Test de ajuste de chi-cuadrado

El test chi cuadrado, si bien se origina en aplicaciones a distribuciones discretas, se extiende a muestras continuas agrupadas en clases. El caso típico es, como siempre, el de una muestra $(X_1,...,X_n)$ de una ley desconocida. Las clases, denotadas $(c_1,...,c_r)$, son una partición del conjunto de los valores posibles. La hipótesis a comprobar tiene que ver con las probabilidades de las clases, para las cuales se toman valores teóricos $P_0(c_1),...,P_0(c_r)$.

$$\mathcal{H}_0 : \mathbb{P}\left[X_i \in c_k\right] = P_0)c_k, \quad \forall k = 1,...,r$$

Bajo la hipótesis $\mathcal{H}_0$, la distribución empírica de la muestra sobre las clases debe estar cerca de la distribución teórica. La distribución empírica es la de las frecuencias de la muestra en las clases:

$$\hat{P}(c_k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{ck}(X_i)$$

Se mide el ajuste de la distribución empírica a la distribución teórica por la *distancia de chi-cuadrado*. Se llama distancia de chi-cuadrado de $\hat{P}$ con respecto a P_0, y se denota por $D_{x^2}(P_0,\hat{P})$, al valor:

$$D_{x^2}(P_0,\hat{P}) = \sum_{j=1}^{r}\left[\frac{\left(P_0(c_j) - \hat{P}(c_j)\right)^2}{P_0(c_j)}\right]$$

La "distancia" de chi-cuadrado es por tanto una media ponderada de las diferencias cuadráticas entre los valores de P_0 y $\hat{P}$. No es una distancia en el sentido usual del término, pues ni siquiera es simétrica. La ley de probabilidad de $D_{x^2}(P_0,\hat{P})$ no tiene una expresión explícita en general. Se emplea entonces el siguiente resultado.

Teorema: Bajo la hipótesis $\mathcal{H}_0$ la ley de la variable aleatoria $nD_{x^2}(P_0,\hat{P})$ converge, cuando n tiende a infinito, a la ley de chi-cuadrado de parámetro r-1.
Si la hipótesis $\mathcal{H}_0$ es falsa, entonces la variable $nD_{x^2}(P_0,\hat{P})$ tiende a infinito. Por tanto aplicaremos un test unilateral a la derecha (rechazo de los valores muy grandes).

El ejemplo clásico del test es la experiencia de Mendel. En los guisantes, el carácter del color esta codificado por un gen que toma dos formas de alelo C y c, correspondientes a los colores amarillo y verde respectivamente. El amarillo es dominante, el verde recesivo. La forma lisa o arrugada es llevada por otro gen con dos alelos R (dominante) y r (recesivo). Si se cruzan dos individuos cuyo genotipo es CcRr, se pueden obtener 16 genotipos equiprobables. Los descendientes serán amarillos y lisos en 9 casos de los 16, amarillos y arrugados en 3 de los 16, verdes y lisos en 3 de los 16, verdes y arrugados en 1 caso de los 16. En sus experiencias Mendel obtuvo los siguientes resultados.

	Amarillo	Amarillo	Verde	Verde
	Liso	Arrugado	Liso	Arrugado
Casos	315	101	108	32
$\hat{P}_0(c_j)$	0.567	0.182	0.194	0.058
$P_0(c_j)$	9/16	3/16	3/16	1/16

El valor que toma el estadístico $nD_{x^2}(P_0, \hat{P})$ es de 0.47. Según el teorema, la región de rechazo debe ser calculada con respecto a la ley de chi-cuadrado $\chi^2(3)$. Por ejemplo, para un umbral 0.05, deberíamos rechazar los valores superiores a $Q_{x^2(3)}(0.95) = 7.81$. El p-valor de 0.47 es $1 - F_{x^2(3)}(0.47) = 0.925$. El resultado es por tanto completamente compatible con $\mathcal{H}_0$.

El ejemplo que se da a continuación tiene que ver con 1000 familias de 4 hijos para las cuales se conoce el número de varones, entre 0 y 4.

El modelo más simple que podemos proponer es que los nacimientos son independientes y los dos sexos son equiprobables. Por tanto la hipótesis nula es que la ley del número de varones para una familia de 4 hijos sigue la ley binomial $B(4,0.5)$. Las frecuencias, observadas y teóricas, son las siguientes.

