
GROUP BY [f18-tercera].[1], [f18-tercera].[2],
[f18-tercera].[3],[f18-tercera].[4];
```

1	2	3	4	Observados	Esperados
5	0	0	0	15175	15212
3	1	0	0	12744	12677
1	2	0	0	1397	1426
2	0	1	0	624	634
0	1	1	0	51	47
1	0	0	1	9	7

Obteniendo χ^2 =2.1034 y un nivel de significancia observado de 0.8347. Por que las probabilidades obtenidas por los dos métodos no se pueden considerar diferentes.

Un archivo de MS-Access con las rutinas, formularios y las consultas descritas en este trabajo le será enviado a quien lo solicite por correo electrónico.

Referencias

The Resampling Project (1994). College of Business and Management, University of Maryland.

Battacharyya G.K., Johnson R.A. (1977). *Statistical Concepts and Methods*. New York: Wiley.

Gifford,D. et al. (1996). *Access 95 Unleashed*. Indianapolis: Sams Publishing.

Estudio estadístico de dos variables críticas en la calidad del agave: peso y contenido de carbohidratos

Dagmar Mariaca Hajducek [1]

Centro de Investigación en Matemáticas, A.C

1. Introducción

Debido a que el agave tequilero forma parte de un proceso industrial, la variación en su crecimiento y desarrollo no sólo se encuentra inducida por la naturaleza, sino también por diversos factores inherentes a dicho proceso. Aunque en ocasiones no sea factible controlar algunos de los factores naturales o del proceso, conocer su efecto puede ser crucial para mejorar la producción. Tener una población homogénea de agave, debidamente estratificada, facilitará las labores de planeación y de manejo implicando la reducción en los costos de producción. Así, la motivación principal de este trabajo tiene que ver con reducir la variación, mediante un estudio estadístico de los factores que afectan a la calidad del agave.

Es importante como primer paso, identificar las variables críticas en la calidad del agave, es decir, aquellas que mejor la definen. Por tradición se ha pensado que el peso se encuentra directamente relacionado con el contenido de carbohidratos (%RT-reductores totales), por lo que ha sido la única variable utilizada en la evaluación de la calidad. Sin embargo, sabemos que el tequila se obtiene precisamente a partir de los azúcares fermentables del agave (var. Tequilana Weber Azul). Con la intención de redefinir el concepto de rendimiento del agave no sólo en términos del peso, sino también del %RT, se estudió la posible relación lineal entre estas variables y, como se ha mencionado, su comportamiento por separado bajo el posible impacto de los diversos factores de variación. Como parte del estudio de la variabilidad, se comparan distintos modelos de crecimiento para la predicción del peso, se propone un método para la estimación del %RT como materia prima del proceso de producción del tequila y se estudian los métodos de laboratorio disponibles para la medición de %RT.

[1] Director de tesis: Fernando Avila Murillo

2. Desarrollo

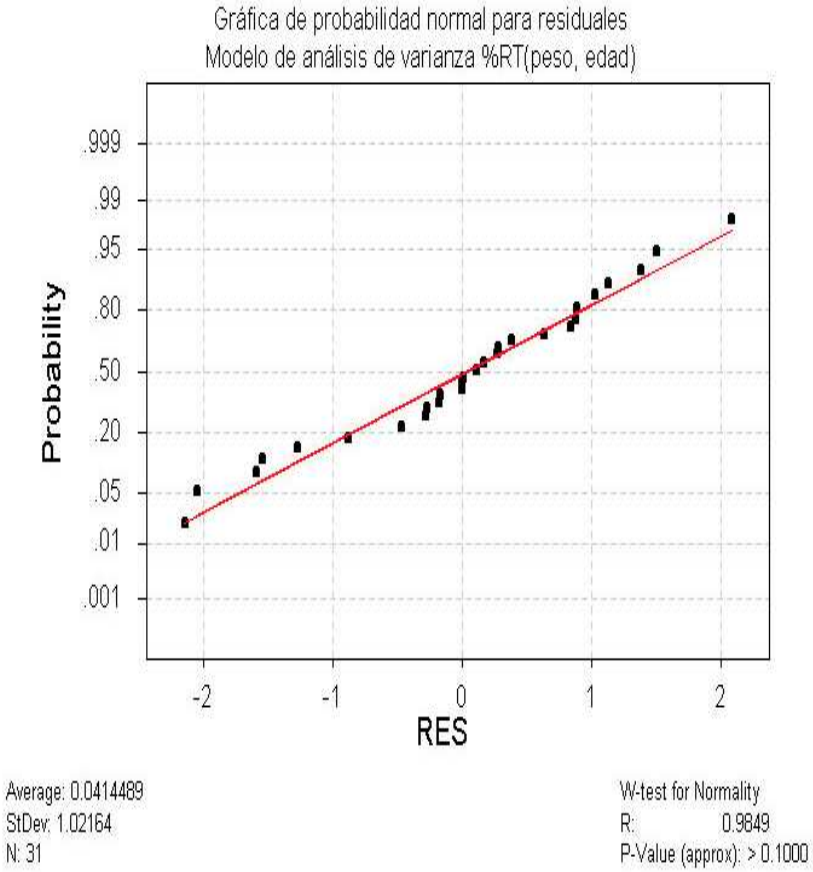
El universo de agave en cuestión comprende las zonas del Centro y de los Altos de Jalisco, así como parte de Nayarit y Guanajuato. Este estudio fue realizado en el año 2002, por lo que las edades mencionadas (3 a 9 años) corresponden a los años de plantación de 1993 a 1999. A continuación se presenta un resumen del estudio, enunciando las diversas causas de variación junto con algunas de las herramientas y los tipos de análisis que se realizados, así como la relevancia práctica de los resultados, obtenidos en algunos casos con base en información histórica y en otros mediante planes de muestreo. Se han omitido los valores de peso y %RT con el fin de preservar la confidencialidad de la información. Los factores de variación estudiados son los siguientes:

2.1. Edad. De la relación entre el peso, la edad y el %RT

A partir de un modelo de análisis de varianza y del archivo histórico de mediciones de %RT de enero de 1999 a agosto de 2002, se concluye que no se encontró evidencia de un efecto significativo del peso ni de la edad en el %RT.

Source	**DF**	**Seq SS**	**Adj SS**	**Adj MS**	**F**	**p**
Peso	19	2.2562	1.8127	0.0954	0.61	0.846
Edad	8	1.0349	1.0349	0.1294	0.82	0.596
Error	14	2.2021	2.2021	0.1573		
Total	41	5.4932				

Debido a la no linealidad de la relación entre el %RT y el peso, no se recomienda utilizar solamente al peso para evaluar la calidad del agave. Este resultado muestra la factibilidad de plantear la hipótesis de no linealidad y diseñar un estudio a futuro, que permita estratificar la información de acuerdo a causas de variación para determinar un modelo que explique esta relación y profundizar el análisis. Tampoco se ha encontrado un efecto significativo de la edad en el %RT. La relevancia de este resultado radica en que tradicionalmente se ha planeado la cosecha del agave en su edad madura, siendo que, determinando una edad óptima de jima, se podría ahorrar considerablemente en sus costos de producción.



