4 Contrastes del Chi 2 de bondad del ajuste

Un contraste de bondad del ajuste es de la forma

$$H_0 : P = P_0 \text{ frente a } H_1 : P \neq P_0$$

o

$$H_0 : P \in \{P_\theta\}_{\theta\in\Theta} \text{ frente a } H_1 : P \notin \{P_\theta\}_{\theta\in\Theta}$$

4.1 Contraste del χ^2 para modelos multinomiales

Consideramos k categoras $C_1, C_2, ..., C_k$ y denotamos p_j la probabilidad de la categora C_j donde

$$\sum_{j=1}^{k} p_j = 1, \text{ y } p_j > 0 \text{ para cada } j = 1, 2, ..., k.$$

- Suponemos que hacemos n experimentos aleatorios independientes cuyos resultados pertenecen a una de las k categoras anteriores.
- Consideramos los estadsticos $N_1, N_2, ..., N_k$, donde N_j es el número de resultados observados en la categora C_j. Tenemos que

$$\sum_{j=1}^{k} N_j = n.$$

  La distribución de $(N_1, N_2, ..., N_k)$ es multinomial $\mathcal{M}(n; p_1, p_2, ..., p_k)$:

$$P\left(N_1 = n_1, N_2 = n_2, ..., N_k = n_k\right) = n! \prod_{j=1}^{k} \frac{p_j^{n_j}}{n_j!}$$

Ejemplo 7 *Damos a n bebes una bola. Pedimos a cada uno que ponga su bola en una de las k cajas de colores que tiene por delante. Este experimento corresponde al modelo Multinomial si los bebes eligen de manera independiente.*

4.1.1 Contraste del χ^2 para una hipótesis simple

$H_0 : p_j = p_j^0$ para cada $j = 1, 2, ..., k$
$H_1 : (no\ H_0)$ existe j tal que $p_j \neq p_j^0$.

Ejemplo 8 *Si $p_j^0 = 1/k$ para cada $j = 1, 2, ..., k$, bajo H_0, estamos suponiendo que los bebes no tienen preferencias de colores (o no las distinguen).*

Para construir una regla de decisión que nos permita contrastar estas dos hipótesis, consideramos el estadstico siguiente

$$\begin{aligned} K_n &= n \sum_{j=1}^{k} \frac{\left(\frac{N_j}{n} - p_j^0\right)^2}{p_j^0} \\ &= \sum_{j=1}^{k} \frac{\left(N_j - np_j^0\right)^2}{np_j^0}. \end{aligned}$$

El estadstico K_n mide la discrepancia entre la frecuencias observadas $\left(\frac{N_j}{n}\right)$ y las probabilidades indicadas bajo H_0 $\left(p_j^0\right)$.

Teorema 4 *La distribución asintotca (cuando n tiende hacia el infinito) de K_n bajo H_0 es un chi 2 con $k-1$ grados de libertad :*

$$K_n \xrightarrow[n\to\infty]{d} \chi^2_{k-1}.$$

Fijaremos el riesgo I del contraste en función de la distribución asintotca de K_n. Regla de decisión del contraste :

$$\phi = \begin{cases} 1 \text{ si } K_n \geq \chi^2_{k-1,\alpha} \\ 0 \text{ sino} \end{cases}$$

Teorema 5 *El contraste precedente es convergente, o sea que su función de potencia (capacidad de rechazar H_0 cuando H_0 es falsa) tiende hacia 1 cuando $n \longrightarrow \infty$, y eso para cualesquiera riesgo I α.*

Prueba: Bajo H_1, existe un j tal que $p_j \neq p_j^0$, entonces tendremos que

$$\frac{\left(\frac{N_j}{n} - p_j^0\right)^2}{p_j^0} \xrightarrow[n\to\infty]{c.s.} \frac{\left(p_j - p_j^0\right)^2}{p_j^0} > 0,$$

puesto que por la ley de los grandes números $\frac{N_j}{n} \xrightarrow[n\to\infty]{c.s.} p_j$. Entonces para n bastante grande

$$n\frac{\left(\frac{N_j}{n}-p_j^0\right)^2}{p_j^0} \xrightarrow[n\to\infty]{c.s.} \infty,$$

Por tanto $K_n \xrightarrow[n\to\infty]{c.s.} \infty$, y

$$P_{H_1}\left(K_n \geq \chi^2_{k-1,\alpha}\right) \xrightarrow[n\to\infty]{c.s.} 1.$$

