ISBN: 84-699-2357-9

HOMOGENEIDAD DE POBLACIONES ESTADISTICAS. EL PROBLEMA DE LA MIXTURA DE COMPONENTES.

AUTORES:

Fajardo Caldera, M.A. (fajardo@unex.es): Perez Mayo, Jesús (jperez@unex.es):

Andrades Caldito, Lydia. Dptº. de Economía Aplicada y Org. de Empresas. Universidad de Extremadura.

RESUMEN:

En este artículo, los autores analizan el problema de la homogeneidad de poblaciones. Este consiste en dividir una población en subpoblaciones y estudiar si la distribución de probabilidad es la misma en ellas. Si esto es afirmativo, entonces podremos trabajar con datos agregados, en caso contrario sería conveniente trabajar con las subpoblaciones. El ignorar la heterogeneidad conduce a conclusiones equivocadas (paradoja de Simpson).

Existen un considerable conjunto de técnicas estadísticas para analizar si una población es homogénea respecto a alguna o varias características cuando estas son observables (Anova, Manova, Regresión multivariante, etc.); el problema surge cuando no conocemos “a priori ” estas características, es decir, son no observables. La aplicación que trataremos en este articulo, será analizar si la variable ingresos totales netos de los hogares españoles en el año 1994 es una distribución homogénea, a través de la técnica estadística conocida con el nombre de análisis de mixtura de componentes y su resolución por el algoritmo EM.

INTRODUCCIÓN.-

El análisis de homogeneidad de poblaciones consiste en dividir una población en subpoblaciones y estudiar si la distribución de una o varias variables aleatorias es la misma en todas ellas. En este caso, se podrá trabajar con los datos agregados. En caso contrario, será conveniente trabajar con las subpoblaciones existentes, Peña y Romo (1).

El ignorar la heterogeneidad debida a la presencia de subpoblaciones puede conducir a conclusiones equivocadas en el análisis, ya que no tenemos una representación clara de la variable, no mejoramos la comprensión del fenómeno en estudio y podemos incurrir en la famosa Paradoja de Simpson (2), quien demostró que al mezclar datos que provienen de distintas poblaciones y, por tanto, son heterogéneos, podemos llegar a conclusiones opuestas a las obtenidas teniendo en cuenta las subpoblaciones.

La Ciencia Estadística ha proporcionado un considerable número de herramientas para poder analizar la homogeneidad de poblaciones cuando la variable grupo y las variables a analizar son observables. Los modelos más conocidos son el Anova (para el estudio de una única variable) y el Manova (para el estudio de un conjunto finito de variables), siempre que se conozca "a priori" la asignación de las observaciones a los grupos, analizándose posteriormente mediante un contraste de igualdad de medias, supuestas que las poblaciones son normales y homocedásticas. En el caso de que se acepte la hipótesis nula de igualdad de medias, entonces diremos que las poblaciones son homogeneas.

El problema surge cuando no disponemos de información " a priori " que nos indique si existe una división de la población en subpoblaciones, es decir, cuando la variable grupo es no observable. Este es el problema que trataremos en este artículo, en el que la variable continua observable viene definida por los ingresos totales netos de los hogares españoles en el año 1994 (Panel de Hogares de la U.E.), y la variable grupo (discreta) es no observable, problema que es conocido, en el campo de la estadística, con el nombre de análisis de mixtura de componentes. Su resolución se basa en el conocido algoritmo EM y en el contraste de hipótesis de homogeneidad a través de los modelos mixtos de variables continuas y discretas, introducidos por Lauritzen y Wermuth en 1989 (3) y su extensión por Edwards en 1990 (4) a los modelos de interacción jerárquica y más tarde construidos por combinación de los modelos log-lineales para variables discretas con los Modelos Gaussianos Gráficos (MGG) para variables continuas por Whittaker (1990) (5) y Edwards (1995) (6).

