2 Contraste de independencia

2.1 Independencia entre variables cualitativas

Consideremos dos variables cualitativas X e Y con I y J modalidades cada una respectivamente, y sea N_{IJ} la tabla de contingencia asociada a la distribución conjunta de ambas variables, notaremos por n_{ij} la frecuencia absoluta correspondiente a la casilla (i,j). Consideremos también la tabla de frecuencias F_{IJ} obtenida dividiendo cada n_{ij} por el total $n_{..}$ es decir $f_{ij} = \frac{n_{ij}}{n_{..}}$

	y_1	$\dots$	y_j	$\dots$	$y_{j'}$	$\dots$	y_J	
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	$n_{1j'}$	$\dots$	n_{1J}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	$n_{ij'}$	$\dots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_{i'}$	$n_{i'1}$	$\dots$	$n_{i'j}$	$\dots$	$n_{i'j'}$	$\dots$	$n_{i'J}$	$n_{i'.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_I	n_{I1}	$\dots$	n_{Ij}	$\dots$	$n_{Ij'}$	$\dots$	n_{IJ}	$n_{I.}$
	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.j'}$	$\dots$	$n_{.J}$	$n_{..}$

	y_1	$\dots$	y_j	$\dots$	$y_{j'}$	$\dots$	y_J	
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	$f_{1j'}$	$\dots$	f_{1J}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	$f_{ij'}$	$\dots$	f_{iJ}	$f_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_{i'}$	$f_{i'1}$	$\dots$	$f_{i'j}$	$\dots$	$f_{i'j'}$	$\dots$	$f_{i'J}$	$f_{i'.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_I	f_{I1}	$\dots$	f_{Ij}	$\dots$	$f_{Ij'}$	$\dots$	f_{IJ}	$f_{I.}$
	$f_{.1}$	$\dots$	$f_{.j}$	$\dots$	$f_{.j'}$	$\dots$	$n_{.J}$	1

En esta última tabla tenemos $1+1+J+I$ tablas unidimensionales, que corresponden a las 2 distribuciones marginales de X y de Y, a las J distribuciones condicionadas de X para cada valor de Y y a las I distribuciones condicionadas de Y para cada valor de X. Las distribuciones condicionadas las podemos escribir:

$$X|y = y_j \longrightarrow \left\{ f_{i|j} = \frac{f_{ij}}{f_{.j}} \; ; \; i = 1, \dots, I \right\} \text{ donde } j = 1, \dots, J \qquad \text{perfiles columna}$$

$$Y|x = x_i \longrightarrow \left\{ f_{j|i} = \frac{f_{ij}}{f_{i.}} \; ; \; j = 1, \dots, J \right\} \text{ donde } i = 1, \dots, I \qquad \text{perfiles fila}$$

2.1.1 Concepto de independencia

Se dice que X es independiente de Y, si la distribución según el carácter X de los individuos que poseen la modalidad y_j, es la misma cualquiera que sea y_j, es decir las distribuciones condicionadas de X para cada valor $y_j \;\; j = 1, \dots, J$, son idénticas, o bien que $\frac{f_{ij}}{f_{.j}} \;\; j = 1, \dots, J$ no es función de j.
Vamos a ver que desde el punto de vista estadístico, la independencia entre variables supone la proporcionalidad entre las columnas de la tabla y comprobaremos que es un concepto simétrico, es decir que si X es independiente de Y, Y también lo es respecto de X, por lo que también se dará proporcionalidad entre las filas de la tabla.
Si X es independiente de Y, tendremos:

$$\begin{array}{ccccccccc}
\frac{f_{11}}{f_{.1}} = & \frac{f_{12}}{f_{.2}} = & \dots & = \frac{f_{1j}}{f_{.j}} = & \dots & = \frac{f_{1j'}}{f_{.j'}} = & \dots & = \frac{f_{1J}}{f_{.J}} \\
\frac{f_{21}}{f_{.1}} = & \frac{f_{22}}{f_{.2}} = & \dots & = \frac{f_{2j}}{f_{.j}} = & \dots & = \frac{f_{2j'}}{f_{.j'}} = & \dots & = \frac{f_{2J}}{f_{.J}} \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{f_{i1}}{f_{.1}} = & \frac{f_{i2}}{f_{.2}} = & \dots & = \frac{f_{ij}}{f_{.j}} = & \dots & = \frac{f_{ij'}}{f_{.j'}} = & \dots & = \frac{f_{iJ}}{f_{.J}} \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{f_{i'1}}{f_{.1}} = & \frac{f_{i'2}}{f_{.2}} = & \dots & = \frac{f_{i'j}}{f_{.j}} = & \dots & = \frac{f_{i'j'}}{f_{.j'}} = & \dots & = \frac{f_{i'J}}{f_{.J}} \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{f_{I1}}{f_{.1}} = & \frac{f_{I2}}{f_{.2}} = & \dots & = \frac{f_{Ij}}{f_{.j}} = & \dots & = \frac{f_{Ij'}}{f_{.j'}} = & \dots & = \frac{f_{IJ}}{f_{.J}}
\end{array}
\Longrightarrow \left\{ \begin{array}{c} \frac{f_{i1}}{f_{.1}} = \frac{f_{i2}}{f_{.2}} = \dots = \frac{f_{ij}}{f_{.j}} = \dots = \frac{f_{ij'}}{f_{.j'}} = \dots = \frac{f_{iJ}}{f_{.J}} \\ i = 1, 2, \dots, I \end{array} \right\}$$

