La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 4

Distancia ji-cuadrado e inercia

Primera edición: julio 2008
ISBN: 978-84-96515-71-0

Traducción: Jordi Comas Angelet
Revisión: Carles M. Cuadras Avellana

www.fbbva.es

CAPÍTULO 4

Distancia ji-cuadrado e inercia

En los capítulos 2 y 3 representamos gráficamente los perfiles, medimos de forma implícita las distancias entre ellos y luego interpretamos sus posiciones. En el AC las distancias entre los perfiles las medimos de forma algo más complicada, mediante la *distancia ji-cuadrado*. Esta distancia es la clave de muchas de las propiedades interesantes del AC. Existen varias maneras de justificar la utilización de la distancia ji-cuadrado. Algunas de ellas, más técnicas, quedan fuera del alcance de este libro, otras son más intuitivas. En este capítulo nos decantamos por estas últimas. Así, empezaremos por la interpretación geométrica del conocido *estadístico ji-cuadrado*, calculado a partir de los datos de una tabla de contingencia. Las ideas que hay detrás del estadístico ji-cuadrado nos llevan al concepto de distancia ji-cuadrado y al concepto de inercia. En el AC medimos la variabilidad de una tabla de datos mediante la inercia, un concepto muy relacionado con la distancia ji-cuadrado.

Contenido

La hipótesis de independencia, o de homogeneidad, en tablas de contingencia

Consideremos, otra vez, los datos de la imagen 3.1. Fijémonos en que, de los 312 individuos de la muestra, 57 (el 18,3%) están situados en el grupo de lectores *C1*, 129 (el 41,3%) en el de lectores *C2* y 126 (el 40,4%) en el de lectores *C3*; así pues, el perfil fila medio está constituido por las proporciones [0,183 0,413 0,404]. Si no existieran diferencias entre los niveles de educación, por lo que concierne al tipo de lectores, los perfiles de todas las filas deberían ser más o menos iguales al perfil fila medio. Las diferencias que observamos se deberían sólo a fluctuaciones del muestreo aleatorio. Si suponemos que no hay diferencias, o dicho de otra manera, si suponemos que los niveles de educación son *homogéneos* con relación al tipo de lectores, ¿cuáles serán, las frecuencias esperadas de la fila E5? El nivel de educación E5 consta de 26 individuos, por tanto esperaríamos que el 18,3% de éstos pertenecieran a la categoría *C1*; es decir, 26 × 0,183 = 4,76 (aunque no tenga sentido considerar 0,76 individuos, para estos cálculos es conveniente mantener los valores decimales). De la misma forma, esperaríamos que 26 × 0,413 = 10,74 individuos de E5, pertenecieran a la categoría *C2*, y 26 × 0,404 = 10,50 a la categoría *C3*. En estadística, el «supuesto de no diferencias» entre las filas de una tabla de contingencia (o de forma similar, entre las columnas) tiene distintas denominaciones, —«hipótesis de independencia» es una de ellas, o quizás, en este contexto, es más adecuada la denominación «supuesto (o asunción) de homogeneidad»—. Mediante el supuesto de homogeneidad, las frecuencias esperadas para la fila E5 serían [4,76 10,74 10,50], sin embargo, los valores observados son [3 7 16]. De forma similar, por el mismo supuesto de homogeneidad, podemos calcular las frecuencias esperadas de las restantes filas. En la imagen 4.1, mostramos, debajo de los valores observados de todas las filas, los correspondientes valores esperados. Llegaríamos exactamente a las mismas frecuencias esperadas si trabajáramos con los perfiles columna, es decir, suponiendo que los grupos de lectores son homogéneos.