2.1.1. Modelos de crecimiento para el peso

El objetivo es predecir el peso en función de la edad. La estimación de la asíntota de los modelos (representada por $\widehat{a}$), permite conocer la edad en la que el agave alcanza su peso máximo y de esta manera reducir los costos por mantenimiento, conocidos en general como costos por inventario. Se compararon distintos modelos de crecimiento, y no se encontró una diferencia significativa entre ellos. Se comprobó que la transformación logarítmica en la variable de respuesta mejora el comportamiento de los residuales y reduce las diferencias entre los modelos ajustados. Sea $\widehat{W}_k$ el peso ajustado en función de la variable $edad_k$ para la k-ésima observación:

Modelo Gompertz:

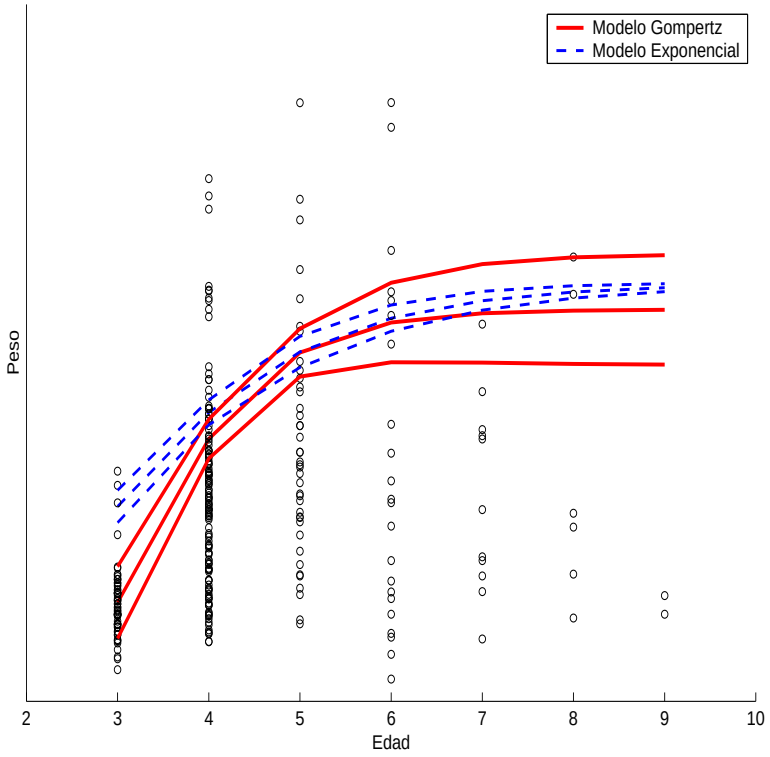
$$\widehat{W_k} = \widehat{a}\exp(-\exp(-\widehat{c}(edad_k - \widehat{b})))$$

Modelo Exponencial:

$$\widehat{W_k} = \widehat{a}/(1+\exp(\widehat{b} - edad_k))/\widehat{c}$$

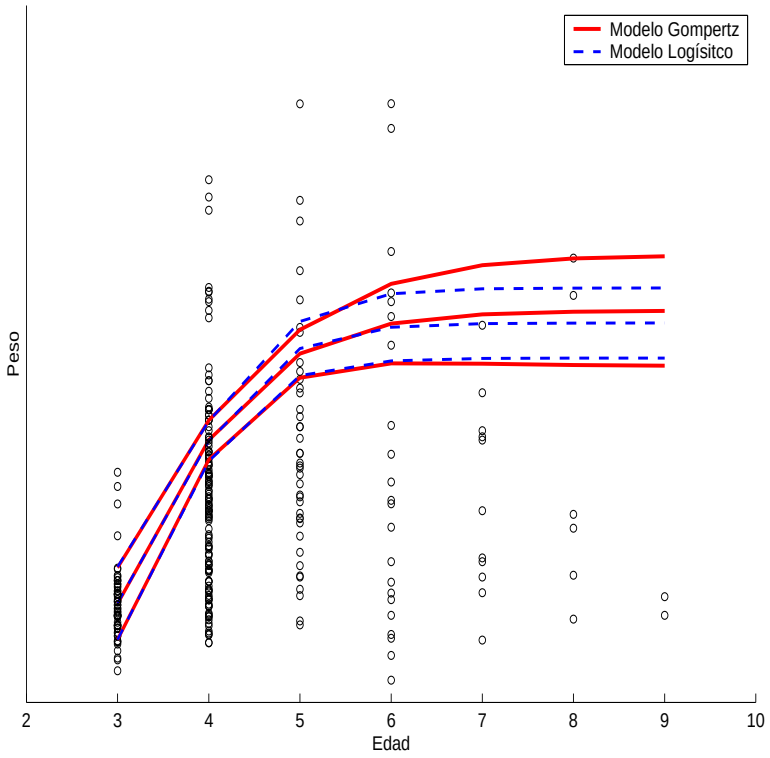
Modelo Logístico:

$$\widehat{W_k} = \widehat{a}(1-\exp(-\widehat{c}(edad_k - \widehat{b}))$$



2.2. Ambiente (clima y suelo). Efecto de estacionalidad en el %RT.

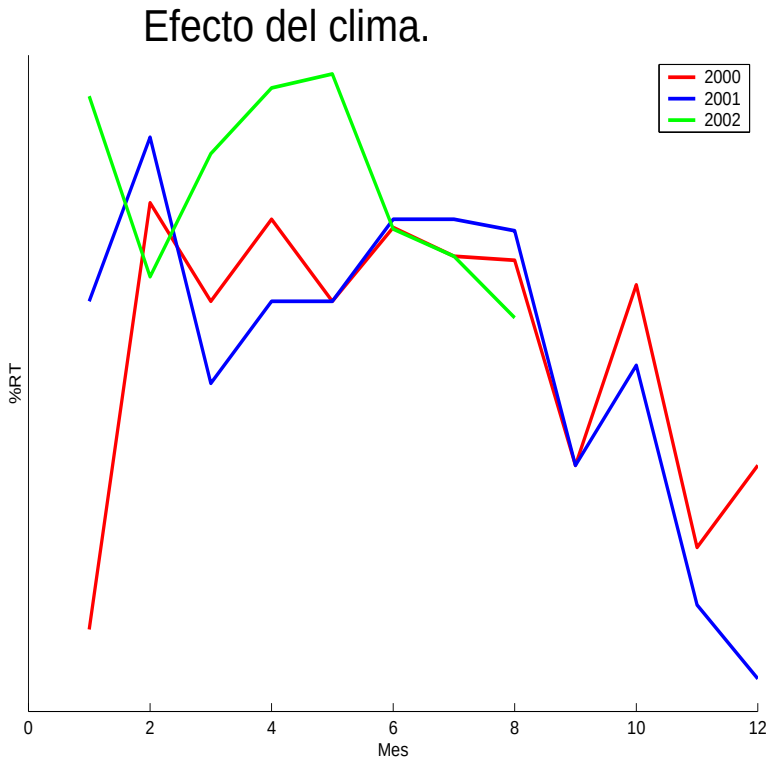
Es sabido que después de la temporada de lluvias el agave aumenta su peso debido al agua. Por otro lado, en la gráfica siguiente se advierte un patrón de estacionalidad en la reducción del %RT después de la temporada de lluvias. Esto implica que el rendimiento del agave en



cuanto a %RT disminuye a partir de una fecha específica y es posible planear la temporada de jima con mayor eficiencia.Por otro lado, esto refuerza la redefinición de la calidad en %RT y peso en conjunto.

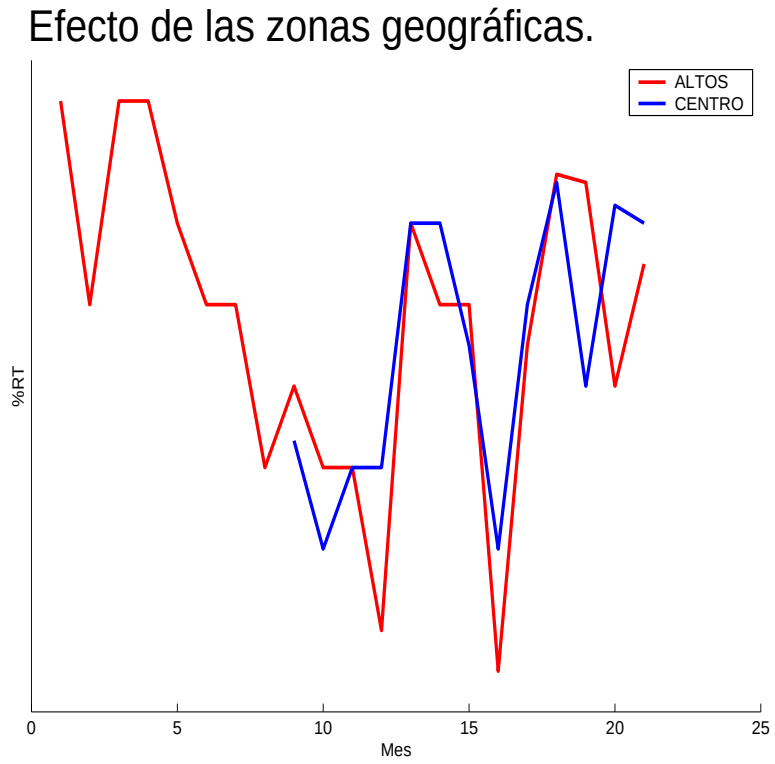
2.3. Espacio (zonas geográficas). Efecto de las zonas de plantación

El universo del agave estudiado comprende predios ubicados primordialmente en las zonas Centro y de los Altos de Jalisco. La gráfica anterior no presenta evidencia de que exista una diferencia importante entre las zonas de plantación. Esta gráfica fue realizada con información histórica, y se propuso como trabajo futuro la elaboración de un plan de investigación que permita detectar fuentes de variación controlables y obtener conclusiones más precisas.