Notemos que X independiente de Y implica que cada modalidad x_i de X es independiente de Y de lo que podemos deducir que las columnas j y j' son proporcionales, es decir:

$$\left.\begin{array}{rcl}
\frac{f_{1j}}{f_{.j}} = \frac{f_{1j'}}{f_{.j'}} & \Longrightarrow & \frac{f_{1j}}{f_{1j'}} = \frac{f_{.j}}{f_{.j'}} \\
\frac{f_{2j}}{f_{.j}} = \frac{f_{2j'}}{f_{.j'}} & \Longrightarrow & \frac{f_{2j}}{f_{2j'}} = \frac{f_{.j}}{f_{.j'}} \\
\vdots & \Longrightarrow & \vdots \\
\frac{f_{ij}}{f_{.j}} = \frac{f_{ij'}}{f_{.j'}} & \Longrightarrow & \frac{f_{ij}}{f_{ij'}} = \frac{f_{.j}}{f_{.j'}} \\
\vdots & \Longrightarrow & \vdots \\
\frac{f_{i'j}}{f_{.j}} = \frac{f_{i'j'}}{f_{.j'}} & \Longrightarrow & \frac{f_{i'j}}{f_{i'j'}} = \frac{f_{.j}}{f_{.j'}} \\
\vdots & \Longrightarrow & \vdots \\
\frac{f_{Ij}}{f_{.j}} = \frac{f_{Ij'}}{f_{.j'}} & \Longrightarrow & \frac{f_{Ij}}{f_{Ij'}} = \frac{f_{.j}}{f_{.j'}}
\end{array}\right\} \Longrightarrow \frac{f_{1j}}{f_{1j'}} = \frac{f_{2j}}{f_{2j'}} = \ldots = \frac{f_{ij}}{f_{ij'}} = \ldots = \frac{f_{i'j}}{f_{i'j'}} = \ldots = \frac{f_{Ij}}{f_{Ij'}} \quad (\forall j, j')$$

Por lo tanto:

$$\forall i, i', j, j' \Longrightarrow \frac{f_{ij}}{f_{ij'}} = \frac{f_{i'j}}{f_{i'j'}} \Longrightarrow \frac{n_{ij}}{n_{ij'}} = \frac{n_{i'j}}{n_{i'j'}}$$

de donde se deduce que las columnas de la tabla de frecuencias absolutas, son proporcionales en el caso de que X sea independiente de Y

Además teníamos que $\forall i = 1, 2, \ldots, I$:

$$\frac{f_{i1}}{f_{.1}} = \frac{f_{i2}}{f_{.2}} = \ldots = \frac{f_{ij}}{f_{.j}} = \ldots = \frac{f_{ij'}}{f_{.j'}} = \ldots = \frac{f_{iJ}}{f_{.J}}$$

de donde aplicando la propiedad de las fracciones de que la suma de los antecedentes dividido por la suma de los consecuentes es igual a cada una de las fracciones, obtenemos:

$$\frac{f_{i1}}{f_{.1}} = \frac{f_{i2}}{f_{.2}} = \ldots = \frac{f_{ij}}{f_{.j}} = \ldots = \frac{f_{ij'}}{f_{.j'}} = \ldots = \frac{f_{iJ}}{f_{.J}} = \frac{f_{i1} + f_{i2} + \cdots + f_{ij} + \ldots + f_{ij'} + \ldots + f_{Ij}}{f_{.1} + f_{.2} + \cdots + f_{.j} + \ldots + f_{.j'} + \ldots + f_{.j}} = f_{i.}$$

$$\Longrightarrow \left\{ f_{i|j} = \frac{f_{ij}}{f_{.j}} = f_{i.} \quad \begin{array}{c} i = 1, \ldots, I \\ j = 1, \ldots, J \end{array} \right\} \Longrightarrow \left\{ f_{ij} = f_{i.} f_{.j} \quad \begin{array}{c} i = 1, \ldots, I \\ j = 1, \ldots, J \end{array} \right\} \text{ lo que indica independencia entre } X \text{ e } Y$$

Además de $f_{ij} = f_{i.} f_{.j}$ como $f_{j|i} = \frac{f_{ij}}{f_{i.}} \Longrightarrow \{ f_{j|i} = f_{.j} \;\; i = 1, \ldots, I \}$ lo que indica independencia entre Y e X Luego también tendremos proporcionalidad entre las filas de la tabla de frecuencias.

