Selección de Atributos y su Relación con el Contraste de Hipótesis. Errores y potencia de un contraste. P-Valores

12 de abril de 2017

0.1. Contrastes de hipótesis.

Muchas investigaciones estadísticas tienen por objeto *contrastar una hipótesis*: dos muestras proceden de la misma población, la vida media de una lámpara son 2.000 horas, la proporción de fumadores es igual en hombres y mujeres, etc.

Una hipótesis se contrasta comparando sus predicciones con la realidad (la muestra). Si ambas coinciden dentro de un margen de error admisible, se acepta la hipótesis, en caso contrario se rechaza en favor de la hipótesis alternativa.

La metodología utilizada en situaciones de incertidumbre cuando las predicciones se hagan en probabilidad constituye la teoría estadística del contraste de hipótesis.

Una *hipótesis estadística* es una suposición que determina total o parcialmente la distribución de una o varias variables aleatorias. Estas hipótesis pueden ser variadas pero la más frecuente es la que especifica un valor o rango de valores para los parámetros que está muy ligada al *contraste de hipótesis*;

El problema se plantea en los siguientes términos:

Se trata de ver si, como consecuencia de los valores muestrales obtenidos - $\vec{x} = (x_1, ..., x_n)$ - se puede aceptar la hipótesis de que la muestra procede de una población $F(x;\theta)$ en que el parámetro toma un conjunto de valores $\theta \in \Theta_0$ o rechazarla en favor de la hipótesis alternativa $\theta \in \Theta - \Theta_0$ donde Θ es el espacio paramétrico. Cuando Θ_0 se reduce a un punto, se habla de *hipótesis simple*, en caso contrario cuando es un conjunto se habla de *hipótesis compuesta.*

Un *contraste de la hipótesis nula*

$$\mathbf{H_0}: \ \theta \in \mathbf{\Theta_0} \tag{1}$$

frente a la *alternativa*

$$\mathbf{H_1}: \ \theta \in \mathbf{\Theta} - \mathbf{\Theta_0} \tag{2}$$

consiste en *clasificar los puntos del espacio muestral en dos regiones complementarias y excluyentes*:

1. una de ***aceptación*** $\mathbf{S_0}$ y
2. otra ***crítica*** o de rechazo $\mathbf{S_1}$

de modo que si $\vec{x} \in \begin{cases} S_0 \rightarrow \text{se acepta } \mathrm{H}_0 \\ S_1 \rightarrow \text{se rechaza } \mathrm{H}_0 \end{cases}$.

Con frecuencia la comprobación se hace de un modo indirecto a través de un estadístico $T(\vec{x})$.

En el caso de selección de atributos nos enfrentamos con *dos clases* a las que pueden pertenecer los elementos de la población (tuits de odio y de no odio) y una serie de *atributos* de cada tuit (p.e. términos) que aparecen con una frecuencia determinada en cada clase. Con estos valores construimos la denominada *tabla de constingencia*, cuando la característica no es cuantitativa.

Clase	C1	C2	
Término 1	n_{11}	n_{12}	$\mathbf{n_{1\bullet}}$
Término 2	n_{21}	n_{22}	$\mathbf{n_{2\bullet}}$
.....	...	...	
Término m	n_{m1}	n_{m2}	$\mathbf{n_{m\bullet}}$
	$\mathbf{n_{\bullet 1}}$	$\mathbf{n_{\bullet 2}}$	$\mathbf{n}$

- n_{tc} es el número de veces que aparece el término t en los documentos de la clase c.
- $n_{t\bullet}$ es el número de veces que aparece el término t en *todos los documentos.*
- $n_{\bullet c}$ es el número de veces que aparecen *todos los términos* en la clase c.
- n es el número de veces que aparecen todos los términos en el corpus.