Imagen 4.1: *Frecuencias observadas, tal como aparecen en la imagen 3.1 junto con las frecuencias esperadas (entre paréntesis) calculadas suponiendo que se cumple el supuesto de homogeneidad*

	Tipo de Lector				
Nivel de educación	*Rápidos C1*	*Minuciosos C2*	*Muy minuciosos C3*	*Total*	*Masas de las filas*
Educación primaria incompleta	5	7	2	*14*	*0,045*
E1	(2,56)	(5,78)	(5,66)		
Educación primaria	18	46	20	*84*	*0,269*
E2	(15,37)	(34,69)	(33,94)		
Educación secundaria incompleta	19	29	39	*87*	*0,279*
E3	(15,92)	(35,93)	(35,15)		
Educación secundaria	12	40	49	*101*	*0,324*
E4	(18,48)	(41,71)	(40,80)		
Educación universitaria incompleta	3	7	16	*26*	*0,083*
E5	(4,76)	(10,74)	(10,50)		
Total	*57*	*129*	*126*	*312*	
Perfil fila medio	*(0,183)*	*(0,413)*	*(0,404)*		

Contraste de la hipótesis de homogeneidad utilizando el estadístico ji-cuadrado

Las frecuencias observadas siempre serán distintas de las frecuencias esperadas. Sin embargo, en estadística queremos saber si estas diferencias son suficientemente grandes como para contradecir la hipótesis de que las filas son homogéneas. Es decir, queremos saber si es poco probable que las discrepancias entre las frecuencias observadas y las frecuencias esperadas se deban sólo al azar. Para responder a esta pregunta calcularemos una medida de discrepancia entre las frecuencias observadas y las frecuencias esperadas. Concretamente, calcularemos las diferencias entre cada par de frecuencias observadas y esperadas, las elevaremos al cuadrado, las dividiremos por las frecuencias esperadas e iremos acumulando los resultados hasta llegar a un valor final —el *estadístico ji-cuadrado*, que simbolizaremos por χ^2—:

$$\chi^2 = \sum \frac{(\text{observado} - \text{esperado})^2}{\text{esperado}}$$

Cálculo de χ^2

Dado que en una tabla de 5 por 3 (5 × 3) hay 15 células, el cálculo constará de 15 términos. En el siguiente cálculo mostramos solamente los tres primeros, correspondientes a la fila E1, y los tres últimos, correspondientes a la fila E5:

$$\chi^2 = \frac{(5-2{,}56)^2}{2{,}56} + \frac{(7-5{,}78)^2}{5{,}78} + \frac{(2-5{,}66)^2}{5{,}66} + \cdots + \frac{(3-4{,}76)^2}{4{,}76} + \frac{(7-10{,}74)^2}{10{,}74} + \frac{(16-10{,}50)^2}{10{,}50} \qquad (4.1)$$

En este cálculo, la suma de los 15 términos es igual a 26,0. Cuanto mayor sea este valor, mayores serán las discrepancias entre las frecuencias observadas y las frecuencias esperadas y, en consecuencia, estaremos menos convencidos de la certeza del supuesto de homogeneidad. Para valorar si 26,0 es grande o pequeño utilizamos las tablas de la distribución ji-cuadrado, con sus correspondientes «grados de libertad». Así, para una tabla de 5 × 3, los grados de libertad son 4 × 2 = 8 (el número de filas menos uno, multiplicado por el número de columnas menos uno). A un valor del estadístico χ^2 de 26,0, con 8 grados de libertad, el valor *p* asociado es de 0,001. Este resultado nos indica que la probabilidad de que las frecuencias observadas en la imagen 4.1 se correspondan con el supuesto de homogeneidad es extremadamente baja —una entre mil—. Es decir, rechazamos la homogeneidad de la tabla y concluimos que es muy probable que existan diferencias reales entre los niveles de educación, en lo concerniente a los perfiles de los tipos de lectura.