2.4. Medición y procedimiento analítico de %RT

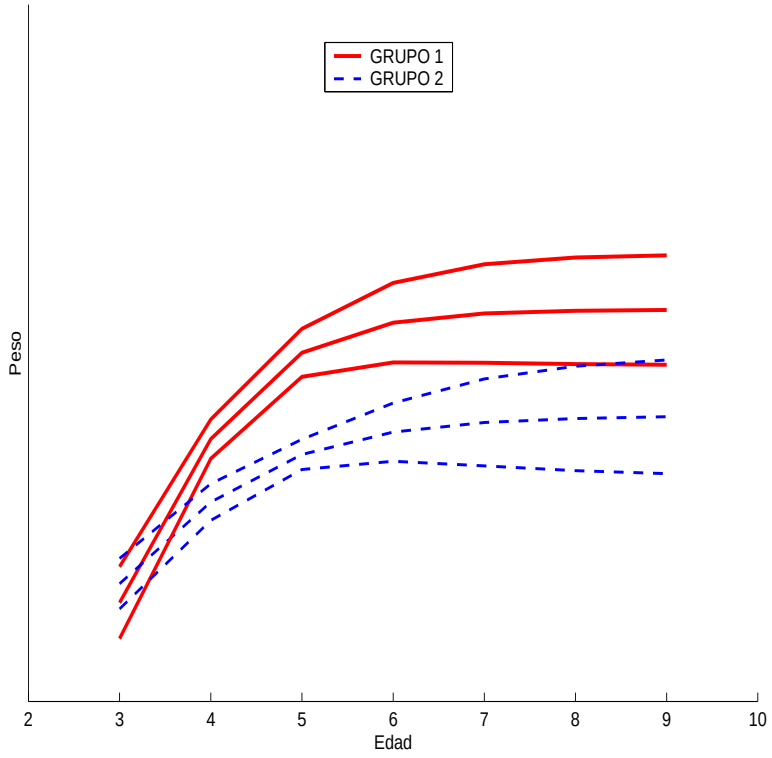
El método de Titulación (EL) es un método tracicional de análisis cuantitativo que se encuentra estipulado de manera normativa, sin embargo se desea validar el método de cromatografía (HPLC), un método novedoso y automatizado para la determinación del %RT, debido a que reduce las fuentes de error y permite que el proceso analítico se realize en menor tiempo. Se utilizaron modelos para comparar los valores de HPLC en función de los de EL y se encontró que el rango de detección del método HPLC es más amplio que el de EL, lo cual motiva a plantear la hipótesis de que HPLC tiene una mejor cobertura. Sin embargo, se encontró una gran variación en los datos, por lo que se recomendó un análisis más detallado de los procedimientos de laboratorio y de preparación de muestras, antes de realizar la validación.



2.5. Muestreo para estimar %RT (toma, preparación y manejo de la muestra)

Se utilizó un diseño anidado para estudiar la variabilidad de %RT dentro de un viaje de agave. El plan de muestreo consiste en particionar visualmente el contenido del viaje en tres secciones aproximadamente del mismo tamaño (ver tabla) y posteriormente seleccionar r grupos formados por una piña grande, dos medianas y dos chicas de cada sección. Se conforma una sola muestra con el material extraido de las 5 piñas y se mezcla con agua. A partir de ésta, se obtuvieron a su vez tres submuestras. Se encontró que el universo de las piñas tiene esta distribución (2/5 partes son medianas y grandes, 1/5 parte son grandes) y se propuso esta estratificación por tamaño para la disminuir del sesgo durante la selección de las piñas. La variación poco significativa de sección a sección y la variación importante entre grupos motivaron a proponer un muestreo estratificado por sección y por tamaños para reducir la variación del %RT promedio estimado.

	Sección 1				Sección 2				Sección 3			
	Grupo				Grupo				Grupo			
	1	2	...	r	1	2	...	r	1	2	...	r
Submuestra 1	Y_{111}	Y_{112}	...	Y_{11r}	Y_{211}	Y_{212}	...	Y_{21r}	Y_{311}	Y_{312}	...	Y_{31r}
Submuestra 2	Y_{121}	Y_{122}	...	Y_{12r}	Y_{221}	Y_{222}	...	Y_{22r}	Y_{321}	Y_{322}	...	Y_{32r}
Submuestra 3	Y_{131}	Y_{132}	...	Y_{13r}	Y_{231}	Y_{232}	...	Y_{23r}	Y_{331}	Y_{332}	...	Y_{33r}



2.6. Prácticas agrícolas. Del efecto de las prácticas agrícolas en el peso.

El ajuste de un modelo de crecimiento para el agave para distintos grupos de prácticas agrícolas ha permitido encontrar diferencias significativas entre ellos.La implicación práctica de esta conclusión consiste en que será posible:

- La evaluación de las condiciones del agave en cuanto a su peso.
- La evaluación de proveedores (contratistas).

- La revisión de las condiciones contractuales con el agricultor, dueño del predio.
- La reestructuración del universo del agave en cuestión.

3. Conclusiones

1. No se encontró una relación directa entre el %RT y el peso. Aunque la relación entre el peso y el %RT no sea directa, reducir la variación del peso del agave necesariamente reducirá la variación en el %RT. La repercusión práctica se traduce en que al tomar en cuenta ambas variables en el control de la calidad, se obtendrá un mejor control del producto y por consecuencia, un mejor control de los recursos.

2. No se encontró evidencia de que la zona de plantación ni la edad tengan un efecto significativo en el %RT. La posibilidad de adecuar los planes de jima a edades más tempranas implica la obtención de un beneficio sustancial mediante la reducción de los costos de producción, esto tomando en cuenta también las conclusiones sobre la relación entre el peso y el %RT.

3. La estratificación por grupos de prácticas agrícolas en el modelo de predicción del peso, permite una mejor planeación de recursos, mediante la evaluación de las condiciones del agave en cuanto a su peso, la evaluación de proveedores (contratistas), la revisión de las condiciones contractuales con el agricultor.

4. El método de muestreo que se propone permite reducir la variación del %RT dentro de un viaje de agave. Estimar el %RT promedio de un viaje de manera confiable conlleva a determinar también de manera confiable, el rendimiento del mismo dentro del proceso de producción del tequila.

