CAPÍTULO IV

4. ANÁLISIS DE CONTINGENCIA Y ANÁLISIS DE CORRESPONDENCIA: TEORÍA.

4.1 Análisis de Tablas de Contingencia.

El análisis de tablas de contingencia es una aplicación del análisis de tablas r X c. La hipótesis nula que se desea probar por medio de las tablas de contingencia es que una variable X es independiente de una variable Y. En general, si θ_{ij} es la probabilidad de que un elemento caerá en la celda que pertenece al iésimo renglón y la jésima columna, $\theta_{i.}$ es la probabilidad de que un elemento caerá en el iésimo renglón, y $\theta_{.j}$ es la probabilidad de que un elemento caerá en la jésima columna, la hipótesis nula que queremos probar es

$$\theta_{ij} = \theta_{i.} \cdot \theta_{.j}$$

para i =1,2,...,r y j =1,2,...c. Correspondientemente, la hipótesis alternativa es $\theta_{ij} \neq \theta_{i.}.\theta_{.j}$

para al menos un par de valores de i y j.

Denotaremos la frecuencia observada en el iésimo renglón y la jésima columna con f_{ij}, los totales de los renglones con $f_{i.}$, los totales de las columnas con $f_{.j}$, y el gran total, la suma de todas las frecuencias de las celdas, con f. Con esta notación, estimamos las probabilidades $\theta_{i.}$ y $\theta_{.j}$ como:

$$\hat{\theta}_{i.} = \frac{f_{i.}}{f} \qquad \text{y} \qquad \hat{\theta}_{.j} = \frac{f_{.j}}{f}$$

Y bajo la hipótesis nula de independencia se obtiene:

$$e_{ij} = \hat{\theta}_{i.} \cdot \hat{\theta}_{.j} \cdot f = \frac{f_{i.}}{f} \cdot \frac{f_{.j}}{f} \cdot f = \frac{f_{i.} \cdot f_{.j}}{f}$$

para la frecuencia esperada para la celda en el iésimo renglón y la jésima columna. Advierta que e_{ij} así obtenida al *multiplicar el total del renglón al cal pertenece la celda por el total de la columna a la cual pertenece y después dividir entre el gran total.*

Una vez que se ha calculado la e_{ij}, basamos nuestra decisión en el valor de

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

y rechazamos la hipótesis nula si excede a $\chi^2_{\alpha,(r-1)(c-1)}$.

El número de grados de libertad es (*r*-1)(*c*-1), y en relación con esto hagamos la siguiente observación: siempre que se estimen frecuencias de celdas en fórmulas de ji cuadrada con base en datos de conteo muestrales, el número de grados de libertad es *s*-*t*-1, donde *s* es el número de términos en la suma y *t* es el número de parámetros independientes reemplazados por estimadores. Cuando se hace la prueba para independencia en una tabla de contingencia *r* X *c* se tiene que *s* = *rc* y *t* = *r*-*c*-2, puesto que los *r* parámetros $\theta_{i.}$ y los *c* parámetros $\theta_{.j}$, no son todos independientes; sus sumas respectivas deben ser igual a 1. Así se obtiene *s*-*t*-1 = *rc* - (*r* - *c* - 2) -1 = (*r*-1)(*c*-1).

Puesto que la estadística de prueba que se ha descrito sólo tiene aproximadamente una distribución ji cuadrada con (*r*-1)(*c*-1) grados de libertad, es costumbre usar esta prueba sólo cuando ninguna de las e_{ij} es menor que 5; esto algunas veces requiere que se combinen algunas de las celdas con una pérdida correspondiente en el número de grados de libertad.