Una expresión alternativa del estadístico χ^2, en términos de perfiles y de masas

De hecho, estamos más interesados en la capacidad del χ^2 para medir la falta de homogeneidad, es decir, para medir la heterogeneidad entre los perfiles, que en la prueba estadística de homogeneidad que acabamos de describir. Vamos a expresar de otra forma el estadístico χ^2. Para ello dividiremos el numerador y el denominador de cada uno de los tres términos de cada fila por

el cuadrado del total de la fila. Por ejemplo, fijémonos en los tres últimos términos del cálculo del estadístico χ^2 que mostramos en (4.1); dividimos el numerador y el denominador de cada uno de estos tres términos por el cuadrado del total de la fila E5, es decir, por 26^2, de esta forma, en vez de tener las frecuencias absolutas originales, tenemos los perfiles observados y los perfiles esperados;

$$\begin{aligned}\chi^2 &= 12 \text{ términos similares} \cdots + \frac{\left(\frac{3}{26}-\frac{4{,}76}{26}\right)^2}{\frac{4{,}76}{26^2}} + \frac{\left(\frac{7}{26}-\frac{10{,}74}{26}\right)^2}{\frac{10{,}74}{26^2}} + \frac{\left(\frac{16}{26}-\frac{10{,}50}{26}\right)^2}{\frac{10{,}50}{26^2}} \\ &= 12 \text{ términos similares} \cdots + \\ &+ 26 \times \frac{(0{,}115-0{,}183)^2}{0{,}183} + 26 \times \frac{(0{,}269-0{,}413)^2}{0{,}413} + 26 \times \frac{(0{,}615-0{,}404)^2}{0{,}404} \end{aligned} \quad (4.2)$$

Fijémonos en que hemos eliminado uno de los 26 que aparecía dividiendo en el denominador de cada uno de los tres términos y que ahora aparece como un factor que multiplica a cada término. De esta manera hemos conseguido expresar los términos como perfiles. Así, los 15 términos los calcularíamos de la manera siguiente:

$$\text{total de la fila} \times \frac{(\text{perfiles observados de la fila} - \text{perfiles esperados de la fila})^2}{\text{perfiles esperados de la fila}}$$

La inercia (total) es igual al estadístico χ^2 dividido por el tamaño de la muestra

Hagamos una modificación más en el cálculo del estadístico χ^2 que mostramos anteriormente, con el fin de ponerlo en sintonía con los conceptos que hasta ahora hemos visto en el AC. Dividamos los dos lados de la ecuación (4.2) por el tamaño total de la muestra, de manera que en cada término de la derecha de la ecuación aparezca en primer lugar un factor que, en vez de corresponder al total de la fila, corresponda a la masa de la misma.

$$\begin{aligned}\frac{\chi^2}{312} &= 12 \text{ términos similares} \cdots + \\ &+ 0{,}083 \times \frac{(0{,}115-0{,}183)^2}{0{,}183} + 0{,}083 \times \frac{(0{,}269-0{,}413)^2}{0{,}413} + 0{,}083 \times \frac{(0{,}615-0{,}404)^2}{0{,}404} \end{aligned} \quad (4.3)$$

donde 0,083 = 26/312 es la masa de la fila E5 (imagen 4.1). En el AC, llamamos *inercia total*, o simplemente *inercia*, al valor χ^2/n de la izquierda, donde n es el total de la tabla. Este valor es una medida de la varianza total de la tabla independiente de su tamaño. En estadística, este valor recibe diferentes nombres, uno de ellos es el de «coeficiente medio cuadrático de contingencia». A su raíz cuadrada la denominamos «coeficiente phi» (ϕ); por tanto, podemos expresar la inercia como ϕ^2. Si en la expresión (4.3) agrupamos los tres términos de cada fila, obtenemos la siguiente expresión para la inercia:

$$\frac{\chi^2}{312} = \phi^2 = 4 \text{ grupos similares de términos} \cdots + \\ + 0{,}083 \times \left[\frac{(0{,}115 - 0{,}183)^2}{0{,}183} + \frac{(0{,}269 - 0{,}413)^2}{0{,}413} + \frac{(0{,}615 - 0{,}404)^2}{0{,}404} \right] \qquad (4.4)$$

Ahora, los cinco grupos de términos de esta fórmula, uno de cada fila de la tabla, son iguales a la masa correspondiente de su fila (por ejemplo 0,083 para la fila E5), multiplicada por un valor al cuadrado, entre paréntesis, que tiene el aspecto de una distancia (para ser más precisos, el cuadrado de una distancia).