Es claro que si hay términos (gay, puto, moro, ...) más frecuentes en una clase que en otra - p.e. el término *puto* más frecuente en odio - se verificará que la frecuencia relativa de tal término será mayor en la clase de odio que en la otra, es decir:

$$\mathbf{f}_{puto,odio} = \frac{n_{puto,odio}}{n_{\bullet,odio}} > \mathbf{f}_{puto,neutro} = \frac{n_{puto,neutro}}{n_{\bullet,neutro}}$$

En nuestro caso, suponemos que la distribución conjunta del número de veces que aparece un térrmino t_i en la clase C_j será una multinomial de modo que:

$$P(n_{1j}, ..., n_{mj}) = \frac{n!}{n_{1j}!...n_{mj}!} (p_{1j})^{n_{ij}} ... (p_{mj})^{n_{mj}} \quad j = 1, 2$$

y *tratamos de averiguar si, como consecuencia de los valores muestrales obtenidos* - $\vec{x} = (n_{11}, n_{12}, ..., n_{m2})$ - *se puede aceptar la hipótesis de que la muestra procede de una población* $F(x; \vec{\theta})$ *en que el parámetro toma un conjunto de valores*[1] $\vec{\theta} \in \Theta_0 = (p_1, ..., p_m)$ *o rechazarla en favor de la hipótesis alternativa* $\vec{\theta} \in \Theta - \Theta_0 \neq (p_1, ..., p_m)$.

Utilizaremos la siguiente *medida de discrepancia* entre valores teóricos y muestra:

$$\chi^2_{(m-1)(c-1)} = \sum_{i=1}^{m} \sum_{j=1}^{2} \frac{(n_{ij} - t_{ij})^2}{t_{ij}} \tag{3}$$

donde n_{ij} es la frecuencia observada yt_{ij}la teórica si fuera independiente de la clase a que pertenece en cuyo caso $t_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ puesto que, por la independencia,

$$\mathbf{f}(\textbf{término}, \textbf{clase}) = f(términol|clase) \cdot f(clase) = \mathbf{f}(\textbf{término}) \cdot \mathbf{f}(\textbf{clase}) = \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n}$$

Profundizando algo más en las hipótesis,

1. Hipótesis nula H_0: es la que contrastamos y se mantendrá a menos que los datos la rechacen. La única forma de confirmarla es observando la totalidad de la población, pero, sin embargo, puede ser falsada por los datos experimentales. La metodología empleada tiende a favorecer esta hipótesis. Si el problema enfrenta a dos hipótesis que, a priori, se juzgan equivalentes, la metodología es discutible y, si las consecuencias de los errores son cuantificables, entramos en la *teoría de la decisión.*
2. Hipótesis alternativa H_1: es la que aceptamos implícitamente cuando rechazamos H_0.

Los tipos de hipótesis más comunes son:

H_0	H_1
$\theta = \theta_0$ simple	$\theta \neq \theta_0$ bilateral
$\theta = \theta_0$ simple	$\theta < \theta_0$ $\theta > \theta_0$ unilaterales

[1] *O sea que* $p_{i1} = p_{i2}\ \forall i = 1, ..., m$

0.2. Errores y potencia de un contraste.

Al realizar un contraste de hipótesis hay cuatro situaciones posible que se muestran en la tabla, similar a la *matriz de confusión*:

		Decisión	
		Aceptar H_0	Rechazar H_0
Hipótesis Cierta	H_0	Correcto	***Error tipo I***
	H_1	***Error tipo II***	Correcto

en que se ve que existen dos tipos de errores posibles:

1. Rechazar la hipótesis nula, cuando esta es cierta (*Falso Negativo*) . A esto se le denomina *error de tipo I*. De este error se deriva **el *nivel de significación***[2] de un contraste que se define como la *probabilidad de cometer un error de tipo I*, es decir.. La probabilidad de que este suceso ocurra, se denomina *nivel de significación*, y se representa con la letra α. Típicamente se elige un valor pequeño, 5 % o 1 %. Esto puede expresarse como,
$$\mathbf{P(S_1/H_0)} = \alpha = \mathbf{err_I} : \text{(o nivel de significación)} \tag{4}$$