4.2. Análisis de Correspondencias.

4.2.1. Análisis de Correspondencias Simple.

Desarrollado por los franceses, el análisis de correspondencia es un procedimiento gráfico para representar asociaciones en una tabla de frecuencias o conteo. El análisis de correspondencia

simple se concentra en una tabla de frecuencia de dos vías o tabla de contingencia. Si la tabla de contingencia tiene *r* renglones y *c* columnas, el gráfico de puntos producido por el análisis de correspondencias simple contiene dos conjuntos de puntos: un conjunto de *r* puntos correspondiente a los renglones y un conjunto de *c* puntos correspondiente a las columnas. La posición de los puntos refleja las asociaciones.

Los puntos renglones que están muy próximos indican que los renglones tienen perfiles similares (distribuciones condicionales) a lo largo de las columnas. Los puntos columna que se encuentran próximos indican columnas con perfiles similares (distribuciones condicionales) por filas. Finalmente, puntos fila que están muy cercanos a puntos columna representan combinaciones que ocurren con mayor frecuencia que si se esperara un modelo de independencia, es decir, un modelo en el cual las categorías indicadas en las filas no están relacionadas con las categorías expresadas en las columnas.

El resultado usual de un análisis de correspondencia simple incluye la “mejor” representación bidimensional de los datos y una medida (denominada *inercia*) de la cantidad de información retenida en cada dimensión.

Desarrollo algebraico del análisis de correspondencia simple.

Supóngase que **X**, con elementos x_{ij}, es una tabla de contingencia I X J. En esta discusión tomemos I<J y asumamos que X es de rango columna completo J. Las filas y las columnas de la tabla de contingencia **X** corresponden a diferentes categorías de dos diferentes características.

Es conveniente basar la representación gráfica de asociación en una tabla de contingencia, en una matriz sutilmente centrada y escalada. Si *n* es el total de las frecuencias en **X**, primero se construirá una matriz de proporciones **P** = { p_{ij} } dividiendo cada elemento de **X** por *n*. Entonces:

$$p_{ij} = \frac{x_{ij}}{n}, \; i = 1,2,\ldots, I, \; j = 1,2,\ldots, J, \quad \text{o} \quad \underset{(IXJ)}{P} = \frac{1}{n} \underset{(IXJ)}{X} \tag{4.1}$$

La matriz P se denomina matriz de correspondencia. Luego, P es centrada mediante la sustracción del producto entre el total de los renglones y el total de las columnas para cada entrada. Esta operación produce:

$$\tilde{p}_{ij} = p_{ij} - r_i c_j, \; i = 1,2,\ldots, I, \quad j = 1,2,\ldots, J, \quad \text{o} \quad \tilde{\mathrm{P}} = \mathrm{P} - \mathrm{rc}' \tag{4.2}$$

Donde

$$r_{ij} = \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n}, \quad i = 1,2,\ldots I, \qquad o \qquad \underset{(IX1)}{\mathbf{r}} = \underset{(IXJ)}{\mathbf{P}} \quad \underset{(IXJ)}{\mathbf{1}}$$

(4.3)

$$c_{ij} = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n}, \quad j = 1,2,\ldots J, \qquad o \qquad \underset{\mathbf{(IX1)}}{\mathbf{c}} = \underset{(IXJ)}{\mathbf{P}'} \quad \underset{(IXJ)}{\mathbf{1}}$$

y 1'=[1,1,…1]. Notemos que rango ($\tilde{\mathbf{P}}$) ≤ J – 1 partiendo de que $\tilde{\mathbf{P}}\mathbf{1} = \mathbf{P1} - \mathbf{rc'1} = \mathbf{r} - \mathbf{r} = \mathbf{0}$.

Definamos las matrices diagonales

$\mathbf{D_r}$= diag (r_1, r_2,…, r_I) y $\mathbf{D_c}$ = (c_1, c_2,…, c_J)

(4.4)

y construyamos la matriz escalada

$$\underset{\mathbf{(IxJ)}}{\mathbf{P}^*} = \underset{\mathbf{(IXI)}}{\mathbf{D}_\mathbf{r}^{-1/2}} \quad \underset{\mathbf{(IXJ)}}{\tilde{\mathbf{P}}} \quad \underset{\mathbf{(JXJ)}}{\mathbf{D}_\mathbf{C}^{-1/2}}$$

(4.5)

de tal manera que la celda (*i, j*) ésima de **P*** es

$$p_{ij}^* = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}, \quad i = 1,2,\ldots I, \; j = 1,2,\ldots J$$

(4.6)

A continuación se presentan los pasos que guían a un gráfico de asociación en una tabla de dos vías.