Ejemplo 9 *Se ha estimado que el número de accidentes diarios en cada regimiento del ejército sigue una distribución de Poisson de parámetro* 2. *Un determinado regimiento ha recogido, durante* 200 *das, los siguientes datos:*

nº de accidentes	0	1	2	3	4	5	6	7
nº de das	22	53	58	39	20	5	2	1

con los cuales se quiere contrastar si se ajusta a la distribución indicada.Bajo H_0 *la probabilidad de que haya* j *accidentes en un da es* $p_j^0 = \exp(-2)2^j/j!$, *y queremos comparar esas probabilidades con las frecuencias observadas : aqu* N_j *es el número de das con* j *accidentes. Para hallar el valor del estadstico* K_{200} *calculamos*

nº de accidentes	0	1	2	3	4	≥ 5
frec. observada N_j/n	*0.11*	*0.265*	*0.29*	*0.195*	*0.10*	*0.04*
frec. esperada p_j^0	*0.13*	*0.27*	*0.27*	*0.19*	*0.09*	*0.05*

Por tanto

$$\begin{aligned} K_{200} &= n\sum_{j=0}^{5}\frac{\left(\frac{N_j}{n}-p_j^0\right)^2}{p_j^0} \\ &= 200\left(\frac{0.02^2}{0.13}+...+\frac{0.01^2}{0.05}\right) \simeq 1.57 \end{aligned}$$

Tenemos que K_{200} *sigue aproximadamente un* χ^2 *con* 5 *grados de libertad (puesto que* $k = 6$*). Para* $\alpha = 5\%$, *obtenemos que* $\chi^2_{5,5\%} = 11.07$. *Entonces el valor hallado de* K_{200} *no permite rechazar* H_0 *con un nivel* α *igual a* 5%.

4.1.2 Contraste del χ^2 para una hipótesis compuesta

- $H_0 : p_j = p_j(\theta)$ para cada $j = 1, 2, ..., k$ y donde $\theta \in \Theta \subset R^m$, $(m < k-1)$.
- $H_1 : (\text{no } H_0)$

Las funciones $p_j(.)$ son tal que para cada $\theta \in \Theta$, $p_j(\theta) > 0$ y $\sum_{j=1}^{n} p_j(\theta) = 1$.

Ejemplo 10 *$k = 4$, bajo H_0 suponemos que para $j \in \{0, 1, 2, 3\}$,*

$$p_j = p_j(\theta) = \binom{j}{3} \theta^j (1-\theta)^{3-j}$$

donde $\theta \in \Theta =]0, 1[$.

Introducimos el estadstico K_n^* que mide la discrepancia entre H_0 y la realidad observada:

$$\begin{aligned} K_n^* &= \sum_{j=1}^{k} \frac{\left(N_j - np_j(\widehat{\theta})\right)^2}{np_j(\widehat{\theta})} \\ &= n \sum_{j=1}^{k} \frac{\left(\frac{N_j}{n} - p_j(\widehat{\theta})\right)^2}{p_j(\widehat{\theta})}, \end{aligned}$$

donde $\widehat{\theta}$ es el estimador de máxima verosimilitud de θ en el modelo multinomial bajo H_0.

Teorema 6 *Si las funciones $p_j(.)$ definida sobre Θ se pueden derivar dos veces, entonces la distribución asintotca de K_n^* bajo H_0 es un χ^2 con $k-m-1$ grados de libertad :*

$$K_n^* \xrightarrow[n\to\infty]{d} \chi^2_{k-m-1}.$$

La regla de decisión para contrastar H_0 frente H_1 será:

$$\phi = \begin{cases} 1 \text{ si } K_n^* \geq \chi^2_{k-m-1,\alpha} \\ 0 \text{ sino} \end{cases}$$

Ejemplo 11 *(continuación) Suponemos ahora que sólo sabemos que el número diario de accidentes sigue una distribución de Poisson de parámetro θ (desconocido). Aqu tenemos que $m = 1$, y el estimador de máxima verosimilitud de la media θ es el número medio observado de accidentes diarios $\widehat{\theta} = \sum_{j=0}^{7} j\frac{N_j}{n} = 2.05$.*

n^o de accidentes	0	1	2	3	4	≥ 5
frec. observada: N_j/n	*0.11*	*0.265*	*0.29*	*0.195*	*0.10*	*0.04*
frec. esperada: $p_j(\widehat{\theta})$	*0.129*	*0.264*	*0.271*	*0.185*	*0.095*	*0.057*

Obtenemos $K^*_{200} \simeq 2.04$. Puesto que $k-m-1 = 6-1-1 = 4$, para contrastar H_0 con un nivel $\alpha = 5\%$ utilizaremos el cuantil 95% de un χ^2 con 4 grados de libertad : $\chi^2_{4,5\%} = 9,48$. Por tanto aceptamos H_0.