2. No rechazar la hipótesis nula, cuando esta es falsa (*Falso Positivo*). Lo denominamos *error de tipo II*. Esto puede escribirse como
$$\mathbf{P(S_0/H_1)} = \beta = \mathbf{err_{II}} \tag{5}$$
y es equivalente a que su probabilidad complementariac la de rechazar H_0 cuando es flasa (o aceptar H_1 cuando es cierta), sea:
$$\mathbf{1} - \beta = \mathbf{1} - \mathbf{P(S_0/H_1)} = \mathbf{P(S_1/H_1)} \tag{6}$$

Cuando es necesario diseñar un contraste de hipótesis, sería deseable hacerlo de tal manera que las probabilidades de ambos tipos de error fueran lo más pequeñas posible. Sin embargo, con una muestra de tamaño prefijado, disminuir la probabilidad del error de tipo I, α, conduce a incrementar la probabilidad del error de tipo II, β.

Podemos definir la *curva característica del contraste* como la variación del error de tipo II con el verdadero valor del parámetro:
$$\beta(\theta) = \mathbf{P(S_0}/\theta)$$

Como puede apreciarse fácilmente, si la hipótesis nula es *simple* $\theta = \theta_0$, se verifica que
$$\beta(\theta_0) = \mathbf{P(S_0}/\theta_0) = \alpha$$

De modo similar, definiremos la ***potencia del contraste*** como el valor de l*a probabilidad de rechazar H_0 para un valor determinado del parámetro.*

$$\mathbf{Pot}(\theta) = \mathbf{1} - \beta(\theta) = \mathbf{P(S_1}/\theta) \tag{7}$$

La figura (1) muestra dos gráficos en que aparecen las curvas características y de potencia de dos contrastes sobre el valor de la media θ de una distribución normal.

1. El gráfico superior se refiere a la contrastación de la hipótesis *simple* $\mathbf{H_0} : \theta = 0$ frente a $\mathbf{H_1} : \theta \neq 0$
2. El inferior contrasta la hipótesis unilateral *compuesta* $\mathbf{H_0} : \theta \leq 0$ frente a $\mathbf{H_1} : \theta > 0$. En este caso, el campo de variación de θ para H_1 se limita al semieje positivo.

[2] a su complementario - (**1-**α) - se le llama ***nivel de confianza.***

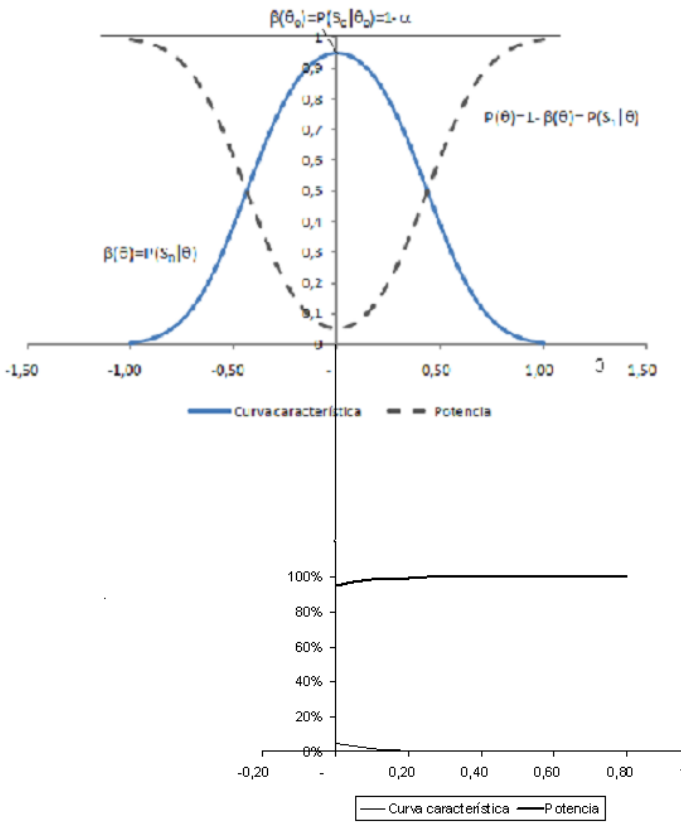


Figura 1:

0.3. Hipótesis simples.

El concepto de potencia nos permite valorar cual entre dos contrastes con la misma probabilidad de error de tipo I, α, es preferible. Si se trata de contrastar dos hipótesis simples sobre un parámetro desconocido, θ, del tipo:

$$\left.\begin{array}{l}\theta = \theta_0 \\ \theta = \theta_1\end{array}\right\}$$

escogeríamos de entre todos los contrastes posibles con α prefijado *aquel que tiene mayor potencia*, esto es, menor probabilidad β de incurrir en el error de tipo II.

0.4. Contrastes de significación, p-valor.

La metodología expuesta (de *Fisher*) consiste en:

1. Definir las *hipótesis nula y alternativa*:

$$\mathbf{H_0}: \ \theta \in \mathbf{\Theta_0}$$

$$\mathbf{H_1}: \ \theta \in \mathbf{\Theta_1}$$

2. Definir una *medida de discrepancia entre los datos muestrales $\vec{x}$ y la hipótesis nula.* En contrastes paramétricos, esta medida de discrepancia puede expresarse como función de los valores $\theta_0 \in \Theta_0$ especificados para el parámetro y el valor estimado en la muestra $\hat{\theta}$: $\mathbf{d}(\theta_0; \hat{\theta})$. Esta discrepancia debe tener una *distribución conocida cuando H_0 es cierta* $F(d/H_0)$, siendo el caso más común que la hipótesis nula sea simple $\theta = \theta_0$. Las medidas de discrepancia que pueden utilizarse son variadas:

   - desviaciones cuadráticas: $d(\theta_0; \hat{\theta}) = (\hat{\theta} - \theta_0)^2$
   - discrepancia tipificada (adimensional): $d(\theta_0; \hat{\theta}) = \dfrac{\hat{\theta} - \theta_0}{\sqrt{V(\hat{\theta})}}$
   - razón de verosimilitudes $\dfrac{l(\hat{\theta})}{l(\theta_0)}$, en cuyo caso la discrepancia no se basa directamente en los valores del parámetro, sino en los de las funciones de verosimilitud como veremos más adelante.

3. Fijado el nivel de significación α, se determina la *región crítica* como aquella en que $\mathbf{P}[\mathbf{d}(\theta_0; \hat{\theta}) > \mathbf{d}_\alpha / \mathbf{H_0}] = \alpha \rightarrow \mathbf{S_1} = \{\vec{\mathbf{x}} | \mathbf{d}(\theta_0; \hat{\theta}) > \mathbf{d}_\alpha\}$

4. Por último, se estima $\hat{\theta}(\vec{x})$ y

   - si $d(\theta_0; \hat{\theta}) \leq d_\alpha$ se acepta H_0
   - si $d(\theta_0; \hat{\theta}) > d_\alpha$ se rechaza H_0

p-valor.

Este modo de actuar - seleccionar una región de rechazo *mediante el nivel de significación* - está sujeto a dos críticas principales:

1. El resultado del test puede depender del nivel de *significación elegido.* Como se ve en la figura (2) el valor muestral obtenido$\hat{\theta}(\vec{x})$ puede rechazar la hipótesis nula con el nivel α_1 y aceptarla con α_2.$\begin{cases} d(\theta_0; \hat{\theta}) > d_{\alpha_1} \\ d(\theta_0; \hat{\theta}) \leq d_{\alpha_2} \end{cases}$