2.1.2 Contraste de independencia χ^2

La hipótesis de independencia se puede escribir

$$H_0 : \;\; f_{ij} = f_{i.} f_{.j} \;\; \left\{ \begin{array}{l} i = 1, \ldots, I \\ j = 1, \ldots, J \end{array} \right.$$

Veamos un procedimiento para contrastar dicha hipótesis, como $f_{ij} = \frac{n_{ij}}{n_{..}}$ sabemos que el valor observado $n_{ij} = n_{..} f_{ij}$, mientras que el valor esperado bajo la hipótesis de independencia sería $e_{ij} = n_{..} f_{i.} f_{,j}$, por lo que una forma de realizar el contraste sería ver la discrepancia entre los valores observados y esperados, en este caso si sumamos las diferencias $(n_{ij} - e_{ij})$, los valores positivos se compensarían con los negativos por lo que una posible solución es elevar al cuadrado dichas diferencias $(n_{ij} - e_{ij})^2$, aún así convendría normalizar las discrepancias, calculándolas en valores relativos, $\frac{(n_{ij} - e_{ij})^2}{e_{ij}}$, lo que permite definir el estadístico χ^2:

$$\chi^2_{exp} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{..}f_{ij}-n_{..}f_{i.}f_{.j})^2}{n_{..}f_{i},f_{.j}} = n_{..}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(f_{ij}-f_{i.}f_{.j})^2}{fi.f_{.j}}$$

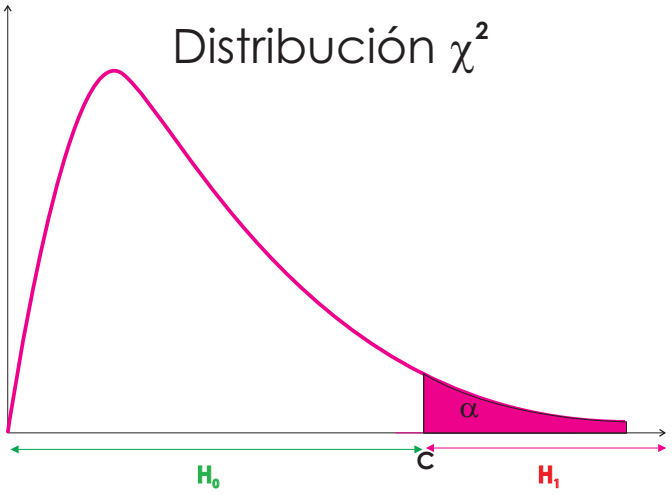
Por lo tanto siendo la cantidad χ^2_{exp} una medida de la discrepancia entre los valores observados y esperados, aceptaremos H_0 si el valor es pequeño y la hipótesis alternativa H_1 en caso de que sea grande.

Sin embargo necesitamos un criterio para decidir bajo qué valores del estadístico χ^2 aceptamos la hipótesis de independencia H_0 entre las variables. Tal criterio exige conocer la distribución de probabilidad del estadístico χ^2_{exp} y en este sentido Pearson demostró que asumiendo que las frecuencias observadas n_{ij} siguen una distribución multinomial, el estadístico χ^2 para grandes tamaños muestrales sigue una distribución χ^2 con $(I-1)\cdot(J-1)$ grados de libertad.

$$\chi^2_{exp} \rightsquigarrow \chi^2_{(I-1)\cdot(J-1)}$$

Los grados de libertad de una tabla de contingencia pueden considerarse como el número de celdas de la tabla que se pueden fijar libremente cuando se fijan los totales marginales, es decir la diferencia entre el número de casillas de la tabla y el número de restricciones impuestas: $ij - ((i-1)+(j-1)) - 1$.

Conocida la distribución del estadístico χ^2, para contrastar con un nivel de significación α la hipótesis H_0 de independencia entre X e Y, hacemos lo siguiente:
Calculamos el valor crítico C de una distribución $\chi^2_{(I-1)\cdot(J-1)}$ tal que $P[\chi^2_{(i-1)\cdot(j-1)} > C] = \alpha$.
Si el valor del estadístico $\chi^2 > C \Longrightarrow P[\chi^2_{(i-1)\cdot(j-1)} > \chi^2] < \alpha$, lo que significa que estamos ante una muestra rara y rechazaríamos H_0, aceptando H_1.
Si el valor del estadístico $\chi^2 < C \Longrightarrow P[\chi^2_{(i-1)\cdot(j-1)} > \chi^2] > \alpha$, lo que nos llevaría a aceptar H_0.



