Análisis estadístico de las preferencias de jerarquización de varias opciones*

Robert J. Flowers †
Universidad Juárez Autónoma de Tabasco, DACB

María de Lourdes Navarrete Martínez
CAM de la SE

Para determinar las preferencias entre un grupo de opciones, una muestra de personas tiene que jerarquizar cada una de las opciones. Usando estas jerarquizaciones se define una media para cada opción. Se supone una distribución multinomial para el número de personas quienes escogieron cada jerarquización. Se define un algoritmo de máxima verosimilitud para probar si hay una diferencia significativa entre éstas medias, y así establecer si existe una preferencia significativa para una de las opciones.

A group of persons is asked to rank four options according to their preferences. It is necessary to determine if there is a clear difference in preferences. It is assumed that the number of persons choosing each ranking follows a multinomial distribution. Means are defined for each option based upon the ranks assigned. A maximum likelihood test is defined for determining if there is a significant difference in the means for each option.

Palabras clave: Distribución multinomial, Estimadores de máxima verosimilitud, Análisis de preferencias de jerarquización.
Keywords: Multinomial distribution, Maximum likelihood estimation, Rank choice analysis.

1. Introducción

En este artículo se considera el problema de analizar datos donde n personas han ordenado h opciones según sus preferencias. Para cada opción, se puede definir una media basado en el ordenamiento dado por los jueces. Forthofer y Lehnen [4] explicaron cómo analizar datos de preferencia usando el método de Grizzle [6]. Este procedimiento usa cuadrados mínimos ponderados para obtener los estimadores para los coeficientes de regresión. La variable dependiente se define por una transformación del vector de proporciones observados. Para el caso de los modelos lineales se puede obtener modelos de la misma forma usando un algoritmo de máxima verosimilitud como el presentado en [1, 3] o un algoritmo de ji-cuadrada mímima como en [2]. En los dos procedimientos de Flowers se usa un algoritmo iterativo de cuadrados mínimos ponderados para obtener los estimadores. Los dos procedimientos de Flowers y el de Grizzle, Starmer y Koch dan estimadores que son óptimos asintóticamente normales. En el problema estudiado aquí, las h medias son dependientes. Esto es debido a que la suma de las medias es igual a $(h + 1)/2$. Por esta razón, en este artículo recomendamos un análisis de medias donde primero se prueba si todas las medias son

*Recibido el 12 de enero de 2007 y aceptado el 6 de marzo de 2007
†**Dirección postal**: Carr. Cunduacán-Jalpa Km 1, Cunduacán Tabasco, México. A.P. 24 C.P. 86690. Tel.(+52)914 336-0928. **Correo electrónico**: robert.flowers@basicas.ujat.mx

iguales. Si se rechaza la hipótesis nula de que todas las medias son iguales, entonces se hace todas las comparaciones de pares de medias. Para controlar el error de tipo I, se puede usar un nivel de significancia de para el valor crítico donde k es el número de pruebas. Para hacer las pruebas de hipótesis usaremos modelos de la forma $Cm = r$.

2. Metodología

En ésta sección definiremos una clase de modelos de la forma $Cm = r$, donde m es el vector de medias, C es una matriz de transformaciones, y r es un vector de restricciones. Para este modelo, se hará el desarrollo suponiendo una distribución multinomial para los elementos de un vector. Gokhale y Kullback [5] usaron modelos de esta forma, pero ellos obtuvieron los estimadores de información discriminante mínima, mientras en este artículo usamos el método de máxima versimilitud.

Obsérvese que, si Y_i sigue una distribución de Poisson con media m_j, $j = 1, 2, ..., n$, entonces la distribución condicional de $Y_1, Y_2, \ldots, Y_n$ dada tiene una distribución multinomial. Se consigue ésta restricción al poner todos los elementos de la primera fila de C igual a 1 y el primer elemento del vector r igual a N. Las demás filas de C y r se definen en tal forma para definir las pruebas de interés.

Se pueden obtener los estimadores de máxima verosimilitud al maximizar $y' \ln(m) - \iota' m$ sujeta a la condición que $Cm = r$ donde $y' = (y'_1, y'_2, \ldots, y'_k)$, $m' = (m'_1, m'_2, \ldots, m'_k)$, e ι es un vector $n \times 1$ unidades.