Análisis de sensibilidad de un modelo de supervivencia semiparamétrico

Luis E. Nieto-Barajas[1]
Departmento de Estadística,
Instituto Tecnológico Autónomo de México

1. Introducción

El modelo de riesgos proporcionales de Cox (1972) es ampliamente utilizado para incluir información de covariables en el análisis de supervivencia. Este modelo supone que la función de riesgo entre dos individuos es proporcional, teniendo una función de riesgo común multiplicada por un factor que depende de las covariables. Para especificar el modelo, sea T_i una variable aleatoria no negativa que representa el tiempo de ocurrencia del evento de interés (tiempo de falla) del individuo i, y $Z_i(t) = (Z_{i1}(t), \ldots, Z_{ip}(t))$ el vetor de p covariables dependientes del tiempo. Entonces, la función de riesgo del individuo i es

$$h_i(t) = h_0(t) \exp\{Z_i(t)'\theta\}, \tag{1}$$

donde $h_0(t)$ es una función de riesgo base y $\theta' = (\theta_1, \ldots, \theta_p)$ es un vector de coeficientes de regresión. En este caso, la función de supervivencia del individuo i toma la forma

$$S_i(t) = \exp\left\{-\int_0^t h_i(s)\mathrm{d}s\right\}.$$

Cox (1972) propuso este modelo dejando sin especificar la función de riesgo base $h_0(\cdot)$. Por lo tanto, este modelo es de naturaleza semiparamétrico. Desde un punto de vista clásico, los estadísticos se han concentrado en estimar los coeficientes de regresión θ, maximizando una verosimilitud parcial (Cox, 1995), sin ningún o poco interés en la función de riesgo base. Por otro lado, en un enfoque Bayesiano, el interés se ha puesto en ambos, la función de riesgo base $h_0(t)$ y los coeficientes de regresión θ. Kalbfleisch (1978), por ejemplo, modeló la función de riesgo base con una inicial proceso gamma. Laud et al. (1998) usó una inicial proceso beta para modelar la función de riesgo acumulado. Más recientemente, Mezzetti & Ibrahim (2000)

[1] lnieto@itam.mx

usaron la inicial proceso de Markov gamma de Nieto-Barajas & Walker (2002) para modelar la función de riesgo base y lo llamaron proceso gamma correlacionado.

Sin embargo, ninguno de los análisis Bayesianos semiparamétricos del modelo de riesgos proporcionales han considerado explicitamente el caso de que las covariables varien con el tiempo. El objetivo de este artículo es estudiar el caso en el que las covariables son dependientes del tiempo usando la incial proceso de Markov gamma de Nieto-Barajas & Walker (2002).

En la Sección 2, la construcción de la inicial proceso de Markov gamma es revisada. La Sección 3 presenta el modelo semiparámetrico y las distribuciones posteriores son obtenidas. Finalmente, en la Sección 4 un análisis de sensibilidad es llevado a cabo usando un conjunto de datos simulado.

2. Inicial proceso de Markov gamma

Sea $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$ una partición del eje del tiempo en intervalos, y h_k la función de riesgo constante en el intervalo $(\tau_{k-1}, \tau_k]$, es decir,

$$h(t) = \sum_{k=1}^{\infty} h_k I_{(\tau_{k-1},\tau_k]}(t).$$

Entonces, la inicial proceso de Markov gamma se define como un proceso de Markov de primer orden en $\{h_k\}$ a través del uso de un proceso latente $\{u_k\}$ de la siguiente manera (Nieto-Barajas & Walker 2002):

$$h_1 \sim \text{Ga}(\alpha_1, \beta_1), \;\; u_k|h_k \sim \text{Po}(c_k h_k) \;\; \& \;\; h_{k+1}|u_k \sim \text{Ga}(\alpha_{k+1} + u_k, \beta_{k+1} + c_k),$$

para $k = 2, 3, \ldots$. Si $\alpha_k = \alpha_1$ y $\beta_k = \beta_1$ para todo k entonces el proceso $\{h_k\}$ es estacionario y $\text{Corr}(h_k, h_{k+1}) = c_k/(\beta_1 + c_k)$.

Aunque el proceso $\{h_k\}$ esta definido en tiempo discreto, la inicial proceso de Markov gamma $h(t)$ esta definida en tiempo continuo y asigna probabilidad uno al conjunto de funciones de distribución continuas.

3. Modelo semiparamétrico

A diferencia de la mayoría de los análisis Bayesianos del modelo de riesgos proporcionales, los cuales modelan la función de riesgo acumulado mediante un proceso estocástico, en este artículo modelamos la función de riesgo mediante un proceso estocástico. Considerando la función de riesgo (1) y usando la inicial proceso de Markov gamma descrito en la sección anterior para modelar la función de riesgo base $h_0(t)$, la función de riesgo acumulado para el individuo i toma la forma

$$H_i(t) = \sum_{k=1}^{\infty} h_k W_{i,k}(t, \theta),$$

donde,

$$W_{i,k}(t,\theta) = \begin{cases} \int_{\tau_{k-1}}^{\tau_k} e^{Z_i(s)'\theta} \mathrm{d}s & \text{si } t > \tau_k \\ \int_{\tau_{k-1}}^{t} e^{Z_i(s)'\theta} \mathrm{d}s & \text{si } t \in (\tau_{k-1}, \tau_k] \\ 0 & \text{e.o.c..} \end{cases}$$

Dada una muestra de observaciones posiblemente censuradas por la derecha $T_1, \ldots, T_n$, donde $T_1, \ldots, T_{n_u}$ son observaciones exactas y $T_{n_u+1}, \ldots, T_n$ son censuradas, la distribución posterior condicional para los parámetros del modelo semiparamétrico son:

- $f(h_k|\,\text{data}, u, \theta) = \text{Ga}(h_k|\alpha_k + u_{k-1} + u_k + n_k, \beta_k + c_{k-1} + c_k + m_k(\theta))$,
  donde, $n_k = \sum_{i=1}^{n_u} I(\tau_{k-1} < t_i \leq \tau_k)$ y $m_k(\theta) = \sum_{i=1}^{n} W_{i,k}(t_i, \theta)$

- $f(u_k|\,\text{data}, h, \theta) \propto \{c_k(c_k + \beta_{k+1})h_k h_{k+1}\}^{u_k} / \{\Gamma(u_k + 1)\Gamma(\alpha_{k+1} + u_k)\}$,
  for $u_k = 0, 1, \ldots$.

- $f(\theta|\,\text{data}, h, u) \propto f(\theta) \exp\{\sum_{i=1}^{n_u} Z_i(t_i)'\theta - \sum_{k=1}^{\infty} h_k m_k(\theta)\}$.

Más aún, si queremos introducir más flexibilidad dentro del proceso inicial $\{h_k\}$, podemos incorporar un proceso super-inicial en las $\{c_k\}$ de tal manera que $c_k \overset{IID}{\sim} \text{Ga}(1, \xi_k)$. El conjunto de distribuciones posteriores condicionales puede ser extendido para incluir

- $f(c_k|\text{data}, h, u, \theta) \propto (\beta_{k+1} + c_k)^{\alpha_{k+1}+u_k} c_k^{u_k} \exp\{-(\lambda_{k+1} + \lambda_k + \xi_k)c_k\}$, para $c_k > 0$.

Inferencia posterior puede ser obtenida implementando un esquema de muestreo de Gibbs. Simular de la distribución posterior condicional de h_k y u_k no tiene ninguna complicación. Si la distribución incial de θ es log-concava en cada argumento, la simulación de la distribución posterior condicional de θ y c_k puede llevarse a cabo usando la técnica de muestreo por rechaco adaptivo de (Gilks & Wild, 1992).

4. Análisis de sensibilidad

En esta sección ilustramos los resultados con un análisis Bayesiano completo y realizamos un análisis de sensibilidad de un conjunto de datos simulado. Este conjunto de datos simulado proviene de un modelo con una sola covariable dependiente del tiempo con las siguientes especificaciones: Sea $Z_i(t) = z_i \log(t)$ la covariable para el individuo i, y $h_0(t) = \lambda t$ la función de riesgo base. No es difícil demostrar que, en este caso, el tiempo de falla $T_i \sim \text{We}(\theta z_i + 2, \lambda/(\theta z_i + 2))$.

Simulamos una muestra de tamaño $n = 100$ con $\lambda = 1$, $\theta = 2$, $z_i = i/50$, para $i = 1, \ldots, 100$. Para especificar la inicial para $h(t)$, tomamos $\alpha_k = 0{,}001$ y $\beta_k = 0{,}01$ para todo k para tener condiciones iniciales relativamente no informativas (ver Nieto-Barajas & Walker, 2002), y tomamos un conjunto de valores $c_k = 0, 1, 5, 10, 20, 50$ para apreciar la sensibilidad de los estimadores posteriores del parámetro θ. Adicionalmente, escogimos una distribución inicial para $\theta \sim \text{No}(\mu_0, \sigma_0^2)$, con $\mu_0 = 0$ y un valor grande para la varianza $\sigma_0^2 = 9$. El muestreador de Gibbs se corrió 50,000 iteraciones con un período de calentamiento de 5,000.