Paso 1. Encontrar la descomposición de valores singulares de **P*.** Se tiene:

$$\underset{(IXJ)}{\mathbf{P}^{*}} = \underset{IX(J-1)}{\mathbf{U}} \quad \underset{(J-1)X(J-1)}{\mathbf{\Lambda}} \quad \underset{(J-1)XJ}{\mathbf{V}'}$$

(4.7)

donde rango ($\tilde{\mathbf{P}}$) ≤ J – 1,

$$\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$$

y la matriz diagonal $\mathbf{\Lambda}$ = diag(λ_1, λ_2,…, λ_{J-1}) contiene los valores singulares, ordenados del más grande al más pequeño, a lo largo de la diagonal principal.

Paso 2. Definir $\tilde{\mathbf{U}} = \mathbf{D}_r^{1/2}\mathbf{U}$ y $\tilde{\mathbf{V}} = \mathbf{D}_C^{1/2}\mathbf{V}$, luego, usando (4.5) y (4.7) la descomposición de valores singulares de $\tilde{\mathbf{P}}$ es:

$$\tilde{\mathbf{P}} = \mathbf{P} - \mathbf{r}\mathbf{c}' = \tilde{\mathbf{U}}\mathbf{\Lambda}\tilde{\mathbf{V}}' = \sum_{j=1}^{J-1} \lambda_j \tilde{\mathbf{u}}_j \tilde{\mathbf{v}}_j'$$

(4.8)

Donde $\tilde{\mathbf{u}}_j$ es el j-ésimo vector columna de $\tilde{\mathbf{U}}$ y $\tilde{\mathbf{v}}_j$ es el j-ésimo vector columna de $\tilde{\mathbf{V}}$. En esta representación, los vectores singulares izquierdo y derecho están normalizados para tener longitud unitaria en las métricas $\mathbf{D}_r^{-1}$ y $\mathbf{D}_c^{-1}$, respectivamente. Es decir:

$$\underset{((J-1)XI)}{\tilde{\mathbf{U}}'} \underset{(IXI)}{\mathbf{D}_r^{-1}} \underset{(IX(J-1)}{\tilde{\mathbf{U}}} = \underset{((J-1)XJ)}{\tilde{\mathbf{V}}'} \underset{(JXJ)}{\mathbf{D}_C^{-1}} \underset{(JX(J-1))}{\tilde{\mathbf{V}}} = \underset{((J-1)X(J-1))}{\mathbf{I}}$$

(4.9)

Las columnas de $\tilde{\mathbf{U}}$ definen las coordenadas para los puntos que representan los perfiles columna de **P**. Similarmente, las columnas

de $\tilde{\mathbf{V}}$ definen las coordenadas para los puntos que representan los perfiles fila de **P**.

Paso 3. Calcular las coordenadas de los perfiles fila.

$$\underset{(IX(J-1))}{\mathbf{Y}} = \underset{(IXI)}{\mathbf{D}_r^{-1}} \underset{(IX(J-1))}{\tilde{\mathbf{U}}} \underset{((J-1)X(J-1))}{\boldsymbol{\Lambda}}$$

(4.10)

y las coordenadas de los perfiles columna.

$$\underset{(IX(J-1))}{\mathbf{Z}} = \underset{(JXJ)}{\mathbf{D}_C^{-1}} \underset{(JX(J-1))}{\tilde{\mathbf{V}}} \underset{((J-1)X(J-1))}{\boldsymbol{\Lambda}}$$

(4.11)

