ANÁLISIS DE CORRESPONDENCIA

Esta técnica estadística es de gran utilidad puesto que la interpretación del resultado puede hacerse de manera sencilla a través de gráficas. Con este procedimiento se puede evidenciar de manera más perceptible el grado de relación entre las categorías de cada variable; de ahí el nombre de mapas perceptuales. Cuando el grado de asociación es alto, éstas aparecerán en el diagrama relativamente juntas (Salvador, 2001).

Surge con el fin de definir, describir e interpretar las relaciones entre variables categóricas a través de un gráfico geométrico.

Un medio descriptivo numérico próximo a estos mapas de percepción, son las tablas de contingencia o también conocidas tablas cruzadas o matriz de tabulación. Por tanto el Análisis de Correspondencia (AC) es una técnica gráfica que representa información contenida en una tabla de contingencia de dos vías la cual representa la totalización (frecuencia) de las observaciones de una muestra dada, para una tabla cruzada de dos variables categóricas. Con el AC se construye una gráfica (mapa perceptual) que señala la interacción de dos variables categóricas a través de la relación de las filas y de las columnas entre sí. Mide el grado de asociación presente entre un conjunto de variables; es decir, construye un diagrama cartesiano o mapa perceptual basado en la relación de dependencia e independencia de los atributos o categorías de las variables. Algunas referencias indispensables en este tema son Greenacre (1984), Jobson (1992, Sección 9.4), Khattree y Naik (1999, Cápitulo 7), Gower y Hand (1996, Cápitulos 4 y 9) y Benzecri (1992). Adicionalmente, se puede relacionar con una técnica más de reducción de variables al transformarlas a un conjunto de variables observables. (Díaz, 2002)

Las variables originales deben conformarse de atributos o categorías. Es decir deben ser cualitativas.

En las tablas de contingencia cada variable con sus correspondientes atributos

es contabilizada contra las otras variables, representando, la frecuencia de asociación de las categorías. La finalidad es poner de manifiesto gráficamente las relaciones de dependencia existentes entre las diversas modalidades de dos o más variables categóricas, es decir cualitativas, a partir de la información proporcionada por sus estas tablas de frecuencias cruzadas o tablas de contingencia.

Las variables categóricas muestran en un mapa su recomposición mediante la asociación de categorías o atributos para conformación de conglomerados a través de la varianza. Esos conglomerados están conformados por categorías de las variables originales y tendrían una varianza mínima internamente y máxima entre ellos.

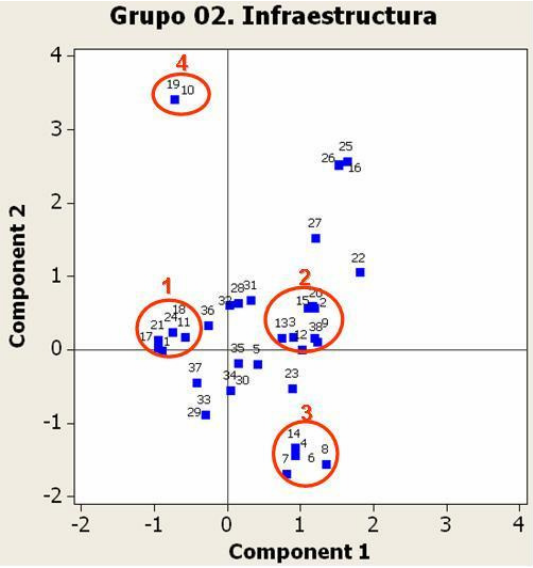
En el análisis de correspondencia, el mapa perceptual muestra un punto por cada fila y un punto para cada columna de la tabla de contingencia. Estos puntos son, en efecto, las proyecciones de las filas y columnas de la tabla de contingencia en un espacio euclidiano de dos dimensiones. El objetivo es preservar tanto como sea posible la relación de las filas (o columnas) a la otra en un espacio de dos dimensiones. Si dos puntos-fila están muy juntos, los perfiles de las dos filas (a través de las columnas) son similares.

Asimismo, dos puntos-columna que están muy juntos representan columnas con perfiles similares a través de las filas. Si un punto de fila está cerca de un punto de la columna, esta combinación de categorías de las dos variables es más frecuente de lo que ocurriría, por casualidad, si las dos variables son independientes. Otro resultado de un análisis de correspondencia es la inercia, o la cantidad de información en cada una de las dos dimensiones en la trama. El mapa perceptual muestra, en fin, los puntos (categorías de las variables observadas) que indican la relación o correspondencia que pudiera existir entre las variables de estudio. Las relaciones se pueden observar cuando se forman algunos conglomerados (concentración de puntos) que describen cierto comportamiento particular (patrón).

Cuando en el gráfico los puntos (variables observadas) se encuentran en el centro del eje, indica que existe colinealidad entre las variables; es decir, existen variables que están fuertemente interrelacionadas, y, por tanto, resulta difícil medir sus efectos individuales sobre la variable respuesta (variable de interés). Las variables redundantes pueden ser identificadas a través de la matriz de correlación y de este modo, se podría mitigar este fenómeno, siempre y cuando las variables sean cuantitativas.