Esto mismo es lo que hacen los diferentes paquetes estadísticos a través del $p-value$, para un nivel de significación $\alpha = 0.05$ si el $p - value < 0.05$ se acepta H_1, en caso contrario si el $p - value > 0.05$, se acepta H_0 la hipótesis de independencia entre las variables.
La relación entre el estadístico χ^2_{exp}, la relación de dependencia y el $p - value$ viene a ser la siguiente:

Si χ^2_{exp} es muy grande $\Longrightarrow$ Existe dependencia entre las variables $\Longrightarrow (p - value) \longrightarrow 0$
Si χ^2_{exp} es muy pequeño $\Longrightarrow$ Existe independencia entre las variables $\Longrightarrow (p - value) \longrightarrow 1$

2.1.3 χ^2 y la Inercia de la nube

$$\chi^2 = n_{..}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(f_{ij}-f_{i.}f_{.j})^2}{fi.f_{.j}} = n_{..}\sum_{i=1}^{I}f_{i.}\sum_{j=1}^{J}\frac{1}{f_{.j}}\left[\frac{f_{ij}}{f_{i.}} - f_{.j}\right]^2 = n_{..}\sum_{i=1}^{I}f_{i.}\sum_{j=1}^{J}\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}}\right)^2 = n_{..}In(N_I)$$

Por un lado tenemos los puntos $\frac{f_{ij}}{f_{i.}}$ afectados de una masa $f_{i.}$ cuyo centro de gravedad sería

$$C.G = \frac{\sum_i^I f_{i.} \frac{f_{ij}}{f_{i.}}}{\sum_i^I f_{i.}} = f_{.j}$$

es decir la suma de las masas por las coordenadas dividido por la suma de las masas, luego la expresión: $\frac{1}{f_{.j}} \left[\frac{f_{ij}}{f_{i.}} - f_{.j} \right]^2$ representaría el cuadrado de las distancias entre los puntos perfiles fila y su centro de gravedad ponderadas por el valor $\frac{1}{f_{.j}}$ y al multiplicar esta expresión por las masas $f_{i.}$ tenemos el concepto de inercia de la nube de perfiles fila (frecuencias condicionadas de $Y|X$), de aquí la relación entre el estadístico χ^2 y la inercia de la nube de puntos fila:

$$\chi^2 = n_{..} In(N_I)$$

Además si introducimos el factor de ponderación $\frac{1}{f_{.j}}$ dentro del cuadrado, obtenemos unos nuevos perfiles fila transformados $\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}$ cuyo centro de gravedad es:

$$C.G = \frac{\sum_i^I f_{i.} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}}{\sum_i^I f_{i.}} = \sqrt{f_{.j}}$$

por lo que también tendríamos el producto de las masas por el cuadrado de las distancias entre los perfiles fila transformados y su centro de gravedad, pero en este caso la distancia ya sería euclídea.

Esta expresión de una distancia ponderada nos permite definir una distancia muy usada en el análisis multivariante como es la distancia χ^2 definida de la forma siguiente:

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left[\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right]^2 = \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{i'.}\sqrt{f_{.j}}} \right)^2$$

donde convertimos la distancia ponderada en distancia euclídea al introducir el factor de ponderación $\frac{1}{f_{.j}}$ dentro del cuadrado, de esta forma la nube de perfiles fila $\frac{f_{ij}}{f_{i.}}$ con la distancia ponderada χ^2 se convierte en la nube de perfiles fila transformados $\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}$ con la distancia euclídea, donde la diferencia entre ambos tipos de perfiles solo consiste en un cambio de escala en el sentido de aumentar los valores de los perfiles fila transformados.

De la misma forma podríamos poner:

$$\chi^2 = n_{..} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - f_{i.}f_{.j})^2}{fi.f_{.j}} = n_{..} \sum_{j=1}^{J} f_{.j} \sum_{i=1}^{I} \frac{1}{f_{i.}} \left[\frac{f_{ij}}{f_{.j}} - f_{i.} \right]^2 = n_{..} \sum_{j=1}^{J} f_{.j} \sum_{i=1}^{I} \left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} - \sqrt{f_{i.}} \right)^2 = n_{..} In(N_J)$$

Lo que nos permite deducir que las inercias de la nube de puntos fila y columna son iguales $In(N_I) = In(N_J)$

Análogamente la distancia simétrica χ^2 entre 2 columnas sería:

$$d^2_{\chi^2}(j,j') = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left[\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right]^2 = \sum_{i=1}^{I} \left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} - \frac{f_{ij'}}{f_{.j'}\sqrt{f_{i.}}} \right)^2$$

2.1.4 Ejemplo

El resultado de un estudio de relación entre el dominio de la vista y el predominio de la mano viene dado en la siguiente tabla:

	Levocular	*Ambiocular*	*Dextrocular*	
Zurdo	34	62	28	124
Ambidextro	27	28	20	75
Dextro	57	105	52	214
	118	195	100	413

Vamos a calcular las siguientes tablas:

- Frecuencias relativas (f_{ij})
- Frecuencias esperadas (e_{ij})
- Perfiles fila ($\frac{f_{ij}}{f_{i.}}$)
- Perfiles columna ($\frac{f_{ij}}{f_{.j}}$)
- Perfiles fila modificados ($\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}$)
- Perfiles columna modificados ($\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}}$)
- Centros de gravedad de todos los perfiles fila y columna ($f_{.j}$, $f_{i.}$, $\sqrt{f_{.j}}$, $\sqrt{f_{i.}}$)
- El estadístico χ^2_{exp}
- El contraste de independencia
- La Inercia de la nube de puntos fila y columna
- Comparación de la inercia con el valor del estadístico χ^2_{exp}
- Representación gráfica de los distintos perfiles.

f_{ij}				$f_{i.}$
	0.082	0.150	0.067	0.299
	0.065	0.067	0.048	0.180
	0.138	0.254	0.126	0.518
$f_{.j}$	0.285	0.471	0.241	1

e_{ij}		
35.193	58.162	29.760
21.186	35.014	17.915
60.971	100.762	51.558

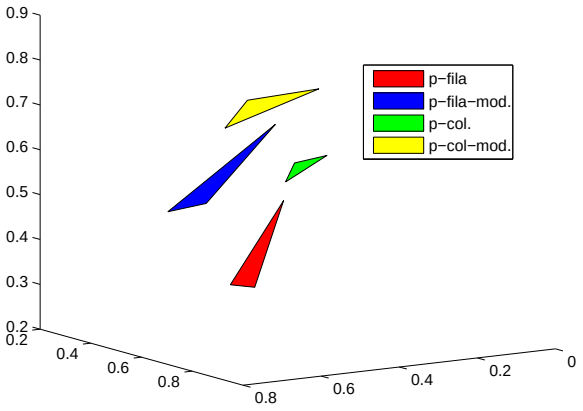
$f_{ij}/f_{i.}$		
0.274	0.501	0.224
0.361	0.372	0.266
0.266	0.490	0.246

$f_{ij}/f_{.j}$		
0.287	0.318	0.278
0.228	0.142	0.199
0.484	0.539	0.522

$f_{ij}/f_{i.}\sqrt{f_{.j}}$			
0.513	0.730	0.456	
0.676	0.542	0.541	
0.498	0.713	0.494	
0.533	0.686	0.490	$\sqrt{f_{.j}}$

$f_{ij}/f_{.j}\sqrt{f_{i.}}$			$\sqrt{f_{i.}}$
0.524	0.581	0.508	0.546
0.537	0.334	0.469	0.424
0.672	0.748	0.725	0.719

Al ser los perfiles fila y columna de este ejemplo, puntos en $\mathbb{R}^3$, podemos representarlos gráficamente y observar en la siguiente gráfica, como los perfiles fila y columna transformados, (en color azul), $\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}$, no suponen más que un cambio de escala respecto de los originales, (en color rojo), $\frac{f_{ij}}{f_{i.}}$, en el sentido de aumentar sus valores.



$$\chi^2_{exp} = \frac{(34-35.193)^2}{35.193} + \frac{(62-58.162)^2}{58.162} + \ldots\ldots \frac{(52-51.558)^2}{51.558} = 4.081$$

$\chi^2_{0.95;4} = 9.492 \Longrightarrow$ Se acepta la independencia de las variables.
De hecho se puede comprobar que existe proporcionalidad entre la 1ª y 3ª filas.

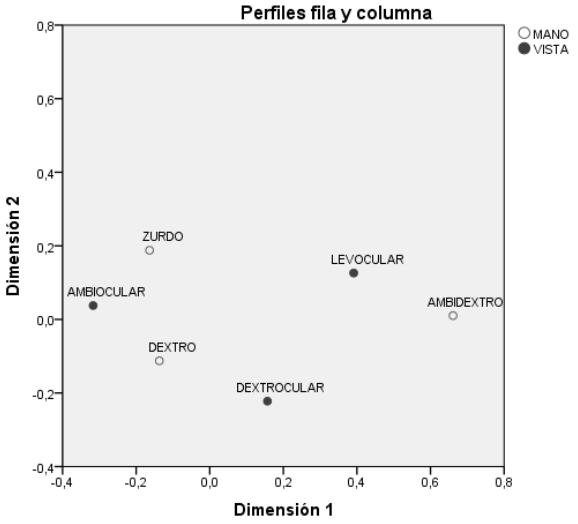
$$\text{Inercia} = \frac{\chi^2}{N} = \frac{4.081}{413} = 0.00988$$