Los estimadores de máxima verosimilitud se obtienen al maximizar la siguiente función

$$L(m) = y' \ln(m) - \iota' m - t'(Cm - r).$$

Nótese que, $y' \ln(m) - \iota' m$ es el núcleo de la función de verosimilitud y t es un vector de multiplicadores de Lagrange; éste último se incluye para obtener las restricciones $Cm = r$.

Derivando la función de verosimilitud con respecto a m, se obtiene

$$D_m^{-1}(y - m) - C't = 0,$$

donde D_m es una matriz diagonal compuesto de los elementos de m. Ahora, se multiplica en el lado izquierdo por D_m y se obtiene la ecuación

$$(y - m) - D_m C't = 0.$$

Multiplicando ésta ecuación en el lado izquierdo por C resulta el sistema de ecuaciones

$$Cy - Cm = CD_m C't.$$

Ya que $Cm = r$, se obtiene

$$Cy - r = CD_m C't.$$

Entonces

$$t = (CD_m C')^{-1}(Cy - r)$$

Ahora,

$$m = y - D_m C' t.$$

Esto nos permite estimar m usando

$$m = y - D_m C'(C D_m C')^{-1}(Cy - r).$$

Como los elementos del vector m tienen valores desconocidos y por lo tanto D_m es una matriz con valores desconocidos, es necesario usar un procedimiento iterativo para estimar m. Para hacer esto hay que definir valores iniciales para los elementos de m. Si se hace $m^{(0)} = y$, entonces $D_m^{(0)} = D_y$. Ahora D_m está definida y se puede estimar m usando

$$m^{(s+1)} = y - D_m^{(s)} C'(C D_m^s C')^{-1}(Cy - r)$$

donde s es el número de la iteración. En el caso de que algunos de los elementos del vector y sean ceros, se puede usar los valores iniciales

$$m^{(0)} = y + \frac{1}{2}\iota.$$

Este proceso continua hasta que los cambios en los elementos de m sean suficientemente pequeños. En el apéndice se da un programa para este procedimiento escrito en `Matlab`. El algoritmo presentado en esta sección puede ser usado para definir varios tipos de modelos como los modelos de simetría y homogeneidad marginal. Aquí, solamente nos interesa obtener estimadores para comparar medias definidos para indicar las preferencias de un grupo de personas a ciertas opciones. Aquí, supondremos que hay solamente una población de interés, pero el procedimiento es suficiente general para poder analizar datos con más de una población.

El estadístico ji-cuadrado de la razón de verosimilitud dado por

$$G^2 = -2\sum_{i=1}^{n} Y_i \ln(\frac{\hat{m}_i}{Y_i}),$$

se puede usar para la prueba de bondad de ajuste del modelo. El número de grados de libertad para esta prueba, es igual al número de filas en la matriz C que son usados para definir los contrastes.

3. Análisis de los datos del DIF

En el Sistema para el Desarrollo Integral de la Familia (DIF) existe la Coordinación general de los programas de nutrición; entre sus funciones se encuentra coordinar el programa alimentario en los 17 municipios del estado. El programa incluye la acción desayunos calientes que busca mejorar la nutrición de los niños menores de 12 años, la acción es implementada por asesores comunitarios capacitando a las madres en cuestiones de nutrición, hábitos de alimentación, higiene y salud.

Los responsables de la Coordinación saben que el entrenamiento de los asesores comunitarios es pieza clave para el éxito del programa, por ello se aplicó una encuesta

de opinión a los asesores que actúan en los 17 municipios. La encuesta estaba conformada por un solo reactivo –ver encuesta en el recuadro – en el cual se solicita la jerarquización de cuatro opciones de capacitación