Para probar la importancia de la asociación de las dos variables categóricas en una tabla de contingencia, podríamos usar una prueba de chi-cuadrado o un modelo log-lineal, los cuales representan una aproximación asintótica. Si una tabla de contingencia tiene algunas celdas frecuencias con valores pequeños o nulos, la aproximación chi-cuadrado no es muy satisfactoria. En este caso, algunas categorías se pueden combinar para aumentar las frecuencias de las celdas y así, disminuir el número de categorías originales. Es importante destacar lo útil de identificar categorías que sean similares; esto permitiría combinarlas y de allí, crear una variable observable que explique mejor los resultados.

Seguidamente se muestran dos mapas perceptuales en los que se representan mediante círculos las nuevas variables después de reagruparse las categorías.

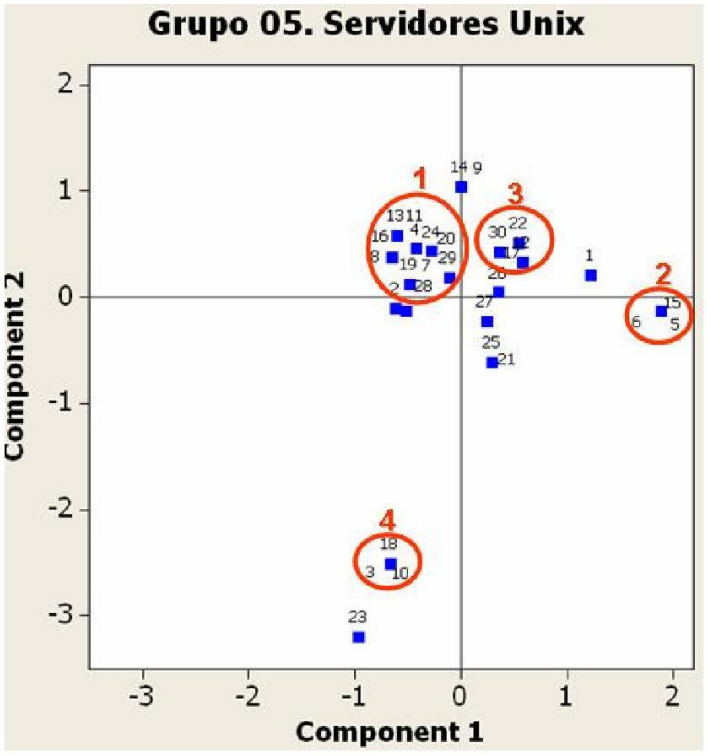


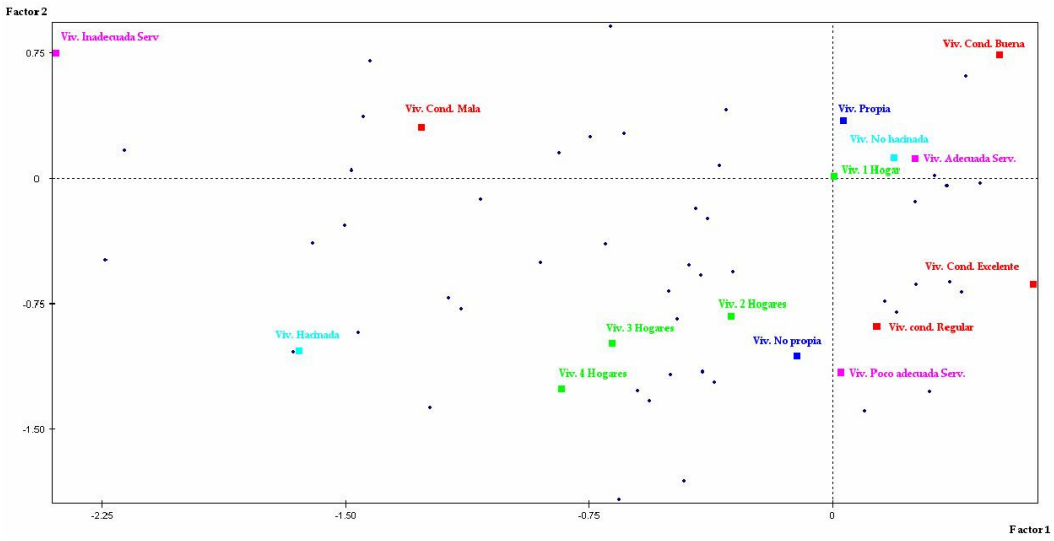
En estos gráficos, usando MINITAB, se observa cada categoría identificada por un número que al mostrar su grado asociación o similitud, estructura estos cluster que conforman la variable respuesta.

A su vez, cada eje, en cada una de sus direcciones identifica una característica no presente en las observaciones originales, marcando en estos mapas un alto grado de similitud hacia esa característica nueva o en su defecto, su opuesto, la disimilitud.

Puede revisarse Hair et al., en el capítulo diez.
De acuerdo a los gráficos aportados por los mapas perceptuales, se asocia a cada modalidad un punto en el espacio **P,** de forma que:

a) Cuanto más alejado del origen de coordenadas está el punto asociado a una modalidad de una variable, más diferente es su perfil condicional del perfil marginal correspondiente a las otras variables.
b) Los puntos correspondientes a dos modalidades diferentes de una misma variable estarán más cercanos cuanto más se parezcan sus perfiles condicionales.
c) Dichos puntos tenderán a estar más cerca de aquellas modalidades con las que tienen una mayor afinidad; es decir, aquéllas en los que las frecuencias observadas de la celda correspondiente tiende a ser mayor que la esperada bajo la hipótesis de independencia de las variables correspondientes.