4.2 Contraste de independencia y simetra

La independencia y la simetra de una tabla de contingencia son hipótesis compuestas.

Ejemplo 12 *Consideramos la tabla siguiente sobre el grado de visión del ojo derecho y izquierdo (clasificado en cuatro grupo $1, 2, 3, 4$ del mejor al peor) de una muestra de 7477 mujeres mayores*

$O_D \backslash\ O_I$	*1*	*2*	*3*	*4*	*$N_{\bullet j}$*
1	*1520*	*266*	*124*	*66*	*1976*
2	*234*	*1512*	*432*	*78*	*2256*
3	*117*	*362*	*1772*	*205*	*2456*
4	*36*	*82*	*179*	*492*	*789*
$N_{i\bullet}$	*1907*	*2222*	*2507*	*841*	*7477*

Queremos a partir de estos datos estudiar:
(i) La independencia de los ojos (¡parece mala!)
(ii) Simetra global de los ojos (simetra de la tabla).

4.2.1 Contraste de independencia:

Sea la tabla de contingencia:

	C_1	C_2	...	C_c
L_1				
L_2				
$\vdots$				
$\vdots$				
L_l				

Si queremos contrastar la independencia de las columnas y las filas, la hipótesis de independencia H_0 se define por

$$H_0 : P(L_i \cap C_j) = P(L_i)P(C_j)$$

para cada $i = 1, 2, ..., l$ y $j = 1, 2, ..., c$.
Aqu $k = l \times c$ y H_0 es compuesta :

$$P(L_i \cap C_j) = p_{ij} = p_{ij}(\theta) = p_{i\bullet}p_{\bullet j}$$

donde $p_{i\bullet} = P(L_i)$, $p_{\bullet j} = P(C_j)$ y

$$\theta = \left(p_{1\bullet}, p_{2\bullet}, ..., p_{(l-1)\bullet}, p_{\bullet 1}, p_{\bullet 2}, ...p_{\bullet(c-1)}\right)$$

Por tanto, $m = \dim \theta = l - 1 + c - 1 = l + c - 2$.

El estadstico K_n^* se escribe aqu:

$$K_n^* = \sum_{i,j=1}^{l,c} \frac{\left(N_{ij} - np_{ij}(\widehat{\theta})\right)^2}{np_{ij}(\widehat{\theta})}$$

- Calculo de $p_{ij}(\widehat{\theta})$:
  En cada casilla de la tabla observamos N_{ij}, número observado de mujeres que pertenecen a la categora L_i y C_j. Bajo H_0, la probabilidad que una realización pertenezca a L_i y C_j es $p_{ij}(\theta) = p_{i\bullet}p_{\bullet j}$, por tanto estimar θ es estimar $p_{i\bullet} = P(L_i)$ y $p_{\bullet j} = P(C_j)$. Los estimadores máximó verosmil de estas dos probabilidades son $\widehat{p}_{i\bullet} = \frac{N_{i\bullet}}{n} = \frac{N_{i1}+N_{i2}+...N_{ic}}{n}$ y $\widehat{p}_{\bullet j} = \frac{N_{\bullet j}}{n} = \frac{N_{1j}+N_{2j}+...N_{fj}}{n}$ y por tanto $p_{ij}(\widehat{\theta}) = \frac{N_{i\bullet}}{n}\frac{N_{\bullet j}}{n}$. As que

$$K_n^* = \sum_{i,j=1}^{l,c} \frac{\left(N_{ij} - \frac{N_{i\bullet}N_{\bullet j}}{n}\right)^2}{\frac{N_{i\bullet}N_{\bullet j}}{n}}$$

- K_n^* tiene como distribución asintotca un χ^2_{k-m-1} donde

$$\begin{aligned} k-m-1 &= lc-(l+c-2)-1 \\ &= (l-1)(c-1) \end{aligned}$$

Entonces la regla de decisión del contraste de independencia (con nivel asintotco α) será :