ANALISIS DE MIXTURA DE DISTRIBUCIONES ESTADÍSTICAS.-

El análisis de las distribuciones mixtas para datos agrupados consiste matemáticamente en el estudio de una función de densidad de probabilidad mixta, la cual es una suma ponderada de k funciones de densidad componentes, donde k es asumido a priori para ser conocido, es decir,

$$f(x/\mu,\sigma) = p_1 f(x/\mu_1,\sigma_1) + p_2 f(x/\mu_2,\sigma_2) + \ldots\ldots + p_k f(x/\mu_k,\sigma_k) \qquad [1]$$

Las densidades componentes pueden ser normales, lognormal, gamma, exponencial o Weibull. Los parámetros son las proporciones de la mixtura, las medias y las desviaciones estándar de las distribuciones componentes. Diversas restricciones pueden ser impuestas a los parámetros.

El caso que vamos a desarrollar es el de una muestra aleatoria, $x_1, x_2, x_3, \ldots\ldots\ldots, x_n,$ extraída de una población con función de densidad dada en [1], con las distribuciones componentes normales y homocedásticas.

Dada la muestra, y fijado un k "a priori ", estimaremos los parámetros y posteriormente contrastaremos la igualdad de medias de las distribuciones componentes, es decir, si los datos están descritos adecuadamente por una componente.

La distribución a posteriori de que un elemento con respuesta x, pertenezca a la clase j=1,2,....,k , viene dada por :

$$h(j/x) = p_j\, f(x/j)/\, f(x) \qquad j=1,2,.....k \qquad [\,2\,]$$

donde las f(x/j) son normales $N(\mu_j, \sigma)$, para todo j = 1,2,....k.

El logaritmo de la función de verosimilitud viene dada por:

$$l = \log \prod_i f(x_i) = \sum_i \log f(x_i) = \sum_i \log [p_1\, f(x_i /\mu_1,\sigma_1) + p_2\, f(xi /\mu_2,\sigma_2) + \ldots\ldots$$

$$+ p_k\, f(xi /\mu_k,\sigma_k)] \qquad [\,3\,]$$

Dado que la suma de las proporciones han de ser igual a 1, es decir, $\sum_i p_i = 1$, tendremos que maximizar la función:

$$\phi = \sum_i \log [p_1 f(x_i /\mu_1,\sigma_1) + ..+ p_k f(xi /\mu_k,\sigma_k)] + \theta (\sum_i p_i -1) \qquad [4]$$

Para la obtención de los estimadores maximoverosimiles, resolvemos las ecuaciones:

$$\partial \phi / \partial p_j = \sum_i f(xi /j) / f(xi) + \theta = 0 \qquad [5]$$

$$\partial \phi / \partial \mu_j = \sum_i p_j f(x_i/j) / f(xi) = 0 \qquad [6]$$

$$\partial \phi / \partial \sigma^2 = \partial /\partial \sigma^2 [\sum \log [p_1 f(x_i /\mu_1,\sigma_1) ++ p_k f(xi /\mu_k,\sigma_k)] /f(x_i) = 0 \quad [7]$$

De las ecuaciones anteriores se deducen los siguientes estimadores en función de las probabilidades a posteriori f(j /xi):

$$\hat{p}_j = \sum_i f(j/ xi) / n \qquad j= 1,2,...,k \qquad [8]$$

$$\hat{\mu}_j = \sum_i xi\, f(j/ xi) / \sum_i f(j/ xi). \quad j= 1,2,....k \qquad [9]$$

$$\hat{\sigma}^2 = \sum_i [\sum_j (xi - \mu_j)^2 f(j/ xi)] / n. \qquad [10]$$

Sin embargo, si f(j / xi) fuese conocida, sería muy fácil resolver las ecuaciones [8], [9] y [10] para obtener las estimaciones de los parámetros, pero ésta es bastante complicada de calcularla, ya que su definición viene dada por:

$$h(j/x) = p_j f(x/j) / f(x) = p_j (2\pi \sigma^2)^{-1/2} \exp -1/2\{ (x- \mu_j)^2 /\sigma^2 \} / [p_1 f(x /\mu_1,\sigma_1) + p_2 f(x/\mu_2,\sigma_2) + ... + p_k f(x /\mu_k,\sigma_k)]. \qquad [11]$$

Para ello, es más útil aplicar el EM algoritmo, el cual tiene la ventaja de conseguir estimaciones de los parámetros de la siguiente forma :

1) Elegimos un conjunto de valores iniciales para las probabilidades a posteriori $\{f(j/xi)\}$.
2) Utilizando las ecuaciones [8], [9] y [10] obtenemos las primeras aproximaciones de los estimadores de p_j, μ_j y de σ^2.
3) Sustituimos estos valores estimados de nuevo en [11], para obtener mejores estimaciones de $\{f(j/xi)\}$.
4) Volviendo al paso 2), obtenemos segundas aproximaciones para los parámetros y continuamos el ciclo hasta alcanzar la convergencia.