Las primeras columnas de **Y** contienen los pares de coordenadas de los puntos fila en la mejor representación bidimensional de los datos. Las primeras dos columnas de **Z** contienen los pares de coordenadas de los puntos columna en la mejor representación en dos dimensiones de los datos. Los puntos correspondientes a estos dos conjuntos de coordenadas pueden ser superpuestos en el mismo gráfico. Para un conjunto de puntos fila, o para un conjunto de puntos columna, la distancia Euclidiana en el gráfico de dos dimensiones corresponde a la distancia estadística entre pares de perfiles columna (fila) en los datos originales. Es importante recordar que no existe relación de distancia *directa*

entre un punto que representa a un perfil fila y otro punto que representa a un perfil columna.

Paso 4. La Inercia es el cuadrado de los valores singulares correspondiente a cada dimensión. La inercia total se define como la suma de los cuadrados de todos los valores singulares diferentes de cero.

$$Inercia\ Total = \sum_{i=1}^{K} \lambda_i^2$$

(4.12)

donde $\lambda_1 \geq \lambda_2 \geq \ldots, \lambda_K > 0$ son los elementos de la diagonal de $\mathbf{\Lambda}$ diferentes de cero. Aquí, *K* = rango ($\tilde{\mathbf{P}}$) y, ordinariamente, rango ($\tilde{\mathbf{P}}$) = min (*I* -1,*J* -1).

Análisis de Asociación Ji cuadrado y el Análisis de Correspondencia.

El estadístico χ^2 para medir el grado de asociación entre las variables fila y columna en una tabla de contingencia de dos vías con *I* filas y *J* columnas es

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

(4.13)

Donde $f_{ij} = x_{ij}$ es la frecuencia observada para la *(i,j)* ésima celda y $e_{ij} = nr_ic_j$ es la frecuencia esperada en la *(i,j)* ésima celda si la variable fila es independiente de la variable columna.

Después de una pequeña manipulación, y usando (4.6), se puede escribir:

$$\chi^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_ic_j)^2}{r_ic_j} = n\sum_{i,j} p_{ij}^{*2}$$

(4.14)

En notación matricial,

$$\frac{\chi^2}{n} = traza(\mathbf{D_r^{-1}(P - rc')D_c^{-1}(P - rc')'}) = traza(\mathbf{P * P*'}) = \sum_{i,j} p_{ij}^{*2}$$

(4.15)

Inercia.

El *i*-ésimo perfil fila $\tilde{\mathbf{r}}_\mathbf{i}$ con el *j*-ésimo elemento

$$\tilde{r}_{ij} = \frac{x_{ij}/n}{\left(\sum_j x_{ij}\right)/n} = \frac{p_{ij}}{r_i}, \quad j = 1,2,\ldots J$$

es la *i*-ésima fila de **X** dividida para su suma. Por lo tanto, la matriz de perfiles fila está dada por:

$$\underset{(IXJ)}{\tilde{\mathbf{R}}} = \begin{bmatrix} \tilde{\mathbf{r}}_1 \\ \tilde{\mathbf{r}}_2 \\ \vdots \\ \vdots \\ \tilde{\mathbf{r}}_I \end{bmatrix} = \underset{(IXI)}{\mathbf{D}_r^{-1}} \underset{(IXJ)}{\mathbf{P}}$$

(4.16)

Similarmente, los perfiles columna $\tilde{\mathbf{c}}_j$, j =1,2,…, J, son las columnas de **X** dividida por sus sumas, de tal manera que el i-ésimo elemento de $\tilde{\mathbf{c}}_j$ es

$$\tilde{c}_{ij} = \frac{p_{ij}}{c_i}, \quad i = 1,2,\ldots I$$

En notación matricial

$$\underset{(JXI)}{\tilde{\mathbf{C}}} = \begin{bmatrix} \tilde{\mathbf{c}}_1 \\ \tilde{\mathbf{c}}_2 \\ \vdots \\ \vdots \\ \tilde{\mathbf{c}}_J \end{bmatrix} = \underset{(JXJ)}{\mathbf{D}_c^{-1}} \underset{(JXI)}{\mathbf{P}'}$$

(4.17)

Consideremos el promedio ponderado $\tilde{\mathbf{R}}'\mathbf{r}$ de los perfiles fila, o centroide fila.