Rechazar H_0 si $K_n^* > \chi^2_{(l-1)(c-1),\alpha}$

Ejemplo 13 *(Continuación) Aqu $l = c = 4$, por tanto el estadstico K^*_{7477} (para la independencia) sigue aproximadamente una distribución del χ^2_9. Obtenemos $K^*_{7477} \simeq 3500$ y consultando la tabla del χ^2_9 hallamos $\chi^2_{9,0.05} = 16.92$. Por tanto rechazamos la independencia con el test del χ^2_9 para un nivel de significación (asintótico) del 5%.*

4.2.2 Contraste de Simetra de una tabla

	C_1	C_2	...	C_r
L_1				
L_2				
$\vdots$				
$\vdots$				
L_r				

Si queremos contrastar la simetra de la tabla, H_0 se define por

$$H_0 : P(L_i \cap C_j) = P(L_j \cap C_i)$$

para cada $i = 1, 2, ..., r$ y $j = 1, 2, ..., r$.
Aqu $k = r \times r$ y H_0 es compuesta :

$$P(L_i \cap C_j) = P(L_j \cap C_i) = p_{ij}(\theta) = p_{ij}$$

donde

$$\theta = (p_{ij})_{i \leq j} \setminus \{p_{rr}\}$$

y $m = \dim \theta = r(r+1)/2 - 1$

- Calculo de $p_{ij}(\widehat{\theta})$:
  Estimamos $p_{ij}(\theta)$ bajo H_0 por $p_{ij}(\widehat{\theta}) = \frac{N_{ij}+N_{ji}}{2n}$

$$\begin{aligned} K_n^* &= \sum_{i,j=1}^{r} \frac{\left(N_{ij} - \frac{N_{ij}+N_{ji}}{2}\right)^2}{\frac{N_{ij}+N_{ji}}{2}} \\ &= \sum_{i,j=1}^{r} \frac{\left(\frac{N_{ij}-N_{ji}}{2}\right)^2}{\frac{N_{ij}+N_{ji}}{2}} \end{aligned}$$

- K_n^* tiene como distribución asintotca un χ^2_{k-m-1} donde

$$\begin{aligned} k - m - 1 &= r^2 - r(r+1)/2 \\ &= r(r-1)/2 \end{aligned}$$

  Entonces la regla de decisión del contraste de simetra (con nivel asintotco α) será : Rechazar H_0 si $K_n^* > \chi^2_{r(r-1)/2,\alpha}$

Ejemplo 14 *(continuación) Aqu $r = 4$ por tanto el estadstico K^*_{7477} (para la simetra) sigue aproximadamente una distribución del χ^2_6. Obtenemos $K^*_{7477} \simeq 11.25$ y consultando la tabla del χ^2_6 hallamos $\chi^2_{6,0.05} = 12.6$. Por tanto aceptamos la simetra de la tabla con el test del χ^2_9 de nivel* 5%.

5 Contraste de Kolmogorov-Smirnov de bondad del ajuste

El método de los test χ^2 consiste en comparar un histograma de la distribución de los datos con la distribución teórica bajo H_0 (frecuencia observada versus frecuencia de acuerdo con H_0).
Problema: El histograma supone una discretización de los datos (partitión en categoras). Por tanto, si la distribución de los datos es continua perdemos información.
Alternativa: Comparar las funciones de distribuciones muestrales y teóricas en lugar de los histogramas.

Sea $X_1, X_2, \ldots, X_n$ una muestra de datos de una función de distribución continua F desconocida. La función de distribución muestral F_n se define por

$$\begin{aligned} F_n(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{X_i \leq x\}} \qquad \text{(proporción de datos} \leq x) \\ &= \begin{cases} 0 & \text{si } x \leq X_{(1)} \\ i/n & \text{si } X_{(i)} \leq x \leq X_{(i+1)} \\ 1 & \text{si } x \geq X_{(n)} \end{cases}, \end{aligned}$$

donde $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ son los elementos de la muestra ordenada.

Teorema 7 (Glivenko-Cantelli) *Cuando n tiende hacia el infinito*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \stackrel{c.s}{\to} 0$$

Ahora si consideramos el contraste de bondad del ajuste $H_0 : F = F_0$, siendo F_0 una distribución continua conocida, podemos utilizar el estadstico

$$\Delta_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

para medir la distancia entre la realidad (observada) y la hypótesis H_0. De hecho, por el teorema anterior, cuando n tiende hacia el infinito,

$$\Delta_n \stackrel{c.s}{\to} \sup_{x \in \mathbb{R}} |F(x) - F_0(x)| \begin{cases} = 0 & \text{, si } H_0 \text{ es cierto} \\ > 0 & \text{, si } H_0 \text{ es falso} \end{cases}$$

El test de Kolmogorov-Smirnov (KS) se basa en el estadstico Δ_n y rechaza H_0 cuando $\Delta_n > u$ ("Δ_n grande"). Para controlar el nivel del test necesitamos conocer la distribución de Δ_n.