La Figura 1 contiene un histograma de la distribución posterior de θ para los diferentes c_k's. De la figura se puede observar que todas las distribuciones tienen aproximadamente la misma localización, pero con varianzas ligeramente diferentes. Este comportamiento se resume también en la siguiente tabla.

Resúmenes posteriores de θ para distintos valores de c_k.

c_k	$\mathrm{E}(\theta\|t)$	$\sqrt{\mathrm{Var}(\theta\|t)}$	95 % HDI
0	2.05	0.57	(1.00, 3.16)
1	2.13	0.32	(1.54, 2.76)
5	2.26	0.41	(1.45, 3.06)
10	2.21	0.43	(1.36, 3.06)
20	2.17	0.47	(1.30, 3.10)
50	2.30	0.45	(1.44, 3.21)

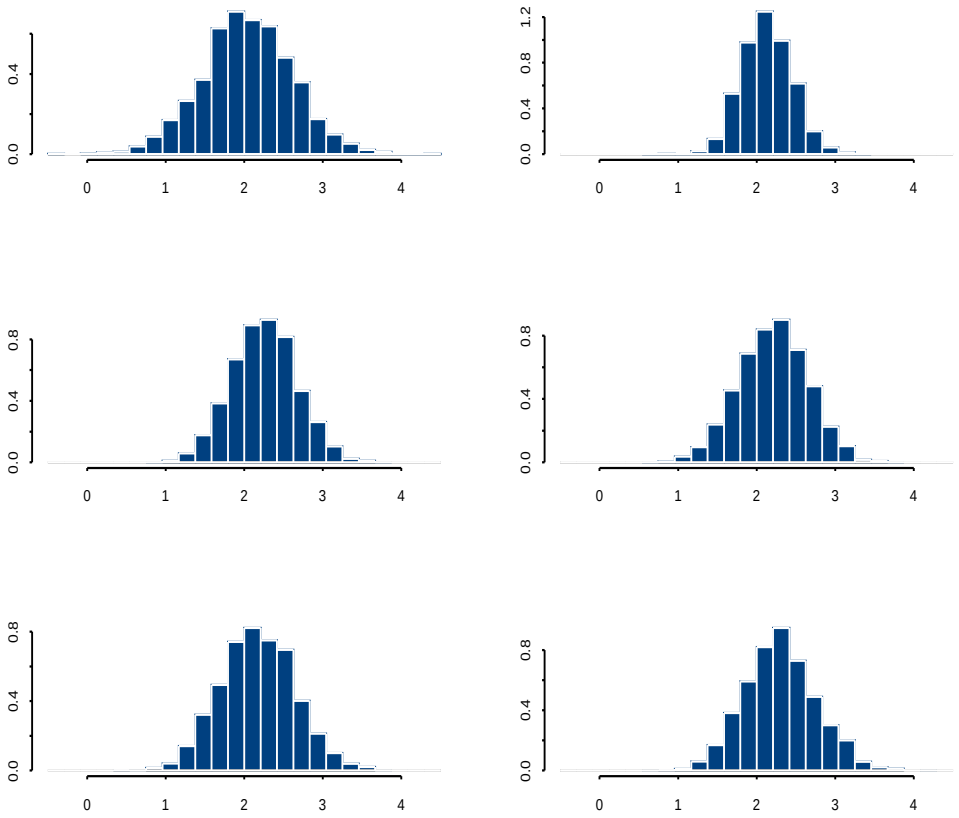


Figura 1: Distribución posterior de θ. De arriba a abajo y de izquierda a derecha: $c_k = 0$, $c_k = 1$, $c_k = 5$, $c_k = 10$, $c_k = 20$, $c_k = 50$.

De la tabla anterior podemos observar que la media posterior de θ no cambia dramáticamente para los distintos valores de c_k y no se alejan mucho del valor $\theta = 2$. La desviación estándar posterior es también estable y los intervalos de alta densidad al 95 % todos claramente

contienen el verdadero valor de θ. Por lo tanto, se puede decir que los estimadores puntuales para θ no son muy sensitivos a la elección del parámetro c_k. Con respecto a la función de riesgo, un comentario final es que conforme c_k crece, los estimadores posteriores están más cercanos de la verdadera función.

Referencias

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-202.

Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.

Gilks, W.R. and Wild, P. (1992). Adaptive Rejection sampling for Gibbs sampling. *Applied Statistics - Journal of the Royal Statistical Society, Series C* **41**, 337-348.

Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* **40**, 214-221.

Laud, P.W., Damien, P. and Smith, A.F.M. (1998). Bayesian nonparametric and covariate analysis of failure time data. In *Practical nonparametric and semiparametric Bayesian statistics*. D. Dey, P. Müller and D. Sinha (Eds). Springer. New York.

Mezzetti,M. and Ibrahim,J.G. (2000). Bayesian inference for the Cox model using correlated gamma process priors. *Technical report, Department of Biostatistics, Harvard School of Public Health.*

Nieto-Barajas, L.E. and Walker, S.G. (2002). Markov beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics* **29**, 413-424.

Análisis bayesiano de la distribución von Mises-Fisher

Gabriel Nuñez Antonio[1]
Instituto Tecnológico Autónomo de México
Eduardo Gutiérrez-Peña[2]
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México

1. Introducción

El objetivo de este trabajo es presentar un análisis Bayesiano del modelo von Mises-Fisher. Se dice que un vector aleatorio unitario p-dimensional tiene una distribución von Mises-Fisher, con vector de dirección media $\boldsymbol{\mu}$ y parámetro de concentración κ, denotada por $vMF(\boldsymbol{x}|\boldsymbol{\mu},\kappa)$, si su función de densidad de probabilidad está dada por

$$f(\boldsymbol{x}|\boldsymbol{\mu},\kappa)=\left(\frac{\kappa}{2}\right)^{p/2-1}\frac{1}{\Gamma(p/2)I_{p/2-1}(\kappa)}\exp\{\kappa\boldsymbol{\mu}^t\boldsymbol{x}\}I_{\mathbb{S}^p}(\boldsymbol{x})I_{\mathbb{S}^p}(\boldsymbol{\mu})I_{(0,\infty)}(\kappa), \tag{1}$$

donde $\mathbb{S}^p=\{\boldsymbol{x}\in\mathbb{R}^p:\boldsymbol{x}^t\boldsymbol{x}=1\}$ y $I_b(\cdot)$ denota la función de Bessel modificada de primer tipo de orden b. La distribución correspondiente para datos circulares, $p=2$, es conocida como distribución von Mises. Esta es la distribución más importante en el análisis de datos circulares. Además, es el análogo natural sobre el círculo de la distribución normal sobre la recta real. En el caso de datos esféricos, $p=3$, la distribución von Mises-Fisher es conocida como distribución Fisher.

Los datos direccionales aparecen de manera natural en varias áreas . Para una revisión desde el punto de vista frecuentista el lector puede consultar Fisher *et al.* (1987) y Mardia y Jupp (2000). La literatura Bayesiana es menos extensa y en cierto sentido no es del todo adecuada. Lo anterior debido a las dificultades que presenta trabajar con las distribuciones de probabilidad asociadas con el análisis de datos direccionales, en particular con la von Mises-Fisher.

[1] gabriel@itam.mx
[2] eduardo@sigma.iimas.unam.mx

El intento más reciente de un análisis Bayesiano completo para la distribución von Mises se debe a Damien y Walker (1999), quienes desarrollan un muestreo de Gibbs basado en el uso de variables auxiliares. Sin embargo, la implementación de su metodología no es muy eficiente en algunos casos. Específicamente, la autocorrelación de las muestras de la distribución final es demasiado alta para valores grandes del parámetro de concentración (digamos $\kappa \geq 7$), lo cual resulta en una convergencia lenta del algoritmo propuesto.