Grupo 05. Servidores Unix
Component 2
Component 1
1
2
3
4


Factor 2
Factor 1
Viv. Inadecuada Serv
Viv. Cond. Mala
Viv. Cond. Buena
Viv. Propia
Viv. No hacinada
Viv. Adecuada Serv.
Viv. 1 Hogar
Viv. Cond. Excelente
Viv. 2 Hogares
Viv. cond. Regular
Viv. Hacinada
Viv. 3 Hogares
Viv. No propia
Viv. 4 Hogares
Viv. Poco adecuada Serv.

PERFILES DE FILAS Y COLUMNAS

Una tabla de contingencia con *a* filas y *b* columnas se muestra en la tabla más abajo. El entradas n_{ij} son los totales o frecuencias para cada combinación de fila y columna (cada celda). Los totales marginales se muestran usando la notación familiar de puntos: $n_{i.} = \sum_{j=1}^{b} n_{ij}$ y $n_{.j} = \sum_{j=1}^{a} n_{ij}$, indican la suma de todas las columnas y de todas las filas respectivamente. La frecuencia total general se denota por *n* en lugar de *n..* por simplicidad: $n = \sum_{i,j} n_{ij}$.

Las frecuencias n_{ij} en una tabla de contingencia se puede convertir a las frecuencias relativas p_{ij} al dividir entre *n*; esto es, $p_{ij} = n_{ij}/n$. La matriz de frecuencias relativas se llama matriz de correspondencia y se denota por **P**:

Tabla de contingencia con *a* filas y *b* columnas

		columnas				
		1	*2*	…..	*b*	**Total fila**
filas	*1*	n_{11}	n_{12}	…..	n_{1b}	$n_{1.}$
	2	n_{21}	n_{22}	…..	n_{2b}	$n_{2.}$
	:	*:*	*:*		*:*	*:*
	a	n_{a1}	n_{a2}	…..	n_{ab}	$n_{a.}$
	Total columna	$n_{.1}$	$n_{.2}$	…..	$n_{.b}$	*n*

Matriz de correspondencias de frecuencias relativas **P**

		columnas				
		1	*2*	…..	*b*	**Total fila**
filas	*1*	p_{11}	p_{12}	…..	p_{1b}	$p_{1.}$
	2	p_{21}	p_{22}	…..	p_{2b}	$p_{2.}$
	:	*:*	*:*		*:*	*:*
	a	p_{a1}	p_{a2}	…..	p_{ab}	$p_{a.}$
	Total columna	$p_{.1}$	$p_{.2}$	…..	$p_{.b}$	*p*

$$\mathbf{P} = (p_{ij}) = p_{ij} / p$$

La última columna de la tabla anterior contiene la suma de las filas $p_{i.} = \sum_{j=1}^{b} p_{ij}$. Este vector columna se representa por **r** y se pueden obtener como

$$\mathbf{r} = \mathbf{Pj} = (p_{1.}, p_{2.}, \ldots, p_{a.})' = (n_1/n, n_{2.}/n, \ldots, n_{a.}/n)'$$

donde **j** es un vector ax1 de 1(s). De manera similar la última fila de la tabla anterior contiene la suma de las columnas $p_{.j} = \sum_{i=1}^{a} p_{ij}$. Este vector columna se representa por **c** y se pueden obtener como

$$\mathbf{c'} = \mathbf{j'P} = (p_{.1}, p_{.2}, \ldots, p_{.b})' = (n_{.1}/n, n_{.2}/n, \ldots, n_{.b}/n)'$$

Los elementos de los vectores **r** y **c** también se le conocen como *filas y columnas masas*. La matriz de correspondencia y los totales marginales pueden ser expresados de acuerdo a una matriz ampliada

$$\begin{bmatrix} \mathbf{P} & \mathbf{r} \\ \mathbf{c}' & \mathbf{1} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdot\cdot & p_{1b} & p_{1.} \\ p_{21} & p_{22} & \cdot\cdot & p_{2b} & p_{2.} \\ \vdots & \vdots & \cdot\cdot & \vdots & \vdots \\ p_{a1} & p_{a2} & \cdot\cdot & p_{ab} & p_{a.} \\ p_{.1} & p_{.2} & \cdot\cdot & p_{.b} & 1 \end{bmatrix}$$

La definición de los perfiles de cada fila y columna de P es como sigue. La i-ésima fila-perfil $\mathbf{r_i}'$, *i=1,2,…,a,* se define dividiendo la i-ésima fila de cualquiera de las tablas anteriores entre su total marginal:

$$\mathbf{r_i}' = \left(\frac{p_{i1}}{p_{i.}}, \quad \frac{p_{i2}}{p_{i.}}, \quad \ldots, \quad \frac{p_{ib}}{p_{i.}} \right) = \left(\frac{n_{i1}}{n_{i.}}, \quad \frac{n_{i2}}{n_{i.}}, \quad \ldots, \quad \frac{n_{ib}}{n_{i.}} \right).$$

Los elementos de cada $\mathbf{r_i}'$ son frecuencias relativas y de ahí que su suma sea 1.

$$\mathbf{r}_i'\mathbf{j} = \begin{pmatrix} \frac{n_{i1}}{n_{i.}} & \frac{n_{i2}}{n_{i.}} & \dots & \frac{n_{ib}}{n_{i.}} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ : \\ 1 \end{pmatrix} = \sum_{j=1}^{b} \frac{n_{ij}}{n_{i.}} = \frac{n_{i.}}{n_{i.}} = 1 .$$