Lema 1 *Bajo H_0, el estadstico Δ_n tiene la misma distribución que*

$$\max\left\{ \max_{1 \leq i \leq n} \left(\frac{i}{n} - U_{(i)} \right) ; \max_{1 \leq i \leq n} \left(U_{(i)} - \frac{i-1}{n} \right) \right\}$$

donde $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$ es una muestra ordenada de una uniforme en $(0, 1)$.

5.1 Cálculo de Δ_n

Para contrastar H_0, necesitamos calcular

$$\Delta_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|$$

- Si denotamos $X_{(0)} = -\infty$ y $X_{(n+1)} = +\infty$, tenemos que

$$\begin{aligned}\sup_{x\in\mathbb{R}} [F_n(x) - F_0(x)] &= \max_{1\le i\le n} \sup_{X_{(i)}\le x\le X_{(i+1)}} \left[\frac{i}{n} - F_0(x)\right] \\ &= \max_{1\le i\le n} \left[\frac{i}{n} - F_0(X_{(i)})\right]\end{aligned}$$

- De manera similar,

$$\begin{aligned}\sup_{x\in\mathbb{R}} [F_0(x) - F_n(x)] &= \max_{0\le i\le n} \left[F_0(X_{(i+1)}) - \frac{i}{n}\right] \\ &= \max_{1\le i\le n} \left[F_0(X_{(i)}) - \frac{i-1}{n}\right]\end{aligned}$$

Por tanto

$$\Delta_n = \max\left(\max_{1\le i\le n}\left[\frac{i}{n} - F_0(X_{(i)})\right], \max_{1\le i\le n}\left[F_0(X_{(i)}) - \frac{i-1}{n}\right]\right)$$

Teorema 8 *Si $X \sim F_0$ y F_0 es continua en $\mathbb{R}$, entonces la variable $U = F_0(X)$ sigue una uniforme en $(0,1)$.*

5.2 Contraste de normalidad de Lilliefors

Consideramos la hipótesis simple

$$H_0 : F = \Phi_{\mu_0,\sigma_0^2}$$

donde Φ_{μ,σ^2} es la función de distribución de la normal $N(\mu,\sigma^2)$.
El test de KS es entonces: "Rechazar H_0" si

$$\Delta_n = \sup_{x\in\mathbb{R}} \left|F_n(x) - \Phi_{\mu_0,\sigma_0^2}(x)\right| > u_\alpha.$$

Pero, en general, no se conoce la media y la varianza de F. La hipótesis de normalidad de F es entonces compuesta:

$$H_0 : \left\{ \begin{array}{l} F \text{ es una distribución normal:} \\ F \in \{\Phi_{\mu,\sigma^2}, (\mu, \sigma^2)\} \end{array} \right\}$$

El test de Lilliefors para este contraste se basa en el estadstico de KS, substituyendo μ y σ^2 por sus estimadores:

$$\Delta_n^* = \sup_{x \in \mathbb{R}} \left| F_n(x) - \Phi_{\overline{X}, S^2}(x) \right|$$

donde $\overline{X}$ y S^2 son respectivamente la media muestral y la varianza muestral. Para un nivel de significación dado α, el umbral critico u_α del test se obtiene mediante la tabla de Lilliefors.

6 Contraste de la mediana

- Disponemos de datos apareados $(X_1, Y_1), \ldots, (X_n, Y_n)$ que provienen de una distribución $P_{X,Y}$.
- Queremos contrastar la hipótesis de simetra:
  $H_0 : P_{X,Y} = P_{Y,X}$ (implica $P_X = P_Y$).

Teorema 9 *Bajo, la hipótesis $H_0 : P_{X,Y} = P_{Y,X}$, la variable $Z = X - Y$ tiene une distribución simétrica.*

