Estadística No Paramétrica

Clase 4: Test de Kolgomorov-Smirnov

Joaquin Cavieres G.

Ingeniería en Estadística

Facultad de Ciencias, Universidad de Valparaíso

Test de Kolgomorov-Smirnov

Dentro de la estadística no paramétrica también podemos encontrar el test de Kolgomorov-Smirnov. Este test tiene la finalidad comparar entre dos distribuciones de probabilidad continuas y unidimensionales en función de una muestra aleatoria con una función de de distribución probabilidad como referencia.

Test de Kolgomorov-Smirnov

El test de Kolgomorov-Smirnov es uno de los test más populares en la estadística no paramétrica. El test cuantifica la distancia entre la función de distribución empírica de la muestra F_n y la función de distribución acumulada de una función de distribución de probabilidad de referencia $F(x)$, esto es:

Definición

La función de distribución empírica F_n para n observaciones $i.i.d$ de una muestra aleatoria X_i esta definida como:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{I}_{(0,\infty)}(X_i) \tag{1}$$

donde $\boldsymbol{I}_{(0,\infty)}(X_i)$ es la función indicadora igual a 1 si $X_i \leq x$ e igual a 0 en otro caso, por lo tanto:

$$D_n = sup_x|F_x(n) - F_X(x)| \tag{2}$$

Test de Kolgomorov-Smirnov

Supongamos que tenemos $X_1, ..., X_n$ una v.a con una función de distribución la cual no conocemos y que denotamos como F_X. A nosotros nos interesaría probar que F_X^0 es la verdadera distribución. Dado lo anterior planteamos lo siguiente hipótesis:

$$H_0 : F_X(x) = F_X^0(x)$$
$$H_1 : F_X(x) \neq F_X^0(x)$$

Test de Kolgomorov-Smirnov

Supongamos que tenemos $X_1, ..., X_n$ una v.a con una función de distribución la cual no conocemos y que denotamos como F_X. A nosotros nos interesaría probar que F_X^0 es la verdadera distribución. Dado lo anterior planteamos lo siguiente hipótesis:

$$H_0 : F_X(x) = F_X^0(x)$$
$$H_1 : F_X(x) \neq F_X^0(x)$$

¿Le parece conocida esta misma hipótesis para funciones de distribución?

Test de Kolgomorov-Smirnov

Como en la prueba X^2 se contrastan hipótesis frente a clases "arbitrarias", el test de Kolgomorov-Smirnov evita esta categorización. Considerando la definición de la función de distribución empírica entonces:

Si asumimos que X_i es una v.a, entonces $F_n(x)$ también lo es, así:

$$\mathbb{E}(F_n(x)) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} \boldsymbol{I}_{(0,\infty)}(X_i)}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\boldsymbol{I}_{(0,\infty)}(X_i)) \tag{3}$$

es un estimador insesgado y consistente para $F_n(x)$

Ya que $\boldsymbol{I}_{(0,\infty)}(X_i) \sim Ber(\mathbb{P}(X_i \leq x)) = Ber(F_X(x))$, entonces $\mathbb{E}(\boldsymbol{I}_{(0,\infty)}(X_i)) = F_X(x)$ para todo i, así.

Test de Kolgomorov-Smirnov

Estimador insesgado

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(I_{(0,\infty)}(X_i)) = \frac{1}{n}n\mathbb{E}(I_{(0,\infty)}(X_i)) = F_X(x) \tag{4}$$

Consistencia del estimador

$$\begin{aligned}\mathbb{VAR}(F_n(x)) =& \mathbb{VAR}\left(\frac{\sum_{i=1}^{n} I_{(0,\infty)}(X_i)}{n}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{VAR}(I_{(0,\infty)}(X_i)) = \frac{F_X(x)(1-F_X(x))}{n}\end{aligned} \tag{5}$$

Test de Kolgomorov-Smirnov

Cuando suponemos observada la muestra aleatoria, F_n puede ser visualizada y se espera F_n se aproxime a la verdadera función de distribución de donde proviene la muestra.

Ejemplo 1

Ver código en `R` simulación de una v.a Normal, construcción de una función de distribución empírica y ajuste a los datos.

Test de Kolgomorov-Smirnov

Parece ser que la información que nos entrega la función de distribución empírica es razonáble para la construcción de un estadístico de prueba para así verificar la bondad de ajuste. Mediante la prueba anterior de insesgamiento y consistencia entonces podemos probar que $F_n(x)$ converge casi seguramente a $F_X(x)$ para cada x.

