EL TEST CHI-CUADRADO

El test Chi-cuadrado es un ejemplo de los denominados test de ajuste estadístico, cuyo objetivo es evaluar la bondad del ajuste de un conjunto de datos a una determinada distribución candidata. Su objetivo es aceptar o rechazar la siguiente hipótesis:

> H_0 = "Los datos de que se dispone son una muestra aleatoria de una distribución $F_X(x)$".

El procedimiento de realización del test Chi-cuadrado es el siguiente:

1) Se divide el rango de valores que puede tomar la variable aleatoria de la distribución en K intervalos adyacentes:

$$[a_0, a_1), [a_1, a_2), \ldots, [a_{K-1}, a_K)$$

Pueden ser $a_0 = -\infty$ y $a_K = \infty$.

2) Sea N_j el número de valores de los datos que tenemos que pertenecen al intervalo $[a_{j-1}, a_j)$.

3) Se calcula la probabilidad de que la variable aleatoria de la distribución candidata $F_X(x)$ esté en el intervalo $[a_{j-1}, a_j)$. Por ejemplo, si se trata de una distribución continua, esa probabilidad sería:

$$p_j = \int_{a_{j-1}}^{a_j} f_X(x)\,dx$$

siendo $f_X(x)$ la función densidad de probabilidad de la distribución candidata. También se puede hacer:

$$p_j = F_X(a_j) - F_X(a_{j-1})$$

Nótese que este es un valor teórico, que se calcula de acuerdo a la distribución candidata y a los intervalos fijados.

4) Se forma el siguiente estadístico de contraste:

$$\Delta = \sum_{j=1}^{K} \frac{(N_j - Np_j)^2}{Np_j}$$

Si el ajuste es bueno, Δ tenderá a tomar valores pequeños. Rechazaremos la hipótesis de la distribución candidata si Δ toma valores "demasiado grandes". Para ello se hace uso de la siguiente propiedad:

> "Si el número de muestras es suficientemente grande, y la distribución candidata es la adecuada Δ tiende a tener a una distribución Chi-cuadrado de $(K - 1)$ grados de libertad"

En realidad, la afirmación anterior sólo es estrictamente cierta si no hay que estimar ningún parámetro en la distribución candidata. Si para definir la distribución candidata hay que estimar algún parámetro (su media, su varianza,…) el número de grados de libertad de la distribución Chi-cuadrado es

(K – 1 – número de parámetros que hay que estimar a partir de los datos)

Tenemos por tanto, que si la distribución candidata es la adecuada, conocemos la distribución del parámetro. Además, si la distribución candidata es la adecuada, el valor del parámetro Δ tenderá a ser pequeño, y si no es adecuada, tenderá a ser grande.

Una forma razonable de fijar un umbral de decisión sería:

“Rechazar la distribución candidata si $\Delta > \chi^2_{gdl,\alpha}$

siendo $\chi^2_{gdl,\alpha}$ el valor que en la distribución Chi-cuadrado de *gdl* grados de libertad deja por encima una masa de probabilidad de α.

Nótese que α (nivel de significación) representa la probabilidad de equivocarse si la distribución candidata es la adecuada, y se fijará a un valor pequeño (típicamente, 0.1, 0.05 ó 0.01).

Es muy importante tener en cuenta que el test está sujeto a error. Acabamos de ver que es posible equivocarse aunque la hipótesis sobre la distribución candidata sea cierta, porque podemos tener la mala suerte de que los valores de Δ salgan grandes. Eso en todo caso sucederá con probabilidad baja (0.1, 0.05 ó 0.01, según acabamos de ver). Asimismo, podríamos equivocarnos también decidiendo que la distribución candidata es la adecuada aunque no sea cierto, debido a que los valores de Δ podrían salir pequeños. El test se basa en la suposición razonable de que si la distribución candidata no es la adecuada, los valores de Δ tenderán a salir por encima del umbral $\chi^2_{gdl,\alpha}$.