En este trabajo implementamos un análisis Bayesiano completo de la distribución von Mises-Fisher considerando todos los parámetros desconocidos. En la sección 2 se obtiene la distribución final de $(\boldsymbol{\mu}, \kappa)$. Se hace uso del hecho que la distribucción von Mises-Fisher pertenece a una familia exponencial regular y se considera la correspondiente distribución inicial conjugada estándar para esta familia (ver Bernardo y Smith (1994)). En la Sección 3 se describe nuestra propuesta para generar muestras de la distribución final conjunta de $(\boldsymbol{\mu}, \kappa)$ vía un simple esquema de muestreo-remuestreo por importancia. Finalmente, en la Sección 4 se dan algunos ejemplos numéricos.

2. La distribución von Mises-Fisher

Sea $\boldsymbol{D}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ una muestra de tamaño n de $f(\boldsymbol{x}|\boldsymbol{\mu}, \kappa)$. La verosimilitud de $(\boldsymbol{\mu}, \kappa)$ está dada por

$$L(\boldsymbol{\mu}, \kappa; \boldsymbol{D}_n) \propto \left(\frac{\kappa^{p/2-1}}{I_{p/2-1}(\kappa)} \right)^n \exp \left\{ \kappa \boldsymbol{\mu}^t \sum_{i=1}^{n} \boldsymbol{x}_i \right\}.$$

De esta expresión se puede derivar fácilmente una distribución inicial conjugada para $(\boldsymbol{\mu}, \kappa)$, la cual se describe en la siguiente subsección .

2.1. Distribución Inicial

Para el modelo von Mises-Fisher se propone la siguiente inicial conjugada

$$f(\boldsymbol{\mu}, \kappa) \propto \{C_p(\kappa)\}^c \exp\{\kappa R_0 \boldsymbol{m}^t \boldsymbol{\mu}\}, \tag{2}$$

donde $C_p(\kappa) = \left(\frac{\kappa}{2}\right)^{p/2-1} \frac{1}{I_{p/2-1}(\kappa)}$, y con $\boldsymbol{m} \in \mathbb{S}^p$, c y R_0 definidos en $(0, \infty)$.

Se debe notar que, bajo esta inicial, la distribución condicional de $\boldsymbol{\mu}$ dado κ es $vMF(\boldsymbol{\mu}|\boldsymbol{m}, \kappa R_0)$.

2.2. Distribución Final

Sea $\boldsymbol{D}_n = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ una muestra de tamaño n de (1) y suponga que $(\boldsymbol{\mu}, \kappa)$ está dado por la inicial conjugada (2). La distribución final de $(\boldsymbol{\mu}, k)$ resulta

$$f(\boldsymbol{\mu}, k|\boldsymbol{D}_n) \propto \{C_p(k)\}^{(c+n)} \exp\{\kappa R_n \boldsymbol{m}_n^t \boldsymbol{\mu}\}, \tag{3}$$

donde R_n es la longitud de la resultante promedio y $\boldsymbol{m}_n$ es el vector de direcciones coseno que se obtienen al combinar la inicial conjugada y los datos observados.

3. Inferencias vía muestreo-remuestreo por importancia

Para obtener muestras simuladas de la distribución final se utilizó una técnica de muestreo conocida como muestreo-remuestreo por importancia (SIR, por sus siglas en inglés) (Rubin, 1988). La idea básica se describe a continuación.

Suponga que se tiene una densidad $g(\varphi)$ de la cual se puede generar una muestra, pero que lo que se desea es una muestra de la densidad

$$h(\varphi) = \frac{f(\varphi)}{\int f(\varphi')d\varphi'}$$

donde sólo se conoce la forma funcional de $f(\varphi)$. El problema consiste en obtener una muestra de $h(\varphi)$ a partir de $f(\varphi)$ y la muestra de $g(\varphi)$. El procedimiento para resolver este problema es el siguiente. Dado $\varphi_1, ..., \varphi_N$ de $g(\varphi)$ obtener

$$q_i = \frac{\omega_i}{\sum_{i=1}^N \omega_i},$$

donde $\omega_i = f(\varphi)/g(\varphi)$, $i = 1, ..., N$. La última expresión induce una distribución discreta sobre $\{\varphi_1, ..., \varphi_N\}$ con $\Pr(\varphi = \varphi_i) = q_i$. Si ahora se simula φ^* de esta distribución discreta, entonces φ^* se distribuirá aproximadamente como una variable aleatoria con distribución $h(\varphi)$. Esta aproximación resultará mejor a medida que $N \to \infty$ (ver, por ejemplo, Bernardo y Smith, 1994).

Para obtener muestras de la distribución final (3) se propone tomar

$$g(\boldsymbol{\mu}, \kappa) = vMF(\boldsymbol{\mu}|\boldsymbol{\mu}_c, \kappa_c)\Gamma(\kappa|a_c, b_c),$$

donde $vMF(\boldsymbol{\mu}|\boldsymbol{\mu}_c, \kappa_c)$ es una densidad von Mises-Fisher con dirección media definida por el vector unitario $\boldsymbol{\mu}_c$ y parámetro de concentración κ_c. Hay que notar que esta selección de la distribución para el muestreo por importancia para $\boldsymbol{\mu}$ es razonable ya que la distribución condicional final verdadera de $\boldsymbol{\mu}$ dado κ es una $vMF(\cdot|\boldsymbol{m}_n, \kappa R_n)$. Una selección alternativa, para valores moderados de κ podría ser tomar $g(\boldsymbol{\mu}, \kappa) = U(\boldsymbol{\mu})\Gamma(\kappa|a_c, b_c)$, donde $U(\cdot)$ es una densidad uniforme sobre $\mathbb{S}^p$.

4. Ejemplos numéricos

Ejemplo 1. En este ejemplo se considera un conjunto de 26 mediciones de residuos magnéticos tomados de muestras recolectadas de yacimientos rojizos (ver Fisher *et al.*, 1987; Apéndice B2). Fisher *et al.* (1987) muestran, usando procedimientos tanto gráficos como formales, que este conjunto de datos puede modelarse razonablemente bien con una distribución Fisher. Para estos datos, los correspondientes estimadores de máxima verosimilitud resultan ser $(\hat{\alpha}, \hat{\beta}) = (147{,}2, 215{,}8)$ y $\hat{\kappa} = 113$. Para este ejemplo se tomo $c = 0$, $\mu_0 = 0$ y $R_0 = 0$. En este caso se consideró $g(\alpha, \beta, \kappa) = U(\alpha|0, \pi)U(\beta|0, 2\pi)\Gamma(\kappa|a_2, b_2)$ con a_1 y b_1 de tal manera que la media y la varianza de κ fueran 110 y 1000, respectivamente. Las distribuciones marginales resultantes de α, β y κ se muestran en la Figura 1.

En este ejemplo se usaron densidades uniformes como parte de la densidad $g(\cdot)$. Aunque una densidad Fisher podría resultar mejor en términos del muestreo-remuestreo por importancia, los requerimientos computacionales serían mayores. Para este ejemplo, se observó que las densidades uniformes propuestas producen resultados satisfactorios.