- Sea $Z_i = X_i - Y_i$, para $i = 1, 2, \ldots, n$, consideramos el estadstico

  $$S_n = \sum_{i=1}^{n} \mathbf{I}_{\{Z_i \leq 0\}}$$

  que bajo H_0 sigue una distribución binomial $B(n, p_0)$, donde $p_0 = P\{Z_i \leq 0\} = 1/2$.
- Por tanto, el test (con nivel α) de este contraste será

  $$\phi = \left\{ \begin{array}{ll} 1 \text{ si} & |S_n - \frac{n}{2}| > u_\alpha \\ \gamma_\alpha \text{ si} & |S_n - \frac{n}{2}| = u_\alpha \\ 0 \text{ si} & |S_n - \frac{n}{2}| < u_\alpha \end{array} \right. ,$$

  donde u_α y γ_α verifican

  $$\alpha = P_{H_0}\left(\left| S_n - \frac{n}{2} \right| > u_\alpha \right) + \gamma_\alpha P_{H_0}\left(\left| S_n - \frac{n}{2} \right| = u_\alpha \right)$$

7 Contraste de homogeneidad

- Disponemos de 2 muestras de datos, independientes entre s: $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$.
- Queremos contrastar si las dos muestras provienen de la misma distribución (homogeneidad): $H_0 : P_X = P_Y$, donde P_X y P_Y son distribuciones desconocidas.

7.1 Contraste de homogeneidad del χ^2

Este contraste está basado en la comparación de los histogramas de las dos muestras.

- Definimos k categoras: $C_1, C_2, \ldots, C_k$; clasificando en ellas los datos de cada muestra.
- Denotamos N_{ij} ($i = 1, 2$ y $j = 1, \ldots, k$) el número observado de datos de la i^{esima} muestra que pertenecen a C_j.
- Denotamos $p_{1j} = P(X \in C_j)$ y $p_{2j} = P(Y \in C_j)$.

La hipótesis H_0 de homogeneidad se traduce en que cada categoras C_j debe tener una probabilidad p_{ij} que no depende de i :

$$H_0 : \begin{cases} p_{1j} = p_{2j} \ (p_{ij}(\theta) = p_j) \\ \text{para cada } j = 1, \ldots, k \end{cases}$$

donde $\theta = (p_1, p_2, \ldots, p_{k-1})$ y $m = \dim(\theta) = k - 1$.
Si H_0 es correcta y las probabilidades p_j fuesen conocidas, el estadstico

$$\sum_{j=1}^{k} \frac{(N_{ij} - n_i p_j)^2}{n_i p_j} \xrightarrow{d} \chi^2(k-1)$$

Por tanto,

$$K_n = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(N_{ij} - n_i p_j)^2}{n_i p_j} \xrightarrow{d} \chi^2(2(k-1)),$$

donde $n = n_1 + n_2$.
Sin embargo, las probabilidades p_j no son conocidas y han de ser sustituidas

por su estimación de máxima verosimilitud $\widehat{p}_j = N_{\bullet j}/n$, dando lugar al estadstico

$$K_n^* = \sum_{i=1}^{2}\sum_{j=1}^{k} \frac{(N_{ij} - n_i N_{\bullet j}/n)^2}{n_i N_{\bullet j}/n}$$

que sigue asintóticamente, según el Teorema 6, una distribución del χ^2 con un número de grados de libertad que se reduce a $2(k-1) - m = (k-1)$. El test (de nivel asintótico) para este contraste será entonces:

$$\text{Rechazar } H_0 \text{ si } K_n^* > \chi^2_{(k-1;\alpha)}$$

7.2 Contraste de homogeneidad de KS

El test de KS para contrastar la hipótesis de homogeneidad $H_0 : F_X = F_Y$ está basado en el estadstico

$$\Delta_{n_1,n_2} = \sup_{x\in\mathbb{R}} |F_{X,n_1}(x) - F_{Y,n_2}(x)|$$

donde

$$F_{X,n_1}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{I}_{\{X_i \le x\}} \text{ y } F_{Y,n_2}(x) = \frac{1}{n_2}\sum_{i=1}^{n_2} \mathbf{I}_{\{Y_i \le x\}}$$

son las distribuciones muestrales de las dos muestras.

- Si H_0 es cierta es probable que $F_{X,n_1}(x)$ y $F_{Y,n_2}(x)$ sean próximas, y que, por tanto, Δ_{n_1,n_2} tenga un valor relativamente pequeño.
- Si en cambio, $F_X \neq F_Y$, puesto que $|F_{X,n_1}(x) - F_{Y,n_2}(x)|$ tenderá a aproximarse a $|F_X(x) - F_Y(x)|$, el valor de Δ_{n_1,n_2} será más elevado.

Esto conduce a rechazar H_0 cuando $\Delta_{n_1,n_2} > u_\alpha$, donde el umbral u_α verifica $P_{H_0}(\Delta_{n_1,n_2} > u_a) = \alpha$, α siendo el nivel (prefijado) del test.