Varones	0	1	2	3	4
$\hat{P}_0(c_j)$	0.0572	0.2329	0.3758	0.2632	0.0709
$P_0(c_j)$	1/16	4/16	6/16	4/16	1/16

El valor que toma el estadístico $nD_{x^2}(P_0, \hat{P})$ es de 34.47. Según el teorema, la región de rechazo debe ser calculada con respecto a la ley chi-cuadrado de parámetro 5-1=4. Por ejemplo para un umbral de 0.05, deberíamos rechazar los valores superiores a $Q_{x^2(4)}(0.95) = 9.49$. El p-valor de 34.47 es $1 - F_{x^2(4)}(34.47) = 5.97 * 10^{-7}$. Podemos, por tanto, rechazar la hipótesis $\mathcal{H}_0$.

El teorema es un resultado asintótico. Para poder usarlo, el orden del tamaño de las muestras debe ser al menos de las centenas. Además la aproximación que describe es menos buena cuando las probabilidades de las clases son débiles. Como regla empírica, se exige que los efectivos teóricos $nP(c_k)$ de cada clase sean al menos iguales a 5. Para alcanzar este objetivo a veces hay que recurrir al reagrupamiento de las clases; se forman nuevas clases uniendo varias de las iniciales, y se suman las frecuencias empíricas y las probabilidades teóricas de las clases agrupadas.

El test de chi-cuadrado se emplea con frecuencia para hacer un test de la bondad de ajuste a una familia particular de leyes que dependen de uno o más parámetros. En este caso, se debe estimar el parámetro a partir de los datos. El teorema no es ya del todo

válido. Si se han estimado h parámetros por el método de máximo de verosimilitud, a partir de las frecuencias de las diferentes clases, se debe remplazar la ley $\chi^2(r-1)$ por la ley $\chi^2(r-1-h)$.

Retomemos el ejemplo del número de varones en una familia de 4 hijos, pero esta vez para probar la hipótesis nula:

$\mathcal{H}_0$: El número de varones sigue una ley binomial $B(4,p)$

El parámetro "**p"** es desconocido y debe ser estimado. El estimador de máximo de verosimilitud (el cual maximiza la probabilidad de los datos observados) es en este caso la proporción total de varones entre los 40000 niños. Encontramos:

$$\hat{p} = 0.5144$$

Ahora aplicamos el test, pero con la distribución teórica calculada teniendo en cuenta el valor estimado del parámetro: la ley $B(4,p)$.

Varones	0	1	2	3	4
$\hat{P}_0(c_j)$	0.0572	0.2329	0.3758	0.2632	0.0709
$P_0(c_j)$	0.0556	0.2356	0.3744	0.2644	0.0700

El valor que toma el estadístico $nD_{x^2}(P_0,\hat{P})$ es ahora de 0.9883. Debe ser comparado con los valores de la ley chi-cuadrado de parámetro 5-1-1=3. El p-valor de 0.9883 es $1-F_{x^2(3)}(0.9883)=0.8041$, lo cual muestra que el resultado es perfectamente compatible con la hipótesis $\mathcal{H}_0$.

Al comparar los resultados de los dos tests precedentes, se puede aceptar la idea que los nacimientos son independientes, pero la proporción de los varones es significativamente superior a 0.5.

Con frecuencia hay que estimar parámetros a partir de datos no agrupados, o por otro método diferente al de máximo de verosimilitud. En ese caso no se dispone de un resultado teórico claro. El valor límite a partir del cual se debe rechazar la hipótesis $\mathcal{H}_0$ al umbral α está comprendido entre $Q_{X^2(r-1-h)}(1-\alpha)$ y $Q_{X^2(r-1)}(1-\alpha)$. En la práctica, después de haber calculado el valor "t" que toma $nD_{x^2}(P_0,\hat{P})$ teniendo en cuenta los h parámetros estimados, es prudente tomar la siguiente actitud:

- Rechazar $\mathcal{H}_0$ si $t > Q_{X^2(r-1)}(1-\alpha)$
- Aceptar $\mathcal{H}_0$ si $t < Q_{X^2(r-1-h)}(1-\alpha)$,
- Si $Q_{X^2(r-1-h)}(1-\alpha) < t < Q_{X^2(r-1)}(1-\alpha)$, no se llega a conclusión