Estadística No Paramétrica

Clase 8: Test de Kolgomorov - Smirnov para testear dos poblaciones

Joaquin Cavieres G.

Ingeniería en Estadística

Facultad de Ciencias, Universidad de Valparaíso

Test de Kolmogorov - Smirnov (K-S)

Recordatorio: El contenido de la segunda parte del curso se centrará sobre dos muestras aleatorias independientes.

Test de Kolmogorov - Smirnov (K-S)

Supuestos

- Las observaciones $X_1, ..., X_m$ son variables aleatorias desde una población 1, esto es, que todas las $X's$ son *i.i.d*. Las observaciones $Y_1, ..., Y_n$ son variables aleatorias desde una población 2, esto es, que todas las $Y's$ son *i.i.d*.
- Las $X's$ e $Y's$ son mutuamente independiente, esto significa que ademas de que las muestras sean independientes, existe independencia entre las dos muestras.

Test de Kolmogorov - Smirnov (K-S)

Hipótesis

Si $X_1, ..., X_m$ y $Y_1, ..., Y_n$ son muestras aleatorias independientes que satisfacen los supuestos de los puntos expuestos previamente y obtenidas desde poblaciones continuas con función de distribución F correspondiente a la población 1 y una función de distribución G correspondiente a la población 2. Por tanto, se plantea:

Bajo estos supuestos, estamos interesados en evaluar si hay alguna diferencia entre las distribuciones de probabilidad para X e Y, esto es en forma general:

$$H_0 : F(t) = G(t) \text{ para todo } t$$

Test de Kolmogorov - Smirnov (K-S)

Forma de calculo

Para calcular un test K-S de ambas colas para dos muestras independientes, que llamaremos J, primero debemos:

- Obtener la función de distribución empírica para las muestras X e Y.

$$\begin{aligned} F_m(t) &= \frac{\text{N de la muestra} X \leq t}{m} \\ G_n(t) &= \frac{\text{N de la muestra} Y \leq t}{n} \end{aligned}$$

- Definir como d el máximo común divisor entre m y n.
- Calcular J como:

$$J = \frac{mn}{d} \max_{(-\infty < t < \infty)} \{|F_m(t) - G_n(t)|\} \tag{1}$$

J es el estadísico K-S para comparar dos muestras independientes.

Forma de calculo

Como calculamos a J desde las muestras X e Y, usando a $F_m(t)$ y $G_n(t)$ funciones escalonadas con valores funcionales sólo para X e Y respectivamente. Por tanto, denotamos a $Z_1 \leq ...., \leq Z_N$ con $N = m + n$ valores ordenados para las muestras combinadas $X_1, ..., X_m$ y $Y_1, ..., Y_n$, entonces podemos re-escribir a J como:

$$J = \frac{mn}{d} \max_{(i=1,...,N)} \{|F_m(Z_i) - G_n(Z_i)|\} \tag{2}$$

Forma de calculo

Para contrastar H_0 (funciones de distribución de X e Y son iguales) con H_1 (las funciones de distribución no son iguales), con un nivel de significancia α,

$$\text{Rechazamos } H_0 \text{ si } J \geq j_\alpha; \quad \text{, de otra manera no rechazamos } H_0 \tag{3}$$

donde la constante j_α es elegida para hacer que la probabilidad del error tipo I sea igual a α.

Aproximación mediante n grande

La aproximación mediante un n grande esta basada en la normalidad asintótica de J correctamente normalizada, como el mínimo de (m, n) tiende al infinito, entonces:

$$J^* = \left(\frac{mn}{N}\right)^{1/2} \max_{(i=1,\ldots,N)} \{|F_m(Z_i) - G_n(Z_i)|\} = \frac{d}{(mnN)^{1/2}} J \qquad (4)$$

$$P_0(J^* < s) \to \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}, 0 \quad \text{para} \quad s > 0, \qquad (5)$$

Definiendo una función $Q(s)$ por:

$$Q(s) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}, 0 \quad \text{para} \quad s > 0, \qquad (6)$$

Aproximación mediante n grande

Ya teniendo la aproximación de J entonces podemos realizar el test de hipótesis como:

$$\text{Rechazamos } H_0 \text{ si } J^* \geq q^*_\alpha; \qquad \text{, de otra manera no rechazamos } H_0 \tag{7}$$

donde q^*_α esta definido como $Q(q^*_\alpha) = \alpha$

Nota: Para encontrar q^*_α podemos utilizar el siguiente comando `qKolSmirnLSA(`α`)`. Por ejemplo, si queremos determinar $q^*_{0.05}$ hacemos: `qKolSmirnLSA(0.05)` y obtenemos $q^*_\alpha = 1.358$.

Test de Kolmogorov - Smirnov (K-S)

Si X e Y son dos variables aleatorias continuas independientes con función de distribución empírica F_X y F_Y respectivamente. El test de hipótesis para contrastar H_0 debería ser:

$$
\begin{aligned}
H_0 &: F_X(t) = F_Y(t), \ \ \forall t \in R \\
H_0 &: F_X(t) \neq F_Y(t), \ \ \text{para al menos un} \ \ t \in R
\end{aligned}
$$

Ya que $X_1, ..., X_m$ y $Y_1, ..., Y_n$ son dos variables aleatorias independientes con función de distribución empírica $\hat{F}_X$ y $\hat{F}_Y$, el K-S test es:

$$D_{n_1,n_2} = \left(\frac{mn}{N}\right)^{1/2} \max |\hat{F}_X - \hat{F}_Y|$$

Test de Kolmogorov - Smirnov (K-S)

Ejemplos

Ver ejemplos en `R`