Algunos resultados encontrados al aplicar un análisis de correspondencias a esta tabla con SPSS son:

Resumen

Dimensión	Valor propio	Inercia	Chi-cuadrado	Sig.	Proporción de inercia		Confianza para el Valor propio	
					Explicada	Acumulada	Desviación típica	Correlación 2
1	,097	,009			,970	,970	,049	-,001
2	,017	,000			,030	1,000	,048	
Total		,010	4,020	,403[a]	1,000	1,000		

a. 4 grados de libertad



Con lo que podemos comprobar el valor obtenido de la inercia, el valor del estadístico χ^2 y la aceptación de la hipótesis nula de independencia al ser el p-value mayor que 0.05. Además en la representación gráfica obtenida al aplicar la técnica, los perfiles de las 2 filas con valores proporcionales, la 1ª y 3ª, (zurdo y dextro), salen juntos, respecto al primer eje.

2.2 Tablas de contingencias tridimensionales N_{IJK}

Dentro de la tabla tridimensional tenemos:

- Tablas marginales unidimensionales:

$$n_{i..} = \sum_j \sum_k n_{ijk} \qquad n_{.j.} = \sum_i \sum_k n_{ijk} \qquad n_{..k} = \sum_i \sum_j n_{ijk}$$

- Tablas marginales bidimensionales:

$$n_{ij.} = \sum_k n_{ijk} \qquad n_{i.k} = \sum_j n_{ijk} \qquad n_{.jk} = \sum_i n_{ijk}$$

2.2.1 Contraste de independencia global o conjunto

La hipótesis nula se puede escribir: $H_0: \quad f_{ijk} = f_{i..}f_{.j.}f_{..k}$ siendo los valores esperados en caso de independencia: $e_{ijk} = n_{...}f_{i..}f_{.j.}f_{..k}$ por lo que el estadístico χ^2 sería:

$$\chi^2_{exp} = \sum_i \sum_j \sum_k \frac{(n_{ijk} - e_{ijk})^2}{e_{ijk}} \rightsquigarrow \chi^2_{ijk-[(i-1)+(j-1)+(k-1)]-1} = \chi^2_{ijk-i-j-k+2}$$

Como en el caso bidimensional, si el valor del estadístico χ^2 es inferior al valor crítico, se acepta H_0 en caso contrario se rechaza la hipótesis nula.

2.2.2 Ejemplo

En Andalucía se quiere contrastar con un nivel de significación $\alpha = 0.05$, la H_0 de independencia global entre las variables, sexo, profesión y utilización de instalaciones deportivas de los municipios de la Comunidad. Para ello se toma una muestra de 854 individuos cuyas observaciones figuran en la siguiente tabla:

Sexo (i)	Profesión (j)	Utilización (k)		
		Usuario	No usuario	Total
Varón	Liberal	62	138	200
	No liberal	176	56	232
	Total	238	194	432
Mujer	Liberal	138	60	198
	No liberal	52	172	224
	Total	190	232	422

- Frecuencias observadas:
  $n_{1..} = 200 + 232 = 432 \qquad n_{2..} = 198 + 224 = 422 \qquad n_{.1.} = 200 + 198 = 398 \qquad n_{.2.} = 232 + 224 = 456$
  $n_{..1} = 238 + 190 = 428 \qquad n_{..2} = 194 + 232 = 426$

- Frecuencias esperadas:
  $e_{111} = 101 \qquad e_{112} = 100 \qquad e_{121} = 116 \qquad e_{122} = 115 \qquad e_{211} = 99 \qquad e_{212} = 98 \qquad e_{221} = 113$
  $e_{222} = 112$

- $\chi^2 = 15.06 + 31.03 + 15.36 + 32.93 + 14.44 + 32.07 + 14.73 + 32.14 + = 185.96 > \chi^2_{0.05,4} = 9.488.$

  Por lo tanto se rechaza H_0, es decir aceptamos la existencia de asociación significativa conjunta de las 3 variables.

  Esta misma información se podría haber dado a través de la siguiente tabla:

Usuario

Profesión

Sexo	Liberal	No liberal	Total
Varón	62	176	238
Mujer	138	52	190
Total	200	228	428

No Usuario

Profesión

Sexo	Liberal	No liberal	Total
Varón	138	56	194
Mujer	60	172	232
Total	198	228	426

A la vista de las tablas vamos a contestar a las siguientes preguntas:

1. De entre los varones de profesión liberal, ¿qué % son usuarios de las instalaciones deportivas?
2. De entre los usuarios de instalaciones deportivas, ¿qué % son varones de profesión liberal?