> Según su opinión y experiencia de trabajo en las comunidades cómo ordenaría estas cuatro opciones de capacitación? para obtener mejores resultados como promotor.
>
> Primera opción Capacitación en el desarrollo de la comunidad: fomento de huertos familiares, integración y funciones de un comité, promoción.
> ( )
> Segunda opción. Capacitación general: estrategias para la administración de cuotas, elaboración de informes y llenado de formas, almacenamiento y manejo de insumos, ...
> ( )
> Tercera opción. Capacitación personal: redacción y ortografía, conocimiento de si mismo, autoestima y toma de decisiones,...
> ( )
> Cuarta opción. Capacitación en nutrición y medida antropométrica: adiestramiento materno, demostración culinaria, control de peso-talla,....
> ( )
>
> Instrucciones: escribe en el paréntesis el número 4 PARA LA CAPACITACIÓN DE MAYOR IMPORTANCIA, después escribe el número 3 para la que sigue en menor grado de importancia, escribe el número 2 para la que sigue y por último escribe el número 1 para la capacitación de menor trascendencia

Se muestran los datos de la encuesta en la tabla 1. Para evaluar las preferencias se puede definir una media para cada opción. Para hacer esto, primero se divide los valores en cada columna por el tamaño de la muestra para obtener los valores c_{1i}, c_{2i}, c_{3i} y c_{4i} . Por ejemplo, para la opción 3, $c_{31} = 3/400, c_{32} = 4/400, \ldots, c_{3,24} = 2/400$. Se puede definir la media para la opción i-ésima por $\bar{y}_i = \sum_{j=1}^{24} c_{ij} y_j$ donde las y_i son las frecuencias observadas mostradas en la tabla 1. Sea μ_i la media esperada para la opción i-ésima. Entonces $\bar{y}_i$ es un estimador de $\mu - i$.

La primera hipótesis que se quiere probar es si todas las medias son iguales. Por la manera de definir las medias, las cuatro medias deben sumar a 10. Por lo tanto, se puede expresar la primera hipótesis como $H_0 : \mu_1 = \mu_2 = \mu_3 = 2.5$. Para esta prueba, la matrices C y r mencionadas en la sección anterior se definen como

$$C = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ c_{11} & c_{12} & \ldots & c_{1,24} \\ c_{21} & c_{22} & \ldots & c_{2,24} \\ c_{31} & c_{32} & \ldots & c_{3,24} \end{bmatrix} \text{ y } r = \begin{bmatrix} 400 \\ 2.5 \\ 2.5 \\ 2.5 \end{bmatrix}$$

Entonces

$$Cy = \begin{bmatrix} 400 \\ \bar{y_1} \\ \bar{y_2} \\ \bar{y_3} \end{bmatrix}.$$

Opción 1	Opción 2	Opción 3	Opción 4	Frecuencia
1	2	3	4	23
1	2	4	3	13
1	3	2	4	25
1	3	4	2	15
1	4	2	3	5
1	4	3	2	11
2	1	3	4	31
2	1	4	3	22
2	3	1	4	36
2	3	4	1	18
2	4	1	3	9
2	4	3	1	3
3	1	2	4	31
3	1	4	2	9
3	2	1	4	32
3	2	4	1	15
3	4	1	2	10
3	4	2	1	7
4	1	2	3	16
4	1	3	2	12
4	2	1	3	18
4	2	3	1	11
4	3	1	2	22
4	3	2	1	6

Tabla 1. Frecuencias de las preferencias personales

La primera columna de C garantiza que la suma de los valores esperados es igual a la suma de los valores esperados. Para ésta prueba, $G^2 = 66.509$ con 3 grados de libertad. En este artículo, se usa un nivel de significancia de $\alpha = 0.05$. Entonces el valor de G^2 es altamente significativo y se puede concluir que hay una diferencia significativa entre las medias. Las medias observadas son $\bar{y}_1 = 2.4550, \bar{y}_2 = 2.2275, \bar{y}_3 = 2.3700$ y $\bar{y}_4 = 2.9475$. Es importante ahora determinar cuales medias difieren significativamente. Entonces se hace todas las comparaciones entre pares de medias. Ya que hay 6 pruebas hacer, usaremos el valor crítico $\chi^2_{1-\alpha/6} = 6.960$ para cada una de las siguientes pruebas. Para probar la hipótesis $H_0 : \mu_1 = \mu_2$, se puede definir C y r como sigue:

$$C = \begin{bmatrix} 1 & 1 & \dots & 1 \\ c_{11} - c_{21} & c_{12} - c_{22} & \dots & c_{1,24} - c_{2,24} \end{bmatrix} \text{ y } r = \begin{bmatrix} 400 \\ 0 \end{bmatrix}.$$

Entonces para esta prueba, $Cy = \begin{bmatrix} 400 \\ \bar{y}_1 - \bar{y}_2 \end{bmatrix}$. Para esta prueba, $G^2 - 7.475$ con un grado de libertad. Entonces se puede concluir que hay una diferencia significativa entre μ_1 y μ_2. Las demás pruebas se realizaron de manera análoga. Se presentan los resultados de estas pruebas en la tabla 2.