Test de Kolgomorov-Smirnov

Convergencia en forma uniforme

Teorema de Glivenko-Cantelli

Sea $X_1, ..., X_n$ una muestra aleatoria de $F_X(x)$ y sea $F_n(x)$ su función de distribución empírica, entonces:

$$D_n = sup_x |F_n(x) - F_X(x)| \to 0 \tag{6}$$

Test de Kolgomorov-Smirnov

Convergencia en forma uniforme

Teorema de Glivenko-Cantelli

Sea $X_1, ..., X_n$ una muestra aleatoria de $F_X(x)$ y sea $F_n(x)$ su función de distribución empírica, entonces:

$$D_n = sup_x |F_n(x) - F_X(x)| \to 0 \tag{6}$$

Esto quiere decir que cuando existe un mayor número de muestra entonces F_n reproduce casi totalmente la verdadera función de distribución.

Test de Kolgomorov-Smirnov

Convergencia en forma uniforme

Teorema de Glivenko-Cantelli

Sea $X_1, ..., X_n$ una muestra aleatoria de $F_X(x)$ y sea $F_n(x)$ su función de distribución empírica, entonces:

$$D_n = sup_x |F_n(x) - F_X(x)| \to 0 \tag{6}$$

Esto quiere decir que cuando existe un mayor número de muestra entonces F_n reproduce casi totalmente la verdadera función de distribución.

La clave en el test de Kolgomorov-Smirnov es que la distribución del supremo no depende de la distribución desconocida de F_X^0 de la muestra, siempre y cuando F_X^0 sea una distribución continua.

Test de Kolgomorov-Smirnov

¿Como evaluar el test de Kolgomorov-Smirnov?

Test de Kolgomorov-Smirnov

¿Como evaluar el test de Kolgomorov-Smirnov?

Si $X_1, ..., X_n \sim F$, entonces $F_n(x)$ es un estimador no paramétrico de F (recordar que F es una función de distribución).

Entonces:

- Una medida de proximidad entre F_n (dado por la muestra) y F_0 (especificado por H_0) debería indicar la "distancia" entre F_n y F_0 para determinar si H_0 es falsa.

Test de Kolgomorov-Smirnov

Si consideramos que:

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

Test de Kolgomorov-Smirnov

Si consideramos que:

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

entonces

Estadístico D_n

Si $H_0 : F = F_0$ se cumple, entonces D_n tiende a ser pequeño. De lo contrario, si D_n tiende a ser grande, entonces $H_1 : F \neq F_0$ y rechazamos H_0.

Test de Kolgomorov-Smirnov

Calculo "a mano" de D_n

$$D_n = \max(D_n^+, D_n^-), \tag{6.2}$$

$$D_n^+ := \sqrt{n} \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - U_{(i)} \right\},$$

$$D_n^- := \sqrt{n} \max_{1 \leq i \leq n} \left\{ U_{(i)} - \frac{i-1}{n} \right\},$$

donde U_j representa para el j-ésimo sorteado $U_i := F_0(X_i), i = 1, ..., n$

Test de Kolgomorov-Smirnov

Distribución bajo H_0

Si H_0 se cumple y F_0 es continua, entonces D_n tiene una función cumulativa asintótica dada por la función K:

$$\lim_{x\to\infty} \mathbb{P}[D_n \leq x] = K(x) := 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2}$$

Test de Kolgomorov-Smirnov

Consideraciones importantes:

- Bajo H_0 el test Kolgomorov-Smirnov no depende de F_0 pero solamente si F_0 es continua y la muestra $X_1, ..., X_n$ también es continua. Si se sumplen estos supuestos entonces una muestra $X_1, ..., X_n \sim F_0$ $i.i.d$ genera muestras $U_1, ..., U_n \sim Uniforme(0, 1)$.
- Como consecuencia al punto anterior, D_n no depende de F_0.
- Si F_0 no es continua entonces K no es la verdadera función de distribución asintótica.

Test de Kolgomorov-Smirnov

Desarrollo en R

Determinar si:

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

Test de Kolgomorov-Smirnov

Ventajas

- La distribución del estadístico de la prueba Kolgomorov-Smirnov en sí no depende de la función de distribución acumulativa subyacente que se esté probando.
- Es una prueba exacta (la prueba de bondad de ajuste de X^2 depende de un tamaño de muestra adecuado para que las aproximaciones sean válidas).

Desventajas

- Solo se aplica a distribuciones continuas.
- Tiende a ser más sensible cerca del centro de la distribución que en las colas.
- Si los parámetros de ubicación, escala y forma se estiman a partir de los datos, la región crítica de la prueba Kolgomorov-Smirnov ya no es válida. Por lo general, debe determinarse mediante simulación.