Para la asignación de las observaciones a las clases o grupos, si la población no es homogénea, podemos proceder calculando las probabilidades a posteriori y estableciendo la siguiente regla de clasificación:

$$h(j/x) = p_j \, f(x/j) / f(x) > h(h/x) = p_h \, f(x/h) / f(x) \qquad [12]$$

de donde,

$$p_j \, f(x/j) / p_h \, f(x/h) > 1$$

tomando logaritmos tenemos:

$$\log p_j - \log p_h + \log [f(x/j) - f(x/jh)] > 0 \qquad [13]$$

Si al sustituir los estimadores en la ecuación anterior [13], obtenemos un valor mayor que cero, entonces la observación x se le asignará a la clase j; en caso contrario a la clase h.

APLICACIÓN PRACTICA.-

Sea la población univariante de los ingresos netos totales de los hogares españoles en 1994, de la cual hemos obtenido una muestra aleatoria de tamaño 6435, cuyas características principales son:

$$\bar{x} = 2.326.147´581$$

y

$$S^2 = 2777986078640$$

El contraste de la normalidad de la población a través de la muestra nos indica un coeficiente de verosimilitud (-2logL=202642´1681) lo que nos permite aceptar la hipótesis de normalidad.

La estimación de los parámetros de la función de f(x) dada en [1], a través del algoritmo EM nos da los siguientes resultados:

$p_1 = 0´504$ $p_2 = 1 - p_1 = 0´496$

A=1 $\mu_1 = 2337042´001$ $\sigma^2 = 2708236211428$

A=2 $\mu_2 = 2308413´743$

Estas estimaciones nos indican que existen dos subpoblaciones, repartidas aproximadamente al 50% y con igual varianza.

Mediante el algoritmo EM podemos realizar una asignación de los elementos de la muestra a las clases, a través de la regla de Bayes de la distribución a posteriori, obteniéndose una clasificación en dos clases con las siguientes características muestrales:

A=1	media= 5053861´081	varianza = 3370791919168	n=1.156
A=2	media= 1728830´512	varianza= 662075428512	m=5.279

Realizando un contraste de hipótesis entre el modelo de homogeneidad e iguales medias contra el modelo de homogeneidad y distintas medias, se acepta la hipótesis del segundo modelo.

CONCLUSIONES:

Del trabajo se obtienen las siguientes conclusiones:

a) Que el análisis de mixtura de distribuciones nos permite discernir si una población es homogénea o heterogénea.

b) Que el algoritmo EM es una herramienta eficaz para la estimación de los parámetros de la distribución de una mixtura de distribuciones.

c) Que la asignación bayesiana de asignación de clases nos permite realizar y comparar las distintas subpoblaciones existentes y no cometer errores de interpretación de la variable en estudio.

d) Que los ingresos netos de los hogares españoles en 1994 están distribuidos en dos subpoblaciones, perfectamente diferenciadas.

BIBLIOGRAFIA

(1).- Peña, Daniel y Romo, Juan(1997). Introducción a la Estadística para las Ciencias Sociales. McGraw-Hill.

(2).- Simpson, C.H. (1951). The interpretation of interaction in contingency tables, J.R.Stat. Soc. B 13: 238-41.

(3).- Lauritzen,S.L. and Vermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann. Stat: 17:31-57.

(4).- Edwards, D. (1990). Hierarchical interaction models (with discussion). J.R. Stat. Soc. B 52:3-20.

(5).- Whittaker, J. (1990). Graphical Models in applied Multivariate Statistics, Wiley.

(6).- Edwards, D. (1995). Graphical modelling. In Krzanowski, W.J. (ed) Recent Advances in Descriptive Multivariate Analysis. Oxford University Press, Oxford, 127-148.