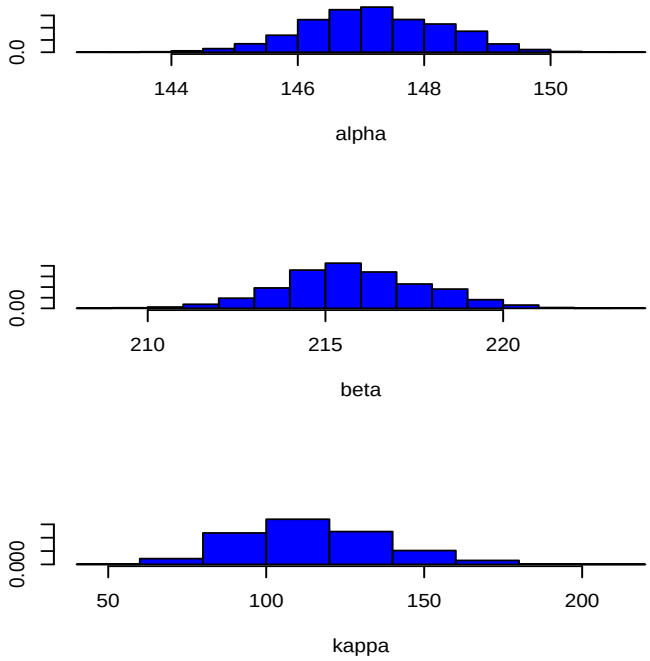


Figura 1: Distribuciones finales de α, β y κ

Ejemplo 2. En este ejemplo se simuló una muestra de tamaño 1000 de una distribución von Mises-Fisher con vector de dirección media $\boldsymbol{\mu} = (0, 0, 0, 0, 1)$ y $\kappa = 50$ usando la propuesta de Wood (1994) basada en el algoritmoVM de Ulrich (1984). Se tomó $c = 0$ y $\boldsymbol{m} = \mathbf{0}$. Para este ejemplo se consideró $g(\boldsymbol{\mu}, \kappa) = vMF(\boldsymbol{\mu}|\boldsymbol{\mu}_c, \kappa_c)\Gamma(\kappa|a_c, b_c)$ con $\boldsymbol{\mu}_c = (0, 0, 0, 0, 1)$, $\kappa_c = 500$ y a_c y b_c de tal manera que la media y la varianza de κ fueran 50 y 1000, respectivamente. Las distribuciones marginales resultantes de $\boldsymbol{\mu}$ y κ se muestran en la Figura 2.

5. Conclusiones

En este trabajo se presentó un análisis Bayesiano completo para datos direccionales vía el modelo von Mises-Fisher considerando todos los parámetros desconocidos.

Aunque en algunos casos otros métodos más complejos pueden ser más atractivos que el presentado aquí, la propuesta basada en el algoritmo de muestreo-remuestreo por importancia

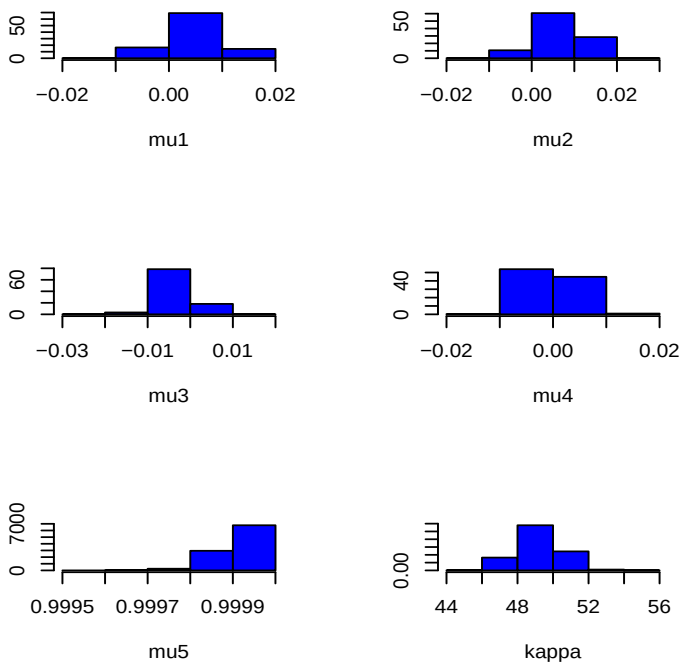


Figura 2: Distribuciones finales de $\boldsymbol{\mu}$ (por componente) y κ

es una buena alternativa y puede ser más eficiente a pesar de su forma tan simple. Por otra parte, este procedimiento no tiene el mismo problema que el muestreo de Gibbs para valores grandes del parámetro de concentración κ.

Consideramos que la metodología presentada en este trabajo ofrece las bases de un análisis Bayesiano completo para datos direccionales. Sin embargo, debe señalarse que la extensión de este análisis a contextos más complejos tales como análisis de regresión puede no resultar tan directa.

Referencias

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

Damien, P. y Walker, S.G. (1999). A full Bayesian analysis of circular data using the von Mises distribution. *Canad. J. Statist.* **27**, 291–298.

Fisher, N.I., Lewis, T. y Embleton, B.J.J. (1987). *Statistical Analysis of Spherical Data*. Cambridge, University Press.

Mardia, K.V. y Jupp, P.E. (2000). *Directional Statistics*. Chichester, Wiley.

Rubin, D.B. (1988). Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley y A.F.M. Smith, eds.). Oxford: University Press, pp. 395–402 (con discusión).

Ulrich, G. (1984). Computer generation of distributions on the m-sphere. *J. Appl. Statist.* **33**, 158–163.

Wood A.T.A. (1994). Simulation of the von Mises-Fisher distribution. *Comm. Statist., Simulation Comput.* **23**, 157–164.

Métodos modernos de optimización en clasificación por particiones

Javier Trejos Z.[1]
CIMPA, Universidad de Costa Rica

1. Introducción

Muchos métodos usuales en Análisis Multivariado de Datos encuentran óptimos locales de los criterios que minimizan. Tal es el caso en clasificación por particiones, escalamiento multidimensional, regresión no lineal y conjuntos burdos. Recientemente, muchas heurísticas de optimización han sido propuestas; su finalidad es encontrar óptimos globales en problemas de optimización discreta. Entre estas heurísticas, están el sobrecalentamiento simulado (SS), la búsqueda tabú (BT), los algoritmos genéticos (AG), las colonias de hormigas (ACO), y los enjambres de partículas (PSO).

Muchos autores han tratado de encontrar mejores soluciones a los problemas de Análisis Multivariado de Datos utilizando estas heurísticas de optimización. En la Universidad de Costa Rica, el equipo PIMAD del Centro de Investigación en Matemática Pura y Aplicada ha abordado distintos problemas: clasificación numérica, clasificación binaria, clasificación bimodal, escalamiento multidimensional métrico, regresión no lineal, selección de variables, y conjuntos burdos.

2. Problemas de clasificación

El objetivo de la clasificación automática (conocida como *clustering* en inglés) es encontrar grupos homogéneos de objetos, de tal forma que objetos similares pertenezcan a la misma clase, y que sea posible distinguir entre objetos que pertenezcan a clases diferentes. En el caso numérico, se tiene el conjunto de objetos $\Omega = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ tal que $\forall i : \mathbf{x}_i \in I\!R^p$, esto es, los objetos son descritos por p variables numéricas o cuantitativas. El criterio más ampliamente

[1] jtrejos@cariari.ucr.ac.cr

usado es la minimización de la varianza o inercia intraclases:

$$W(P) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2 \tag{1}$$

donde K es el número (fijado de antemano) de clases, $P = (C_1, \ldots, C_K)$ es la partición que se busca, y $\mathbf{g}_k$ es el centro de gravedad o vector promedio de C_k. Debe observarse que el criterio W satisface la propiedad de monotonicidad: mín$\{W(P) \in \mathcal{P}_{k+1}^*\} \leq$ mín$\{W(P) \in \mathcal{P}_k^*\}$ donde $\mathcal{P}_k^*$ es el conjunto de todas las particiones de Ω en exactamente k clases no vacías. Esto significa que no tiene sentido comparar particiones con diferente número de clases, y por ello el número de clases es fijado de antemano.

Los métodos clásicos, como k-medias (búsqueda local) o clasificación jerárquica (método voraz) encuentran por lo general mínimos locales de W (Diday et al. (1982)). Por lo tanto, hemos aplicado las heurísticas modernas de optimización que se mencionan a continuación, buscando la mejoría de los resultados dados por los métodos existentes.

En el caso de datos binarios los objetos $\mathbf{x}_i$ pertenecen al conjunto $\{0, 1\}$. El primer problema es definir un criterio numérico de homogeneidad. Nosotros hemos estudiado (ver Piza et al. (2000)) las propiedades de varios criterios aditivos del tipo $W(P) = \sum_{k=1}^{K} \delta(C_k)$ donde $\delta(C)$ es una medida de homogeneidad de la clase C, y los que brindaron mejores resultados fueron la suma de las disimilitudes:

$$\delta_{\text{sum}}(C) = \sum_{\mathbf{x}, \mathbf{x}' \in C} d(\mathbf{x}, \mathbf{x}') \tag{2}$$

y la suma ponderada de las disimilitudes:

$$\delta_{\text{wsm}}(C) = \frac{1}{2|C|} \sum_{\mathbf{x}, \mathbf{x}' \in C} d(\mathbf{x}, \mathbf{x}'). \tag{3}$$

Estos criterios satisfacen la propiedad de monotonicidad.