Por definición,

$$\mathbf{D_r} = \text{diag}(\mathbf{r}) = \begin{bmatrix} p_{1.} & 0 & \dots & 0 \\ 0 & p_{2.} & \dots & 0 \\ : & : & \dots & : \\ 0 & 0 & \dots & p_{a.} \end{bmatrix} \text{ y, } \mathbf{D_r^{-1}} = \begin{bmatrix} 1/p_{1.} & 0 & \dots & 0 \\ 0 & 1/p_{2.} & \dots & 0 \\ : & : & \dots & : \\ 0 & 0 & \dots & 1/p_{a.} \end{bmatrix}.$$

La matriz **R** de filas-perfil puede ser expresada como

$$\mathbf{R} = \mathbf{D_r^{-1}}.\mathbf{P} = \begin{bmatrix} r_1' \\ r_2' \\ : \\ r_a' \end{bmatrix} = \begin{bmatrix} \frac{p_{11}}{p_{1.}} & \frac{p_{12}}{p_{1.}} & \dots & \frac{p_{1b}}{p_{1.}} \\ \frac{p_{21}}{p_{2.}} & \frac{p_{22}}{p_{2.}} & \dots & \frac{p_{2b}}{p_{12.}} \\ : & : & \dots & : \\ \frac{p_{a1}}{p_{a.}} & \frac{p_{a2}}{p_{a.}} & \dots & \frac{p_{ab}}{p_{a.}} \end{bmatrix}.$$

De manera similar sucede para la columna-perfil, $\mathbf{c_j}$, *j=1,2…,b,* se define dividiendo la j-ésima columna entre su total marginal. Esto es

$$\mathbf{c_j} = \begin{pmatrix} \frac{p_{1j}}{p_{.j}} & \frac{p_{2j}}{p_{.j}} & \dots & \frac{p_{aj}}{p_{.j}} \end{pmatrix}' = \begin{pmatrix} \frac{n_{1j}}{n_{.j}} & \frac{n_{2j}}{n_{.j}} & \dots & \frac{n_{aj}}{n_{.j}} \end{pmatrix}'.$$

Los elementos en cada $\mathbf{c_j}$ son frecuencias relativas y su suma es 1.

$$\mathbf{j'c_j} = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} \frac{n_{1j}}{n_{.j}} \\ \frac{n_{2j}}{n_{.j}} \\ \vdots \\ \frac{n_{aj}}{n_{.j}} \end{pmatrix} = \sum_{i=1}^{a} \frac{n_{ij}}{n_{.j}} = \frac{n_{.j}}{n_{.j}} = 1, \text{ y ahora}$$

$$\mathbf{D_c} = \text{diag}(\mathbf{c}) = \begin{bmatrix} p_{.1} & 0 & \dots & 0 \\ 0 & p_{.2} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & p_{.b} \end{bmatrix}$$

Y usando la matriz de columnas-perfil **C**, se tiene

$$\mathbf{C} = \mathbf{P}\mathbf{D}_\mathbf{c}^{-1} = \begin{pmatrix} \mathbf{c_1} & \mathbf{c_2} & \dots & \mathbf{c_b} \end{pmatrix} = \begin{pmatrix} \frac{p_{11}}{p_{.1}} & \frac{p_{12}}{p_{.2}} & \dots & \frac{p_{1b}}{p_{.b}} \\ \frac{p_{21}}{p_{.1}} & \frac{p_{22}}{p_{.2}} & \dots & \frac{p_{2b}}{p_{.b}} \\ \vdots & \vdots & \dots & \vdots \\ \frac{p_{a1}}{p_{.1}} & \frac{p_{a2}}{p.1} & \dots & \frac{p_{ab}}{p_{.b}} \end{pmatrix}.$$

El vector **r** que quedó definido como un vector columna de las suma de filas de **P**, **r = Pj** = $(p_{1.}, p_{2.}, \dots, p_{a.})' = (n_1/n, n_2/n, \dots, n_a/n)'$, se puede expresar como la media ponderada de los columnas-perfil.

$$\mathbf{r} = \sum_{j=1}^{b} p_{.j}\mathbf{c_j}; \text{ o lo que es lo mismo}$$

$$\begin{pmatrix} p_{1.} & p_{2.} & \dots & p_{a.} \end{pmatrix}' = p_{.1}\mathbf{c_1} + p_{.2}\mathbf{c_2} + \dots + p_{.b}\mathbf{c_b},$$

al sustituir a $\mathbf{c_i}$ por sus respectivos vectores y realizar las operaciones correspondientes, quedaría

$$\begin{aligned} p_{1.} &= p_{11} + p_{12} + ... + p_{1b} \\ p_{2.} &= p_{21} + p_{22} + ... + p_{2b} \\ &..... \\ p_{a.} &= p_{a1} + p_{a2} + ... + p_{ab} \end{aligned}$$, que es el vector de las sumas de filas, **r**.

Del mismo modo ocurre para **c'** que es el vector fila de las sumas de columnas de **P**, mediante la expresión $\mathbf{c'} = \sum_{i=1}^{a} p_{i.}\mathbf{r'}$.