3. De entre los usuarios de instalaciones deportivas, ¿qué % son de profesión liberal?
4. De entre los usuarios de instalaciones deportivas, ¿qué % son varones?
5. ¿Qué % hay de mujeres de profesión liberal?
6. ¿Qué % hay de mujeres?
7. ¿Qué % hay de profesión no liberal?
8. ¿Qué % hay de usuarios de instalaciones deportivas de profesión no liberal?
9. ¿Qué % hay de mujeres no usuarios de instalaciones deportivas y de profesión no liberal?
10. De entre las mujeres de profesión liberal, ¿qué % no son usuarias de instalaciones deportivas?
11. De entre las mujeres no usuarias de instalaciones deportivas ¿qué % son de profesión liberal?
12. De entre las mujeres, ¿qué % son de profesión liberal?
13. De entre los varones, ¿qué % son usuarios de instalaciones deportivas?
14. ¿Qué % hay de varones?
15. ¿Qué % hay de usuarios de instalaciones deportivas?

(62/200) , (62/428) , (200/428) , (238/428) , (198/854) , (422/854) , (456/854) , (228/854) , (172/854) , (60/198) (60/232) , (198/422) , (238/432) , (432/854) , (428/854)

2.2.3 Contraste de independencia parcial

El rechazo de la hipótesis de independencia global H_0, no implica que exista asociación entre las 3 variables, sino que puede darse también porque exista:

- Independencia parcial: Hay asociación entre dos de las variables y la 3ª es independiente de ellas.
- Independencia condicional: Dos de las variables son independientes entre sí, para cada nivel de la 3ª, pero pueden estar asociadas cada una de ellas con la 3ª variable.

En una tabla de 3 dimensiones existen 3 posibles tipos de independencia parcial:

$H_{0(1)} \longrightarrow f_{ijk} = f_{i..}f_{.jk}$ (1ª independiente de la 2ª y 3ª mientras la 2ª no es independiente de la 3ª)

$H_{0(2)} \longrightarrow f_{ijk} = f_{.j.}f_{i.k}$ (2ª independiente de la 1ª y 3ª mientras la 1ª no es independiente de la 3ª)

$H_{0(3)} \longrightarrow f_{ijk} = f_{..k}f_{ij.}$ (3ª independiente de la 1ª y 2ª mientras la 1ª no es independiente de la 2ª)

Los valores esperados en caso de independencia son respectivamente:
$e_{ijk} = n_{...}f_{i..}f_{.jk}$, $e_{ijk} = n_{...}f_{.j.}f_{i.k}$, $e_{ijk} = n_{...}f_{..k}f_{ij.}$ por lo que el estadístico χ^2 sería:

$$H_{0(1)}: \quad f_{ijk} = f_{i..}f_{.jk} \qquad \chi^2_{exp} = \sum_i \sum_j \sum_k \frac{(n_{ijk} - e_{ijk})^2}{e_{ijk}} \rightsquigarrow \chi^2_{ijk-[(i-1)+(jk-1)]-1} = \chi^2_{ijk-i-jk+1}$$

$$H_{0(2)}: \quad f_{ijk} = f_{.j.}f_{i.k} \qquad \chi^2_{exp} = \sum_i \sum_j \sum_k \frac{(n_{ijk} - e_{ijk})^2}{e_{ijk}} \rightsquigarrow \chi^2_{ijk-[(j-1)+(ik-1)]-1} = \chi^2_{ijk-j-ik+1}$$

$$H_{0(3)}: \quad f_{ijk} = f_{..k}f_{ij.} \qquad \chi^2_{exp} = \sum_i \sum_j \sum_k \frac{(n_{ijk} - e_{ijk})^2}{e_{ijk}} \rightsquigarrow \chi^2_{ijk-[(k-1)+(ij-1)]-1} = \chi^2_{ijk-k-ij+1}$$

Como en el caso bidimensional, si el valor del estadístico χ^2 es inferior al valor crítico, se acepta la hipótesis nula, en caso contrario se rechaza.

2.2.4 Ejemplo

Se desea estudiar si existe relación entre el nivel de estudios (básico-medio, superior), el medio de comunicación social (prensa, radio y tv) y el sexo de las personas. Para ello se realiza una encuesta a un grupo de 600 personas cuyos resultados son:

Nivel de estudios	Medio de comunicación	Sexo		
		Hombre	Mujer	Total
Básico-medio	Prensa	30	25	55
	Radio	40	35	75
	Tv	90	75	165
	Total	160	135	295
Superior	Prensa	80	65	145
	Radio	50	55	105
	Tv	20	35	55
	Total	150	155	305

En este caso $\chi^2 = 105.8 > \chi^2_{0.05,4} = 9.488$ por lo que rechazamos H_0, es decir los 3 factores no son independientes conjuntamente.
Vamos a plantearnos la hipótesis de independencia parcial, y de entre las 3 posibilidades nos interesa contrastar si el nivel de estudios y el medio de comunicación, son independientes del sexo, existiendo algún tipo de asociación entre las dos primeras.