Ahora, $\tilde{\mathbf{R}}'\mathbf{r} = \mathbf{P}'\mathbf{D}_r^{-1}\mathbf{r} = \mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{1}$ por (4.3). Luego $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{1} = \mathbf{1}$ debido a (4.16), $\mathbf{D}_r^{-1}\mathbf{P}$ tiene elementos p_{ij}/r_i . Finalmente **c=P'1** entonces

$$\mathbf{c} = \mathbf{P}'\mathbf{D}_r^{-1}\mathbf{r}$$

(4.18)

Similarmente, el promedio ponderado $\tilde{\mathbf{C}}'\mathbf{c}$ de los perfiles columna, o centroide columna, es

$$\mathbf{r} = \mathbf{P}'\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{c}$$

(4.19)

Ahora estamos en posición de definir la inercia.

La inercia total es la suma ponderada de las distancias cuadradas de los perfiles fila (o perfiles columna) hacia el centroide. Consecuentemente, es una medida de la variación total, o diferencias, en los puntos que representan los perfiles fila (o perfiles columna). La inercia asociada a los puntos fila es la misma que la inercia asociada a los puntos columna.

Usando la relación

$$\mathbf{P} - \mathbf{rc}' = \sum_{j=1}^{J-1} \lambda_j \tilde{\mathbf{u}}_{\mathbf{j}} \tilde{\mathbf{v}}_{\mathbf{j}}'$$

La escala (4.5), y las condiciones de ortogonalidad para los $\tilde{\mathbf{u}}'s$ y los $\tilde{\mathbf{v}}'s$

$$\text{Inercia} = \frac{\chi^2}{n} = traza\left(\mathbf{D}_{\mathbf{r}}^{-1}(\mathbf{P} - \mathbf{rc}')\mathbf{D}_{\mathbf{c}}^{-1}(\mathbf{P} - \mathbf{rc}')'\right) = \sum_{k=1}^{J-1} \lambda_k^2$$

4.2.2. Análisis de Correspondencias Múltiples.

Se aplica a tablas de contingencias en las que por filas se tienen n individuos y por columnas s variables categóricas con p_i i = 1,..., s categorías mutuamente excluyentes y exhaustivas.

La tabla de datos tiene, por lo tanto, la forma:

$$Z = [Z_1, Z_2 ,..., Z_s]$$

con Z_i matriz nxp_i de forma que

z_{ij} = 1 si el individuo i-ésimo ha elegido la modalidad j

z_{ij} = 1 si el individuo i-ésimo no ha elegido la modalidad j

con i=1,...,n y j=1,..., $p=p_1 + p_2 + ... + p_s$

El Análisis de Correspondencias Múltiples se basa en realizar un Análisis de Correspondencias sobre la llamada matriz de Burt:

$$\mathbf{B = Z'Z}$$

Dicha matriz se construye por superposición de cajas. En los bloques diagonales aparecen matrices diagonales conteniendo las frecuencias marginales de cada una de las variables analizadas. Fuera de la diagonal aparecen las tablas de frecuencias cruzadas correspondientes a todas las combinaciones 2 a 2 de las variables analizadas

Se toman como dimensiones aquellas cuya contribución a la inercia supera 1/p.

Distancias χ2

En este caso vienen dadas por las expresiones

$$d^2(j,j') = \sum_{i=1}^{n} n\left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}}\right)^2$$ Distancia entre modalidades

$$d^2(i,i') = \frac{1}{s}\sum_{j=1}^{p} \frac{n}{z_{.j}}\left(z_{ij} - z_{i'j}\right)^2$$ Distancia entre individuos