Se sabe que para cualquier fila o columna, $\sum_{j=1}^{b} p_{.j} = \sum_{i=}^{a} p_{i.} = 1$, y de ahí, $\mathbf{j'r} = \mathbf{c'j} = \mathbf{1}$, donde la suma de una fila-perfil en **r** con **j** de *ax1* de **P,** es igual a una **b** columna-perfil en **c** con **j** de *bx1* de **P** y es igual 1.

PRUEBA DE INDEPENDENCIA

Como se sabe los datos en una tabla de contingencia pueden ser usados para verificar la asociación de dos variables categóricas. Supóngase dos variables categóricas, x y y, y de acuerdo con lo visto en la sección anterior, la suposición de independencia se puede expresar en términos de probabilidad mediante

$$P(x_i y_j) = P(x_i)P(y_j),\ i=1,2,...,a \ \ y \ \ j=1,2,\ldots,b,$$

Donde x_i y y_j se corresponden a la *i-ésima* fila y *j-ésima* columna de la matriz de correspondencia, se puede estimar

$$p_{ij} = p_{i.}p_{.j} \text{ con } i=1,2,...,a \ \ y \ \ j=1,2,\ldots,b.$$

La chi-cuadrado para probar la hipótesis nula que indica independencia de x y y, al comparar p_{ij} con $p_{i.}$ y $p_{.j}$, está dada por

$$\chi^2 = n\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}},$$

La cual es aproximadamente asintóticamente distribuida como una variable aleatoria chi-cuadrado con *(a-1)(b-1) grados de libertad*.

En función de la cantidad de observaciones en la tabla de contingencia, en lugar de las frecuencias relativa p_{ij}, se puede re-escribir con el total n, el total en cada celda n_{ij} y en cada columna $n_{i.}$ y fila $n_{.j}$,

$$\chi^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(n_{ij} - \dfrac{n_{i.}n_{.j}}{n}\right)^2}{\dfrac{n_{i.}n_{.j}}{n}}.$$

La expresión anterior también se puede re-escribir en función vectorial, tal como se ha visto hasta este momento. En función de $\mathbf{r}$, $\mathbf{r_i}$, $\mathbf{D_c}$, $\mathbf{D_r}$, $\mathbf{c_j}$ y $\mathbf{c}$. Cualquiera de las dos expresiones siguientes son aplicables: mediante la comparación de los vectores $\mathbf{r_i}$ a $\mathbf{c}$ para cada i, y mediante los vectores $\mathbf{c_j}$ a $\mathbf{r}$ para cada j. Cualquiera de estas comparaciones es equivalente a probar la independencia comparando p_{ij} a $p_{i.}p_{.j}$, para todo i, j.

$$\chi^2 = \sum_{i=1}^{a} np_{i.}(\mathbf{r_i} - \mathbf{c})'\mathbf{D_c^{-1}}(\mathbf{r_i} - \mathbf{c}), \text{ o } \chi^2 = \sum_{j=1}^{b} np_{.j}(\mathbf{c_j} - \mathbf{r})'\mathbf{D_r^{-1}}(\mathbf{c_j} - \mathbf{r}).$$

En conclusión, es equivalente la prueba de independencia mediante la chi-cuadrado si se aplica cualquiera de los siguientes tres procedimientos.

a) $p_{ij}=p_{i.}p_{.j}$ para todo i, j, (**P=rc'**).
b) Todas las filas $\mathbf{r_i}$' de **R** son iguales (también iguales a su media ponderada, **c'**).
c) Todas las columnas $\mathbf{c_j}$ de **C** son iguales (también iguales a su media ponderada, **r'**).

De este modo, si x y y fueran independientes, se esperaría que las filas de la tabla de contingencia tendrían perfil similar o de manera equivalente, las columnas tendrían perfil similar.

En forma vectorial, la chi-cuadrado se puede exptresar de la siguiente manera

$$\chi^2 = n \cdot \mathbf{tr}[\mathbf{D_r^{-1}}(\mathbf{P} - \mathbf{rc}')\mathbf{D_c^{-1}}(\mathbf{P} - \mathbf{rc}')'],$$

donde **tr** es la traza de la matriz resultante de la expresión o lo que es lo mismo, la suma de la diagonal de esa matriz y *n* es la suma total de las frecuencias de la tabla de contingencia.

La expresión anterior es equivalente a escribir $n = \sum_{i=1}^{k} \lambda_i^2$, que son los *k* autovalores diferentes de cero de $[\mathbf{D_r^{-1}(P - rc')D_c^{-1}(P - rc')}]'$ y *k* es el rango de $[\mathbf{D_r^{-1}(P - rc')D_c^{-1}(P - rc')}]'$. Hay que recordar que este rango *k* está asociado al min[*(a-1),(b-1)*].

COORDENADAS PARA GRAFICAR LOS PERFILES DE FILAS Y COLUMNAS

En este punto se consideran los aspectos fundamentales para establecer, en general, las coordenadas para un análisis de correspondencia, sea de dos o más variables.

La métrica para los puntos de filas y columnas es la misma y los dos conjuntos de puntos pueden ser superpuestos en el mismo gráfico.

Para obtener estas coordenadas se factoriza la matriz mediante una descomposición espectral. En análisis de correspondencia la matriz **P-rc'** no es simétrica y de ahí el uso de valor de descomposición singular (svd) para obtener las coordenadas.