La distancia euclídea o pitagórica

En la expresión (4.4) que acabamos de ver, si no fuera por el hecho de que dividimos el cuadrado de las diferencias entre los elementos observados y los esperados del perfil por los elementos esperados, el valor entre los corchetes, sería exactamente el cuadrado de la distancia «directa» entre el perfil fila E5 y el perfil fila medio en un espacio físico tridimensional, es decir la *distancia euclídea* o *pitagórica*. Para comprenderlo mejor, vamos a verlo de otra manera, supongamos que representamos gráficamente los dos perfiles [0,115 0,269 0,615] y [0,183 0,413 0,404] con respecto a tres ejes perpendiculares. La distancia entre ellos sería la raíz cuadrada de la suma de los cuadrados de las diferencias entre las coordenadas de cada perfil, es decir:

$$\text{Distancia euclídea} = \sqrt{(0{,}115 - 0{,}183)^2 + (0{,}269 - 0{,}413)^2 + (0{,}615 - 0{,}404)^2} \qquad (4.5)$$

Esta distancia, cuyo valor es 0,264, corresponde exactamente a la distancia entre el punto E5 y la media de los perfiles que representamos gráficamente en la Imagen 3.2

La distancia ji-cuadrado: un ejemplo de distancia euclídea ponderada

Sin embargo, la expresión (4.4) no es la distancia euclídea —contiene un factor extra, en el denominador de cada término al cuadrado—. Dado que este factor redimensiona o *repondera* cada una de las diferencias al cuadrado, denominamos a esta variante de la distancia euclídea, *distancia euclídea ponderada*. En este caso en particular en el que los factores de ponderación que aparecen en el denominador son los elementos esperados del perfil, la denominamos *distancia ji-cuadrado*, o de forma sintética, distancia χ^2. Por ejemplo, la distancia χ^2 entre la fila E5 y el centroide es:

$$\text{Distancia } \chi^2 = \sqrt{\frac{(0{,}115 - 0{,}183)^2}{0{,}183} + \frac{(0{,}269 - 0{,}413)^2}{0{,}413} + \frac{(0{,}615 - 0{,}404)^2}{0{,}404}} \qquad (4.6)$$

su valor es de 0,431, mayor que la distancia euclídea que calculamos en (4.5), ya que los términos contenidos en la raíz cuadrada han aumentado de valor. En el próximo capítulo veremos cómo visualizar las distancias ji-cuadrado.

Interpretación geométrica de la inercia

A partir de (4.4) y (4.6) podemos expresar la inercia de la siguiente forma:

$$\text{Inercia} = \sum_i (i\text{-ésimo masa}) \times \left(\text{distancia } \chi^2 \text{ de } i\text{-ésimo perfil media}\right)^2 \qquad (4.7)$$

efectuando la suma con relación a las cinco filas de la tabla. Dado que la suma de las masas es 1, podemos decir que la inercia es la media ponderada de los cuadrados de las distancias χ^2 entre los perfiles fila y su perfil media. Por tanto, la inercia será alta cuando los perfiles fila presenten grandes desviaciones con relación a su media, y será baja cuando éstos se hallen cerca de la media. En la imagen 4.2 mostramos una secuencia de cuatro pequeñas matrices de datos, de baja a alta inercia total, cada una de ellas con cinco filas y tres columnas. También hemos representado gráficamente cada una de las matrices en coordenadas triangulares. Hemos escogido estos ejemplos, esencialmente, para visualizar incrementos de magnitud de las inercias. Esta secuencia de mapas también ilustra el concepto de asociación, o de correlación, entre las filas y las columnas de una matriz. Cuando la inercia es baja, los perfiles fila presentan poca variación y se hallan cerca de su perfil medio. En tal caso, decimos que existe poca asociación, o correlación, entre las filas y las columnas. Cuanto