- La hipótesis es H_0: $f_{ijk} = f_{..k} f_{ij.}$
- $e_{111} = 28.4$ $e_{112} = 26.6$ $e_{121} = 38.75$ $e_{122} = 36.25$ $e_{131} = 85.25$ $e_{132} = 79.75$
  $e_{211} = 74.9$ $e_{212} = 70.1$ $e_{221} = 54.25$ $e_{222} = 50.77$ $e_{231} = 28.4$ $e_{232} = 26.6$
- $\chi^2_{exp} = 7.362 < \chi^2_{2\cdot3\cdot2-2-6+1} = \chi^2_{0.05,5} = 11.07$.

Luego se acepta H_0, es decir el nivel de estudios y la preferencia por un medio de comunicación social, son independientes del sexo.

Esta misma información se podría haber dado a través de la siguiente tabla:

	Hombre			
	Medios de Comunicación			
Nivel de estudios	Prensa	Radio	Tv	Total
Básico-medio	30	40	90	160
Superior	80	50	20	150
Total	110	90	110	310

	Mujer			
	Medios de Comunicación			
	Prensa	Radio	Tv	Total
Básico-medio	25	35	75	135
Superior	65	55	35	155
Total	90	90	110	290

A la vista de las tablas vamos a contestar a las siguientes preguntas:

1. ¿Qué % de hombres, ven la tv y tienen nivel de estudios superior?
2. De entre los hombres, ¿qué % ven la tv y tienen nivel de estudios superior?
3. De entre los hombres que ven la tv, ¿qué % tienen nivel de estudios superior?

4. ¿Qué % de personas con estudios básicos, leen la prensa?
5. ¿Qué % de personas que escuchan la radio son mujeres?
6. ¿Qué % de personas que escuchan la radio y son mujeres, tienen nivel de estudios superior?
7. ¿Qué % de mujeres que escuchan la radio, tienen nivel de estudios básico?
8. ¿Qué % de mujeres con nivel de estudios superior leen la prensa?
9. ¿Qué % de personas ve la tv?
10. ¿Qué % de mujeres escuchan la radio y tienen nivel de estudios básico?
11. ¿Qué % de personas que escuchan la radio y tienen nivel de estudios básico, son mujeres?
12. ¿Qué % de mujeres escuchan la radio?
13. ¿Qué % de mujeres escuchan la radio y tienen nivel de estudios superior?

(20/600) , (20/310) , (20/110) , (55/295) , (90/180) , (55/90) , (35/90) , (65/155) , (220/600) , (35/600)
(35/75) , (90/600) , (55/600)

2.2.5 Contraste de independencia condicional

En una tabla tridimensional hay 3 supuestos de independencia condicional

$H_{0(23)} \longrightarrow f_{ijk} = f_{ij.}f_{i.k}$ (la 2ª y 3ª son independientes entre sí, para cada nivel de la 1ª variable)

$H_{0(13)} \longrightarrow f_{ijk} = f_{ij.}f_{.jk}$ (la 1ª y 3ª son independientes entre sí, para cada nivel de la 2ª variable)

$H_{0(12)} \longrightarrow f_{ijk} = f_{i.k}f_{.jk}$ (la 1ª y 2ª son independientes entre sí, para cada nivel de la 3ª variable)

Los valores esperados en caso de independencia son respectivamente:

$$e_{ijk} = \frac{n_{ij.} \cdot n_{i.k}}{n_{i..}} \qquad e_{ijk} = \frac{n_{ij.} \cdot n_{.jk}}{n_{.j.}} \qquad e_{ijk} = \frac{n_{i.k} \cdot n_{.jk}}{n_{..k}}$$

Por lo que el estadístico χ^2 sería:

$$H_{0(23)}: \quad f_{ijk} = f_{ij.}f_{i.k} \qquad \chi^2_{exp} \rightsquigarrow \chi^2_{ijk-[(ij-1)+(jk-1)]-1} = \chi^2_{ijk-ij-ik+1}$$
$$H_{0(13)}: \quad f_{ijk} = f_{ij.}f_{.jk} \qquad \chi^2_{exp} \rightsquigarrow \chi^2_{ijk-[(ij-1)+(ik-1)]-1} = \chi^2_{ijk-ij-jk+1}$$
$$H_{0(12)}: \quad f_{ijk} = f_{i.k}f_{.jk} \qquad \chi^2_{exp} \rightsquigarrow \chi^2_{ijk-[(ik-1)+(ij-1)]-1} = \chi^2_{ijk-ik-jk+1}$$

2.2.6 Relaciones de orden dos y superiores

En las tablas tridimensionales, podríamos encontrarnos con que la asociación entre 2 de las variables, difieran en grado o en dirección, para distintas categorías de la 3ª variable, estas relaciones se llaman interacciones de 2º orden y se estudian a través de los modelos logarítmicos lineales.