Se escala **P-rc'** para obtener $\mathbf{Z} = \mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc'})\mathbf{D_c^{-1/2}}$, cuyos elementos de **Z** son $z_{ij} = \dfrac{p_{ij} - p_{i.}p_{.j}}{\sqrt{p_{i.}}p_{.j}}$. Se factoriza a Z mediante svd, $\mathbf{Z} = \mathbf{U}\Lambda\mathbf{V}'$, donde **U** y **V** son autovectores ortonormales y Λ es una matriz diagonal con λ_i, $i=1,\ldots,k$, donde k es el min*[(a-1)(b-1)*] y se corresponden con los valores singulares de **Z**. Las columnas *axk* de **U** y las columnas *bxk* de **V** son auitovectores normalizados de **Z'Z** y de ahí que λ_i^2, $i=1,\ldots,k$, sean los autovalores de **Z´Z**. Note que

$$\begin{aligned}\mathbf{Z´Z} &= \mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc'})\mathbf{D_c^{-1/2}}\mathbf{D_c^{-1/2}}(\mathbf{P} - \mathbf{rc'})'\mathbf{D_r^{-1/2}} \\ &= \mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc'})\mathbf{D}_c^{-1}(\mathbf{P} - \mathbf{rc'})'\mathbf{D_r^{-1/2}} \\ = \mathbf{D_r^{-1/2}}\mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc'})\mathbf{D}_c^{-1}(\mathbf{P} - \mathbf{rc'})' &= = \mathbf{D_r^{-1}}(\mathbf{P} - \mathbf{rc'})\mathbf{D}_c^{-1}(\mathbf{P} - \mathbf{rc'})',\end{aligned}$$

que es la expresión ya conocida para determinar los k autovalores.

Para la descomposición de P-rc', se puede igualando

$$\mathbf{Z} = \mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc'})\mathbf{D_c^{-1/2}} \text{ con } \mathbf{Z} = \mathbf{U\Lambda V'}.$$
$$\mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc'})\mathbf{D_c^{-1/2}} = \mathbf{U\Lambda V}$$
$$(\mathbf{P} - \mathbf{rc'}) = \mathbf{D_r^{1/2}U\Lambda V'D_c^{1/2}} = \mathbf{A\Lambda B'},$$
$$\mathbf{A\Lambda B'} = \sum_{j=1}^{k} \lambda_j \mathbf{a_j b_j}$$

Luego $\mathbf{A} = \mathbf{D_r^{1/2}U}$ y $\mathbf{B} = \mathbf{D_c^{1/2}V}$, $\mathbf{a_i}$ y $\mathbf{b_i}$ son las columnas de **A** y **B**, y $\Lambda = \lambda_i$, *i=1,…,k.* Además, **UU'=I** y **VV'=I**, luego por la expresión anterior, A y B están escaladas y $\mathbf{A'D_r^{-1}A}$ y $\mathbf{B'D_c^{-1}B}$ son iguales a **I.**

Las filas de **P-rc'** están representadas por la combinación lineal de las filas de **B'**, las cuales son las columnas de **B=(b$_1$,b$_2$,…,b$_k$)**. Los coeficientes (coordenadas) para la i-ésima fila de **P-rc'** están en la i-ésima fila de **AΛ** y de la misma manera, las coordenadas para las columnas de **P-rc'** están dadas por las columnas de **AB'**, puesto que **AB'** provee los coeficientes para **A=(a$_1$,a$_2$,….,a$_k$).**

Para encontrar las coordenadas para las desviaciones de las filas **r$_i$'-c'**y las desviaciones en las columnas **c$_j$-r**, se expresa en forma matricial y en función de **P-rc'** de la siguiente manera

$$\mathbf{R} - \mathbf{jc'} = \mathbf{D_r^{-1}}(\mathbf{P} - \mathbf{rc'}) \text{ y}$$
$$\mathbf{C} - \mathbf{rj'} = \mathbf{D_c^{-1}}(\mathbf{P} - \mathbf{rc'})$$

De este modo, las coordenadas para las filas en **R-jc'** con respecto a los ejes **b$_1$,b$_2$,…,b$_k$**, están dados por las columnas de $\mathbf{X} = \mathbf{D_r^{-1}A\Lambda}$, por otro lado, las coordenadas para las columnas de C-rj'n con respecto a los ejes **a$_1$,a$_2$,…,a$_k$**, están dados por $\mathbf{Y} = \mathbf{D_c^{-1}B\Lambda}$.

De allí, se tendría que para graficar las coordenadas para las desviaciones de perfil-filas $\mathbf{R\text{-}jc'=r_i'\text{-}c'}$, $i=1,2,\ldots,a$, en dos dimensiones, para dos columnas de X, sería

$$\mathbf{X}=\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ : & : \\ x_{a1} & x_{a2} \end{pmatrix}$$

Del mismo modo par Y. Las coordenadas para las columnas de las desviaciones del perfil-columnas $\mathbf{C–rj'= c_j\text{-}r}$, $j=1,2,\ldots,b$, en dos dimensiones sería

$$\mathbf{Y}=\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ : & : \\ y_{b1} & y_{b2} \end{pmatrix}.$$

Teniendo en cuenta que cada punto tiene un peso o ponderación iguala su masa (los elementos de los vectores **r** y **c** se le conocen como *filas y columnas masas)*, la inercia sería un estadístico adecuado para medir la dispersión de la nube de puntos. Esta dispersión es el promedio de las distancias de los puntos a su centro de gravedad. Mayores detalles s e pueden consultar en *escalamiento multidimensional.*