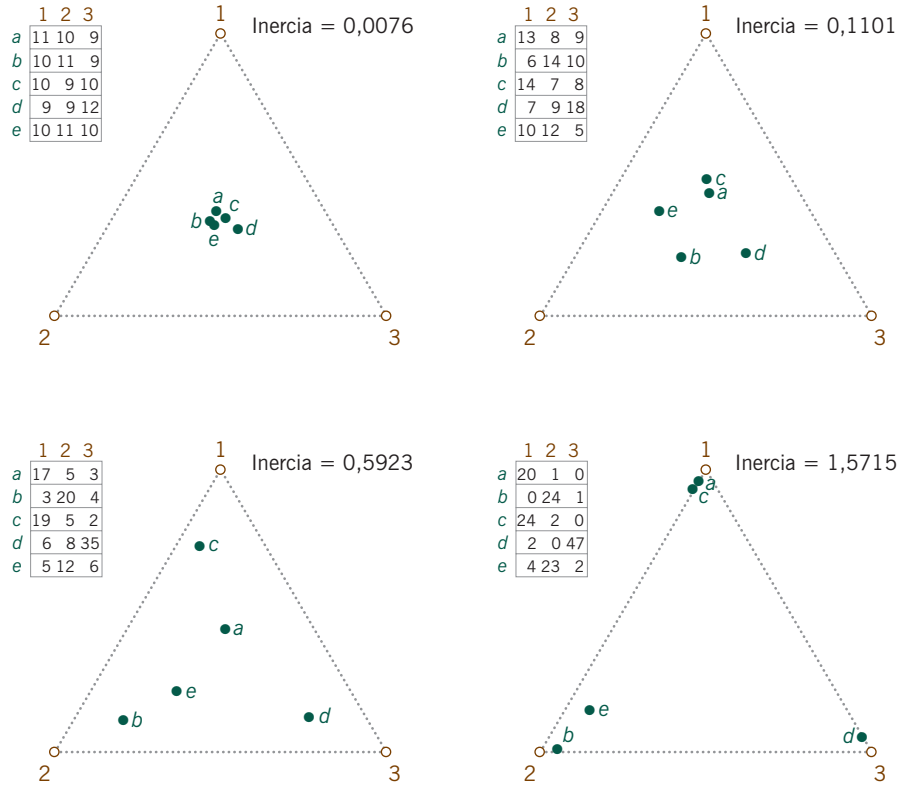

Imagen 4.2: *Serie de tablas de datos con inercia total en aumento. Cuanto mayor sea la inercia total, mayor será la asociación entre las filas y las columnas. Visualizamos este hecho con una mayor dispersión de los puntos en el espacio de perfiles. Hemos escogido los valores de estas tablas de manera que las sumas de las columnas sean todas iguales, y así también lo serán los pesos en la formulación de la distancia χ^2. Por tanto las distancias que observamos en estos mapas son distancias χ^2*

mayor sea la inercia, más cerca se hallarán los perfiles fila de los vértices columna. Es decir, mayor será la asociación entre las filas y las columnas. Más adelante, en el capítulo 8, describiremos de manera más formal la relación existente entre la inercia y el coeficiente de correlación entre las filas y las columnas.

Las inercias máxima y mínima

Si todos los perfiles fueran idénticos, y por tanto todos se hallaran en el mismo punto (su media), todas las distancias ji-cuadrado serían cero, también lo sería la inercia total. Por otro lado, se llegaría a la inercia máxima cuando todos los perfiles se hallaran exactamente en los vértices del espacio de perfiles. En tal caso, la inercia máxima sería igual a la dimensionalidad del espacio (en los ejemplos triangulares de la imagen 4.2, este valor máximo sería igual a 2).