Los resultados de la tabla 2 indican que no hay diferencia significativa entre las preferencias para las opciones 1 y 3, tampoco hay entre las opciones 2 y 3.

H_0	G^2
$\mu_1 = \mu_2$	7.475
$\mu_1 = \mu_3$	0.852
$\mu_1 = \mu_4$	29.506
$\mu_2 = \mu_3$	2.665
$\mu_2 = \mu_4$	63.697
$\mu_3 = \mu_4$	34.411

Tabla 2. Comparaciones entre pares de medias

4. Conclusión

Siendo $\bar{y}_1 = 2.4550, \bar{y}_2 = 2.2275, \bar{y}_3 = 2.3700$, y $\bar{y}_4 = 2.9475$, entonces la cuarta opción "capacitación en nutrición y medida antropométrica" ocupa, en promedio, el primer lugar en los intereses de capacitación por parte de los asesores comunitarios. En la jerarquización siguen la primera y la tercera opción, simultáneamente, ya que no difieren significativamente, o sea, "capacitación en el desarrollo de la comunidad" y "capacitación personal" ocupan el segundo lugar en el interés de los asesores.

En el último lugar se encuentra la segunda y la tercera opción, ya que no difieren significativamente, esto quiere decir que "capacitación en el desarrollo de la comunidad" y "capacitación general" son la menor necesidad sentida por los asesores en cuestiones de adiestramiento. Se puede hacer el análisis como en Forthofer y Lehnen (1981) , pero el procedimiento presentado aquí es más directo. Los dos procedimientos deber ser asintoticamente equivalentes.

Apéndice

```
%Este programa calcula los valores de M, Gcuad de un modelo de la
%forma Cm=R. Hay que dar las matrices Y, C, y R desde la
%pagina inicial de matlab, ahi se llama o se corre el programa.

tol=0.00001;
k=length(Y);
i=ones(k,1);
M0=Y+(0.5*i); %Valor incial para M
Dm=diag(M0);
norma=tol+1;
   while norma>tol;
       V=C*Dm*C';
       Vinv=inv(V);
       T=Vinv*(C*Y-R);
       M=Y-Dm*C'*T; %Estimacion de M
       norma=norm(M-M0);
       M0=M;
       Dm=diag(M0);
   end;
Gcuad=0;
```

```
for j=1:k
    if Y(j)~=0
        Gcuad=Gcuad-2*Y(j)*log(M(j)/Y(j));
    else Gcuad=Gcuad;
    end;
end;
fprintf(1,'la martriz M de los estimadores es: ',M);
%Da la matriz de los estimadores%
M
fprintf(1,'el valor de G-cuad es: ',Gcuad);
%Da el valor de ji-cuadrada de la razon de verosimilitud%
```

Referencias

[1] Flowers, R.J. *A*nalysis of discrete data using linear regression. Universidad y Ciencia, 1,**2** (1984), pp 75-85.

[2] Flowers, R.J. *E*stimadores de ji-cuadrada minima para modelos lineales. Universidad y Ciencia, 4,**7**(1987), pp 83-90.

[3] Flowers, R.J. *P*ruebas estadística para homogeneidad marginal o simetría. Revista de Ciencias Básicas UJAT, 3,**1**(2000), pp 29-44.

[4] Forthofer, R.N. y R.G. Lehnen. *P*ublic program analysis: A new categorical data approach. Belmont: Lifetime Learning Publications. 1981

[5] Gokhale, D.V. y Kullback, S. *T*he information in contingency tables. New York: Marcel Dekker, Inc. 1978

[6] Grizzle, J.E., C.F. Starmer, y G.G. Koch. *A*nalysis of categorical data by linear models. Biometrics, **25**(1969), pp 489-504.