De este modo la media ponderada (ponderada por p_i) de las distancias chi-cudrado $(\mathbf{r_i}-\mathbf{c})\mathbf{D_c^{-1}}(\mathbf{r_i}-\mathbf{c})$ entre las filas-perfil $\mathbf{r_i}$ y y sus media **c** es llamada inercia total y puede ser expresado por

$$\frac{\chi^2}{n}=\sum_{i=1}^{a} p_{i.}(\mathbf{r_i}-\mathbf{c})'\mathbf{D_c^{-1}}(\mathbf{r_i}-\mathbf{c}), \text{ o } \frac{\chi^2}{n}=\sum_{j=1}^{b} p_{.j}(\mathbf{c_j}-\mathbf{r})'\mathbf{D}_r^{-1}(\mathbf{c_j}-\mathbf{r}).$$

Pero como $\sum_i p_{i.} = \sum_j p_{.j} = 1$, entonces, $\frac{\chi_2}{n} = \sum_{i=1}^{k} \lambda_i^2$ y de ahí la contribución de cada una de las primeras dos dimensiones del gráfico al total de inercia es $\frac{\lambda_1^2}{\sum_{l=1}^{k} \lambda_i^2}$ y $\frac{\lambda_2^2}{\sum_{l=1}^{k} \lambda_i^2}$. La combinada de las dos dimensiones sería $\frac{\lambda_1^2 + \lambda_2^2}{\sum_{l=1}^{k} \lambda_i^2}$.

Hay un procedimiento en MATLAB que puede producir este conjunto de estadísticos y valores de pruebas, conjuntamente con las gráficas que ayuden a formular reducciones de de categorías con las combinación de dos o más de ellas, en variables latentes que a juicio particular daría origen a nuevas variables. Del mismo modo, SPSS, SAS, Minitab y otros, ofrecen esta herramienta multivariante de reducción de variables para hacer un análisis, más descriptivo, y especialmente gráfico, de los valores observados.

EJEMPLOS

AC Y GRÁFICA CON DOS VARIABLES

Cantidad de fallas de los aros de pistón en la tres patas				
	Pata del compresor			
Compresor	A	B	C	Total fila
1	17	17	12	46
2	11	9	13	33
3	11	8	19	38
4	14	7	28	49
Total col.	53	41	72	166

Matriz de correspondencia				
	Pata del compresor			
Compres	A	B	C	Total fila
1	0,102	0,102	0,072	0,277
2	0,066	0,054	0,078	0,199
3	0,066	0,048	0,114	0,229
4	0,084	0,042	0,169	0,295
Total col	0,319	0,247	0,434	1

fila-perfil			
0,3696	0,3696	0,2609	1
0,3333	0,2727	0,3939	1
0,2895	0,2105	0,5000	1
0,2857	0,1429	0,5714	1

columna-perfil		
0,3208	0,4146	0,1667
0,2075	0,2195	0,1806
0,2075	0,1951	0,2639
0,2642	0,1707	0,3889
1	1	1

$R=inv(D_r)*P$		
0,3697	0,3697	0,2610
0,3330	0,2724	0,3935
0,2894	0,2104	0,4998
0,2859	0,1429	0,5718

$C=P*inv(D_c)$		
0,3210	0,4146	0,1666
0,2077	0,2195	0,1804
0,2077	0,1951	0,2637
0,2644	0,1707	0,3887

Prueba de independencia

$inv(D_r)*(P-r*c')$		
0,050	0,123	-0,173
0,014	0,026	-0,040
-0,030	-0,036	0,066
-0,034	-0,104	0,138

$inv(D_c)*(P-r*c')'$			
0,044	0,009	-0,021	-0,031
0,138	0,021	-0,034	-0,124
-0,110	-0,018	0,035	0,094

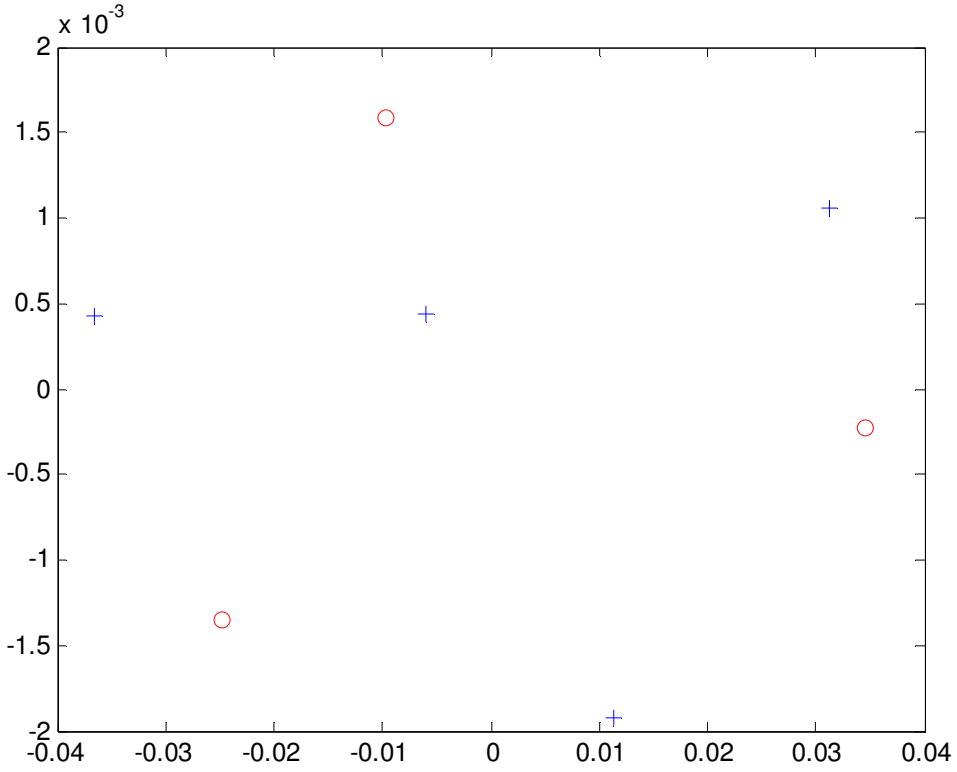
$inv(D_r)*(P-r*c')*inv(Dc)*(P-r*c')'$			
0,038	0,006	-0,011	-0,033
0,009	0,001	-0,003	-0,007
-0,014	-0,002	0,004	0,012
-0,031	-0,005	0,009	0,027