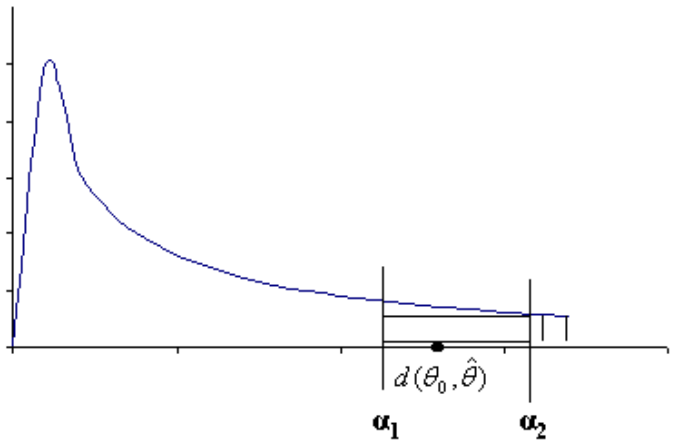


Figura 2:

2. Si únicamente damos el resultado del *test,* esto no nos permite calibrar el grado de evidencia que la muestra $\vec{x}$ presenta contra la hipótesis nula. En la figura (3)puede observarse que tanto el resultado $d(\theta_0; \hat{\theta}_1)$ como el $d(\theta_0; \hat{\theta}_2)$ rechazan la hipótesis nula, pero el primero se encuentra muy cerca del límite.

Por esta razón, se utilizan los llamados *p-valores* (o niveles críticos) definidos de la siguiente manera:

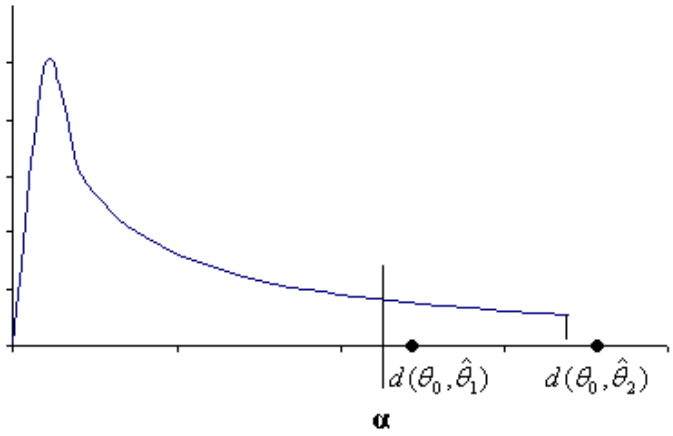


Figura 3:

$$p\text{-}valor = P[d > \hat{d}/H_0] = 1 - F(\hat{d}) \ ,$$

es decir, ***la probabilidad de que se obtenga una discrepancia superior a la observada en la muestra cuando H_0 es verdadera*** *(fig. 4)*. Por lo tanto, los p-valores *no se fijan a priori, sino que se obtienen de la muestra*. Es evidente que cuanto mayor sea el p-valor más plausible es la hipótesis nula.