La inercia de las filas es igual a la inercia de las columnas

Hasta ahora hemos visto los conceptos de perfil, de masa, de distancia χ^2 y de inercia, en términos de las filas de una tabla. Tal como comentamos en el capítulo 3, todo lo que hemos descrito hasta ahora para las filas lo podríamos aplicar, de forma equivalente, a las columnas de la tabla (en la imagen 3.4 podemos ver los valores de los perfiles columna, el perfil columna medio y las masas de las columnas). Podríamos comprobar que el resultado del cálculo de la inercia, según la ecuación (4.7) sería idéntico si lo calculáramos a partir de los perfiles columna. Es decir, la inercia total de la tabla, sería igual a la media ponderada de los cuadrados de las distancias χ^2 entre los perfiles columna y su perfil media, ponderadas ahora con las masas de las columnas.

Algo de notación

Esta sección no es imprescindible para la comprensión de los aspectos prácticos del análisis de correspondencias, por tanto la podemos obviar. Sin embargo, será útil para los lectores que quieran comprender la teoría y la literatura sobre el análisis de correspondencias (utilizaremos esta notación en el capítulo 14). Introduciremos un poco de notación estándar para los conceptos definidos hasta el momento aprovechando los datos de la imagen 3.1 (que repetimos en la imagen 4.1).

- n_{ij}: elemento de la tabla de contingencia situado en la *i*-ésima fila y en la *j*-ésima columna, por ejemplo $n_{21} = 18$.
- n_{i+}: el total de la *i*-ésima fila, por ejemplo $n_{3+} = 87$ (el subíndice + indica suma de los elementos del correspondiente índice).
- n_{+j}: el total de la *j*-ésima columna, por ejemplo $n_{+2} = 129$.
- n_{++}: o simplemente n, el total de la tabla, por ejemplo $n = 312$.
- p_{ij}: n_{ij} dividido por el total de la tabla, así, $p_{21} = n_{21}/n = 18/312 = 0{,}0577$.
- r_i: la masa de la *i*-ésima fila, así $r_i = n_{i+}/n$ (lo que equivale a p_{i+}, la suma de frecuencias relativas de la *i*-ésima fila p_{ij}); así $r_3 = 87/312 = 0{,}279$; simbolizamos al vector de masas como **r**.

- c_j: la masa de la j-ésima columna, es decir $c_j = n_{+j}/n$ (lo que equivale a p_{+j}, la suma de las frecuencias relativas de la j-ésima columna p_{ij}); por ejemplo c_2 = 129/312 = 0,414; simbolizamos al vector de las masas de las columnas como **c**.

- a_{ij}: el j-ésimo elemento del perfil de la fila i, así $a_{ij} = n_{ij}/n_{i+}$; así a_{21} = 18/84 = 0,214; simbolizamos al perfil de la fila i por el vector $\mathbf{a}_i$.

- b_{ij}: el i-ésimo elemento del perfil de la columna j, así $b_{ij} = n_{ij}/n_{+j}$; así b_{21} = 18/57 = 0,316; simbolizaremos el perfil de la columna j por el vector $\mathbf{b}_j$.

- $\sqrt{\sum_j (a_{ij} - a_{i'j})^2/c_j}$: la distancia χ^2 entre el i-ésimo y el i'-ésimo perfil fila, lo simbolizamos por $\|\mathbf{a}_i - \mathbf{a}_{i'}\|_c$; así de la imagen 3.1

  $$\|\mathbf{a}_1 - \mathbf{a}_2\|_c = \sqrt{\frac{(0,357-0,214)^2}{0,183} + \frac{(0,500-0,548)^2}{0,413} + \frac{(0,143-0,238)^2}{0,404}} = 0,374.$$