Como la traza es la suma de la diagonal y ella resulta en 0.071, la χ^2 con (a-1)*(b-1) grados de libertad, donde a=4 y b=3, resulta en 6 g.d.l.

CHI-CUADRADO= n*tr([inv($\mathbf{D_r}$)*(**P-r*c'**)*inv($\mathbf{D_c}$)*(**P-r*c')'**]					
CHI-Cuad.	11,724				

De acuerdo a tabla el valor-p sería 0.085, que acepta la hipótesis nula, la cual hay evidencia de la pérdida de independencia entre ambas variables y por supuesto, se puede establecer alguna asociación.

Métrica para las coordenadas



ACM Y GRÁFICA CON MÚLTIPLES VARIABLES

Lista de 12 observaciones y sus categorias en cuatro variables				
Obs.	Género	Edad	Estado civil	Color pelo
1	M	joven	soltero	castaño
2	M	adulto	soltero	rojizo
3	F	mayor	casado	claro
4	M	adulto	soltero	negro
5	F	mayor	casado	negro
6	F	mayor	soltero	castaño
7	M	joven	casado	rojizo
8	M	adulto	casado	claro
9	M	mayor	soltero	castaño
10	F	joven	casado	negro
11	F	adulto	soltero	castaño
12	M	joven	casado	claro

					G						
Obs.	Género		Edad			Edo. civil		Color pelo			
1	1	0	1	0	0	1	0	0	1	0	0
2	1	0	0	0	1	1	0	0	0	0	1
3	0	1	0	1	0	0	1	1	0	0	0
4	1	0	0	0	1	1	0	0	0	1	0
5	0	1	0	1	0	0	1	0	0	1	0
6	0	1	0	1	0	1	0	0	1	0	0
7	1	0	1	0	0	0	1	0	0	0	1
8	1	0	0	0	1	0	1	1	0	0	0
9	1	0	0	1	0	1	0	0	1	0	0
10	0	1	1	0	0	1	0	0	1	0	0
11	0	1	0	0	1	0	1	1	0	0	0
12	1	0	1	0	0	0	1	1	0	0	0

	G'											
Obs.	1	2	3	4	5	6	7	8	9	10	11	12
Género	1	1	0	1	0	0	1	1	1	0	0	1
	0	0	1	0	1	1	0	0	0	1	1	0
Edad	1	0	0	0	0	0	1	0	0	1	0	1
	0	0	1	0	1	1	0	0	1	0	0	0
	0	1	0	1	0	0	0	1	0	0	1	0
Edo. civil	1	1	0	1	0	1	0	0	1	1	0	0
	0	0	1	0	1	0	1	1	0	0	1	1
Color pelo	0	0	1	0	0	0	0	1	0	0	1	1
	1	0	0	0	0	1	0	0	1	1	0	0
	0	0	0	1	1	0	0	0	0	0	0	0
	0	1	0	0	0	0	1	0	0	0	0	0

MATRIZ DE BURT (G'G)												
Género	M	7	0	3	1	3	4	3	2	2	1	2
	F	0	5	1	3	1	2	3	2	2	1	0
Edad	J	3	1	4	0	0	2	2	1	2	0	1
	M	1	3	0	4	0	2	2	1	2	1	0
	A	3	1	0	0	4	2	2	2	0	1	1
Edo. civil	S	4	2	2	2	2	6	0	0	4	1	1
	C	3	3	2	2	2	0	6	4	0	1	1
Color pelo	Cl	2	2	1	1	2	0	4	4	0	0	0
	Cst	2	2	2	2	0	4	0	0	4	0	0
	N	1	1	0	1	1	1	1	0	0	2	0
	R	2	0	1	0	1	1	1	0	0	0	2

Métrica para las coordenadas

De acuerdo a la tabla de BURT anterior y haber usado la función de MATLAB *analcorr2.m*, se observaron, entre otros estos mapas que se logran percibir algunas combinaciones. Más detalle se puede observar en la segunda

práctica. Sin embargo, el juicio sobre cuáles combinaciones sean apropiadas depende de aspectos: un buen valosr re la chi-cuadrado que así lo justifique y el criterio que se siga para realizar las asociaciones.