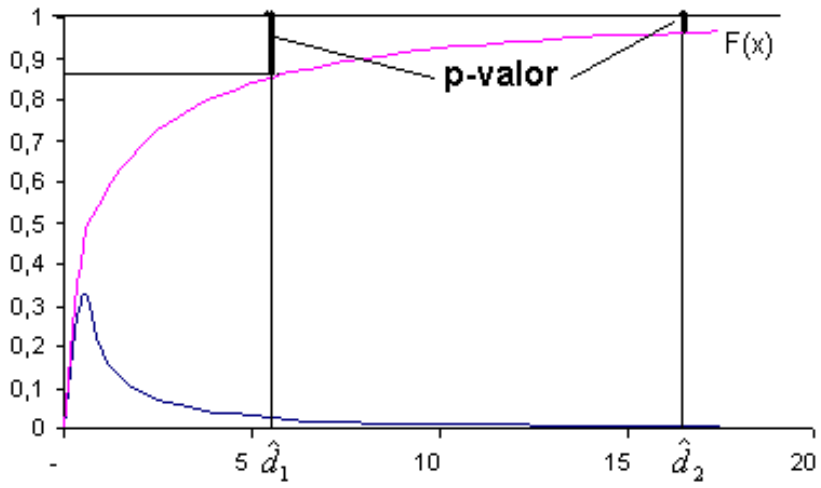
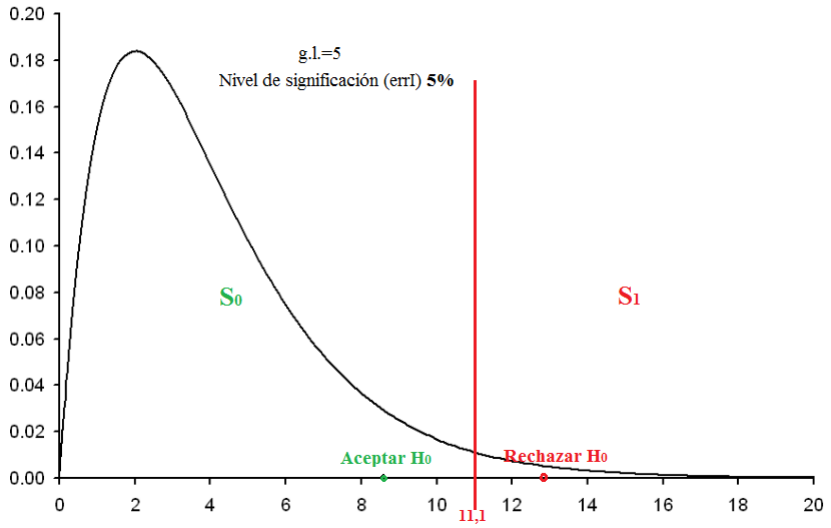


Figura 4:

Como puede apreciarse el p-valor de $\hat{d}_1$ es mayor que el de $\hat{d}_2$ por lo que el primer valor es más creíble que el segundo.

P.e. en el caso de la selección de atributos con m=6 y dos clases tendríamos una χ^2 con (6-1)(2-1)=5 g.l. Si la muestra proporciona un valor de $\chi^2 \leq 11,1$ aceptamos la hipótesis, si no la rechazamos.



Normalmente deseamos limitar el número de atributos y, si fijamos el valor del nivel de significación α podemos elegir entre todos aquellos en que χ^2<11,1. Sin embargo, esto puede hacernos preferir atributos menos influyentes sobre otros que tienen más peso.

Si nos fijamos en los p-valores de tres atributos de los que hemos de seleccionar dos (Fig. 5) parece lógico preferir los que tienen p-valores más elevados (70 % y 60 %).

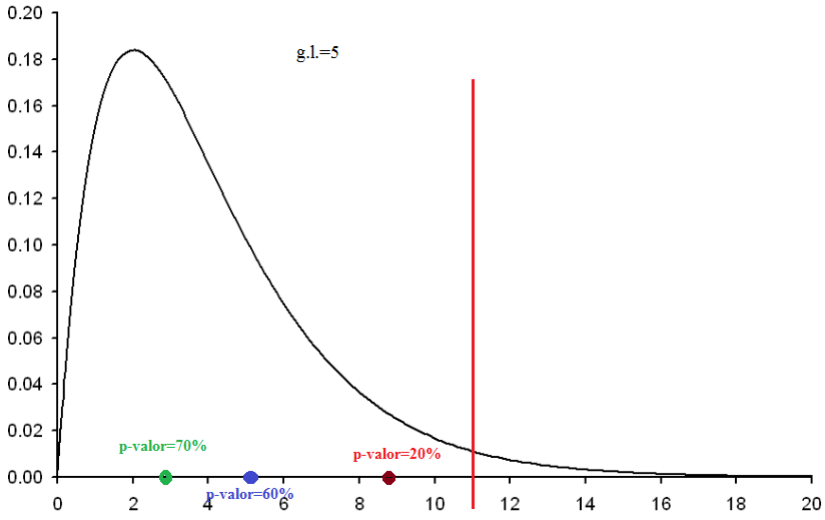


Figura 5:

0.5. Ejemplo.

Doce individuos se clasificaron según manifestaran deseo o no de ver una final de la Europa League:

Tabla de contingencia desea ver partido * SEXO

Clase	Hombre	Mujer	
Sí	6	1	*7*
No	1	4	*5*
	7	*5*	**12**

Ambas variables tienen la distribución conjunta - $f(fútbol, sexo)$- mostrada en la tabla:

Clase	Hombre	Mujer	
Sí	0,5	0,083	*0,583*
No	0,083	0,333	0,413
	0,583	*0,412*	**1**

Las distribuciones condicionadas por el sexo serán las siguientes:

$f(fútbol|sexo) = \dfrac{f(fútbol, sexo)}{f(sexo)}$ que nos lleva a

Clase	f(fútbol|Hombre)	f(fútbol|Mujer)
Sí	$\frac{0,5}{0,583} = 0,857$	$\frac{0,083}{0,412} = 0,20$
No	$\frac{0,083}{0,583} = 0,142$	$\frac{0,333}{0,412} = 0,80$
	1	**1**

Sabemos que si ambas variables - fútbol, sexo - fuesen *independientes* se verificaría que $f(fútbol|Hombre) = f(fútbol|Mujer) = f(fútbol)$ y, por tanto, esta es la hipótesis que contrastaremos: dado que la cariavle es dicotómica, la variable número de personas de sexo s que quieren asistir a la final seguirán binomiales $B(n_s, p_s)$ y contrastaremos la hipótesis:

$$\mathbf{H_0}: \mathbf{p_H = p_M}$$

$$\mathbf{H_1}: \mathbf{p_H \neq p_M}$$

Para ello se utiliza la siguiente *medida de discrepancia* entre vaklores teóricos y muestra:

$$\chi^2_{(m-1)(c-1)} = \sum_{i=1}^{m}\sum_{j=1}^{c} \frac{(n_{ij} - t_{ij})^2}{t_{ij}} \tag{8}$$

donde n_{ij} es la frecuencia observada yt_{ij}la teórica si fuera independiente de la clase a que pertenece en cuyo caso $t_{ij} = \dfrac{n_{i\bullet} n_{\bullet j}}{n}$ puesto que, por la independencia,

$$\mathbf{f(fútbol, sexo)} = f(fútbol|sexo) \cdot f(sexo) = \mathbf{f(fútbol)} \cdot \mathbf{f(sexo)} = \left(\frac{n_{i\bullet}}{n}\right) \cdot \left(\frac{n_{\bullet j}}{n}\right)$$

La variable (8) sigue una Chi cuadrado con (m-1)(c-1) g.l.

Por tanto, según hemos dicho la medida de discrepancia es una χ^2_1. Los cuadros siguientes muestran los valores muestrales y los teóricos:

Muestra

	H	M	
Si	6	1	7
No	1	4	5
	7	5	**12**

Teórico

	H	M	
Si	4,1	2,92	7
No	2,9	2,08	5
	7	5	12

de donde se obtiene un valor de ls discrepancia $\chi^2 = 5,182$. Dado que para un nivel de significación de α=5 % $\chi^2_\alpha = 3,84$ esto quiere decir que si la hipótesis fuera cierta

$$P(\chi^2 > \chi^2_\alpha = 3,84) = 0,05$$

y siendo el valor muestral $5,182 > 3,84$ *rechazamos la hipótesis* de que elegir ver la final sea independiente del sexo.

El *p-valor* para 5,182 es del 2 %, lo que indica que valores superiores a 5,182 solo se obtienen el 2 % de los casos, *si la hipótesis es cierta.*