En el caso de clasificación bimodal o clasificación cruzada, sea $X = (x_{ij})$ una matriz bimodal (por ejemplo, una tabla de contingencia) con $x_{ij} \geq 0$. Se quiere minimizar el criterio de varianza definido por:

$$W(P, Q) = \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i \in A_k} \sum_{j \in B_l} (x_{ij} - g_{kl})^2 \tag{4}$$

donde (P, Q) es una partición bimodal (esto es, una partición $P = (A_1, \ldots, A_K)$ de las filas y una partición $Q = (B_1, \ldots, B_L)$ de las columnas de X), y $g_{kl} = \sum_{i \in A_k} \sum_{j \in B_l} x_{ij}$.

3. Uso de heurísticas de optimización

Se define un problema de optimización combinatoria como la minimización de una función real de costo F, definida sobre un conjunto finito o numerable de estados $\mathcal{S}$. Entre los métodos más sencillos de optimización combinatoria, están la búsqueda local y los métodos voraces,

En Trejos et al. (1998) describimos nuestra aplicación de sobrecalentamiento (recocido) simulado, búsqueda tabú y algoritmos genéticos al problema de clasificación con datos numéricos. Los primeros dos métodos se basan en la definición de vecindarios, que en este caso se trata de particiones generadas a partir de la transferencia de un único elemento de una clase a otra. Luego se aplican las heurísticas de manera similar a los planteamientos iniciales de Kirkpatrick et al. (1983) o Glover et al. (1993), con algunos ajustes y escogencias razonadas de los parámetros. En Piza et al. (2000) se describe el uso de las heurÍsticas en clasificación de datos binarios. Las implementaciones del SS y la BT son como en el caso numérico. Para el AG, la función de fitness se define como $B(P) = W(\Omega) - W(P)$, y el resto de características del algoritmo son como en el caso numérico. Los resultados para SS y BT fueron excelentes, y bastante pobres para AG. No es posible hacer una adaptación del PSO a este caso en vista de que no se dispone de centro de clases en un espacio numérico. La adaptación al uso de ACO aún no se ha hecho. En Trejos y Castillo (2000) describimos el uso de SS, y en Castillo y Trejos (2002) la aplicación de BT a este problema.

Son menos conocidas las heursticas de colonias de hormigas (ver Bonabeau et al. (1999)) y enjambres de partículas (ver Kennedy y Eberhardt (2000)). En ambas se manejan, como en algoritmos genéticos, conjuntos de agentes. En colonias de hormigas (Trejos et al. (2002)), estos agentes se asocian a particiones y las buenas soluciones hacen crecer la probabilidad de asignar objetos a la misma clase. En enjambres de partículas (Trejos et al. (2003)) las particiones están asociadas a sus centros de gravedad, que se mueven en el espacio multidimensional con la particularidad de que los agentes–particiones se comunican de acuerdo con tres principios: inercia, imitación al mejor agente, y regresión a la mejor posición encontrada

anteriormente.

4. Resultados comparativos

Se han llevado a cabo numerosas comparaciones entre los métodos propuestos por los autores usando las heurísticas de optimización aquí descritas, y los métodos clásicos de clasificación, tanto con datos reales como simulados. En Trejos et al. (1998) se presentan comparaciones con tablas de datos reales, siendo el sobrecalentamiento simulado el que obtiene los mejores resultados. Se llevó a cabo también un experimento exhaustivo para comparar SS, BT y AG con k-medias y clasificación jerárquica de Ward. Se generaron tablas con generadores de números semi-aleatorios, y se llevó a cabo un experimento con 4 factores y dos niveles en cada uno. Los factores son: el número n de individuos; se tomó $n = 105$ y $n = 525$, el número K de clases; se tomó $K = 3$ y $K = 7$, la cardinalidad de las clases; se tomó todas las clases con misma cardinalidad, y una clase mayor que el resto (con aproximadamente el 50 % de todos los objetos), la varianza de las clases; se tomó todas las clases con igual varianza, y una clase con el triple de varianza que el resto. En todos los casos las tablas tienen $p = 6$ variables. Los vectores de medias fueron generados al azar en $[0, 1]^6$. Por lo tanto, se tiene 16 casos, y para cada uno se generaron 100 particiones iniciales al azar antes de aplicar los métodos de particionamiento. En este experimento se puede medir el porcentaje de mala clasificación, por la forma en que se construyeron las tablas de datos. En Pacheco (2003) se pueden consultar los valores de los parámetros utilizados en cada heurística, así como los resultados correspondientes; en este caso, el algoritmo genético fue el que brindó los mejores resultados. En cuanto a los tiempos de ejecución, el método de k-medias es mucho más rápido que los demás, y el más lento es el BT. En promedio este último tardó unos 16 minutos por corrida, el AG 2 minutos, el SS unos 30 segundos, y el k-medias muy pocos segundos.

En el caso binario, en Piza et al (2000) se presentan resultados que muestran de nuevo la superioridad del sobrecalentamiento simulado, lo mismo que en Castillo y Trejos (2002) para el caso bimodal.

Referencias

Aarts, E. y Korst, J. (1990). *Simulated Annealing and Boltzmann Machines*. Chichester: Wiley.

Bonabeau, E., Dorigo, M. y Therauluz, G. (1999). *Swarm Intelligence. From Natural to Artificial Systems*. New York: Oxford University Press.

Castillo, W. y Trejos, J. (2002). Two-mode partitioning: review of methods and application of tabu search. In *Classification, Clustering, and Data Analysis* (eds. K. Jajuga et al.), pp. 43-51. Berlin: Springer.

Diday, E., Lemaire, J., Pouget, J. y Testu, F. (1982). *Eléments d'Analyse de Données*. Paris: Dunod.

Glover, F. et al. (1993). Tabu search: an introduction. *Annals of Operations Research*, **41**, 1-28.

Goldberg, D. E. (1989). *Genetic Algorithm in Search, Optimization and Machine Learning*. Reading: Addison-Wesley.

Kennedy, J. y Eberhart, R.C. (2000). *Intelligent Swarm Systems*. New York: Academic Press.

Kirkpatrick, S., Gelatt, D. y Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, **220**, 671-680.

Pacheco, A. (2003). Experimentación para la comparación de métodos de clasificación automática. Preprint CIMPA. San José: Universidad de Costa Rica.

Piza, E., Trejos, J. y Murillo, A. (2000). Clustering with non-Euclidean distances using combinatorial optimisation techniques. In *Science Methodology in the New Millenium* (eds. J. Blasius et al.), CD-Rom paper Nr.P090504, ISBN 90-801073-8-7.

Trejos, J., Piza, E. y Murillo, A. (1998). Global stochastic optimization techniques applied to partitioning. In *Advances in Data Science and Classification*, (eds. M. Rizzi et al.), pp.

185-190. Berlin: Springer.

Trejos, J. y Castillo, W. (2000). Simulated annealing optimization for two-mode partitioning. In *Classification and Information Processing at the Turn of the Millenium* (eds. W. Gaul & R. Decker), pp. 133-142. Berlin: Springer.

Trejos, J., Goddard, J., Cobos, S. y Piza, E. (2002). Clasificación de datos numéricos mediante optimización por enjambres de partículas. *5th International Conference on Operations Research*. La Habana.

Trejos, J., Piza, E. y Murillo, A. (2003). Hormigas que modifican particiones para clasificación automática. *6th International Conference on Operations Research*. La Habana.

Esta publicación consta de 410 ejemplares y se terminó de
imprimir en el mes de enero de 2005 en los talleres gráficos del
Instituto Nacional de Estadística, Geografía e Informática
Av. Héroe de Nacozari Sur Núm. 2301, Puerta 11, Nivel Acceso
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.
México