- $\sqrt{\sum_i (b_{ij} - b_{ij'})^2/r_i}$: distancia χ^2 entre el j-ésimo y la j'-ésimo perfil columna, lo simbolizamos por $\|\mathbf{b}_j - \mathbf{b}_{j'}\|_r$; así de la imagen 3.4

  $$\|\mathbf{b}_1 - \mathbf{b}_2\|_r = \sqrt{\frac{(0,088-0,054)^2}{0,045} + \frac{(0,316-0,357)^2}{0,269} + \ldots\text{etc.}} = 0,323$$

  donde 0,088 = 5/57; 0,054 = 7/129; 0,045 = 14/312; etc.

- $\sqrt{\sum_j (a_{ij} - c_j)^2/c_j}$: distancia χ^2 entre el i-ésimo perfil fila $\mathbf{a}_i$ y el perfil fila medio **c** (el vector de las masas de las columnas), lo simbolizamos por $\|\mathbf{a}_i - \mathbf{c}\|_c$; así de la imagen 3.1

  $$\|\mathbf{a}_1 - \mathbf{c}\|_c = \sqrt{\frac{(0,357-0,183)^2}{0,183} + \frac{(0,500-0,413)^2}{0,413} + \frac{(0,143-0,404)^2}{0,404}} = 0,594.$$

- $\sqrt{\sum_i (b_{ij} - r_i)^2/r_i}$: distancia χ^2 entre el j-ésimo perfil columna $\mathbf{b}_j$ y el perfil columna medio **r** (el vector de las masas de las filas), lo simbolizamos por $\|\mathbf{b}_j - \mathbf{r}\|_r$; así de la imagen 3.4

  $$\|\mathbf{b}_1 - \mathbf{r}\|_r = \sqrt{\frac{(0,088-0,045)^2}{0,045} + \frac{(0,316-0,269)^2}{0,269} + \ldots\text{etc.}} = 0,332.$$

Con esta notación, la fórmula de la inercia total (4.7) es:

$$\phi^2 = \frac{\chi^2}{n} = \sum_i r_i \|\mathbf{a}_i - \mathbf{c}\|_c^2 = \sum_i r_i \sum_j \left(\frac{p_{ij}}{r_i} - c_j\right)^2 / c_j \qquad \text{(por fila)} \tag{4.8}$$

$$= \sum_j c_j \|\mathbf{b}_i - \mathbf{r}\|_r^2 = \sum_j c_j \sum_i \left(\frac{p_{ij}}{c_j} - r_i\right)^2 / r_i \qquad \text{(por columna)} \tag{4.9}$$

y su valor 0,0833, por lo tanto, χ^2 = 0,0833 × 312 = 26,0.

RESUMEN:
Distancia ji-cuadrado e inercia

1. El estadístico ji-cuadrado (χ^2) es una medida global de las diferencias entre las frecuencias observadas y las frecuencias esperadas de una tabla de contingencia. Calculamos las frecuencias esperadas mediante la hipótesis de homogeneidad de los perfiles fila (o de los perfiles columna)
2. La *inercia* (*total*) de una tabla de contingencia es igual al estadístico χ^2 dividido por el total de la tabla.
3. Geométricamente, la inercia mide lo «lejos» que se hallan los perfiles fila (o los perfiles columna) de su perfil medio. Podemos considerar que el perfil medio simboliza la hipótesis de homogeneidad (es decir, de igualdad) de los perfiles.
4. Medimos las distancias entre los perfiles utilizando la *distancia ji-cuadrado* (distancia χ^2). La formulación de esta distancia es similar a la *distancia euclídea* (o *pitagórica*) entre puntos en un espacio físico, salvo por el hecho de que dividimos cada cuadrado de la diferencia entre coordenadas por su correspondiente elemento del perfil medio.
5. Podemos expresar la inercia de manera que la podamos interpretar como una media ponderada de las distancias χ^2 entre los perfiles fila y su perfil medio (de forma similar, entre los perfiles columna y su media).