

Estadística III

Pruebas de bondad de ajuste

Alejandro López Hernández

FES Acatlán - UNAM

March 27, 2020

- 1 Introducción
- 2 Prueba de Kolgomorov-Smirnov
- 3 Prueba de Cramér-von Mises
- 4 Prueba de Anderson-Darling
- 5 Pruebas de 2 Muestras

La idea de las pruebas de bondad de ajuste es comparar la función de distribución de nuestros datos (\hat{F}_n) con una función de distribución dada (F_0). Nuestro objetivo será encontrar estadísticos que nos ayuden a aceptar o rechazar la siguiente prueba:

$$H_0 : F = F_0$$

Ejemplo: Supongamos que tenemos los datos

$$X = 0.254, 1.23, 4.566, 2.165, 1.23, 1.829, 5, 3.23$$

Una pregunta interesante es si los datos tiene una distribución *uniforme*, para probar la hipótesis, es necesario calcular \hat{F}_n y con ella la prueba sería de la forma:

$$H_0 : F = F_0 \sim U(0, 5)$$

Algunas alternativas para medir las diferencias en las distribuciones son:

- $D_n^+ = \sup_{-\infty < t < \infty} (\hat{F}_n(t) - F_0(t))$
- $D_n^- = \sup_{-\infty < t < \infty} (F_0(t) - \hat{F}_n(t))$
- $D_n = \sup_{-\infty < t < \infty} |F_0(t) - \hat{F}_n(t)| = \max(D_n^+, D_n^-)$
- $V_n = D_n^+ + D_n^-$
- $C_n = \int (F_0(t) - \hat{F}_n(t))^2 dF_0(t)$
- $A_n = \int \frac{(\hat{F}_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t)$
- $w_{n,k,g} = \int (F_0(t) - \hat{F}_n(t))^k g(F_0(t)) dF_0(t)$
- $w_{n,k,g} = \int (F_0(t) - \hat{F}_n(t))^k g(F_0(t)) dF_0(t)$

La prueba de Kolgomorov-Smirnov se define como

$$D_n = \sup_{-\infty < t < \infty} |F_0(t) - \hat{F}_n(t)|$$

Se define de esa manera debido al teorema de Gilvenko-Cantelli, que nos dice que $\sup_{-\infty < t < \infty} |F(t) - \hat{F}_n(t)| \rightarrow 0$ a.s, es decir que la máxima distancia entre la distribución empírica y la real tiende a 0. Por lo tanto si nuestra F_0 es la distribución real, se espera que D_n sea pequeño.

Para calcular D_n solo es necesario conocer las observaciones X_1, X_2, \dots, X_n , si $X_{\{i\}}$ es el i -ésimo estadístico de orden, se puede probar que

$$D_n = \max_{1 \leq i \leq n} \max\left(\frac{i}{n} - F_0(X_{\{i\}}), F_0(X_{\{i\}}) - \frac{i-1}{n}\right)$$

Con el siguiente resultado encontraremos la distribución asintótica de D_n

Teorema 1

Sea $X_1, X_2, \dots, X_n \sim F_0$ y sea $D_n = \sup_t |F_0(t) - \hat{F}_n(t)|$ entonces, suponiendo que F_0 es continua;

$$\sqrt{n}D_n \rightarrow \sup_{0 \leq t \leq 1} |B(t)| \quad \text{en distribución}$$

Donde $B(t)$ es un puente Browniano.

Afortunadamente, se puede desarrollar la distribución del puente browniano.

Proposición 1

Bajo la hipótesis H_0 entonces

$$\lim_n \mathbb{P}(\sqrt{n}D_n \leq \lambda) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2}$$

Con esta distribución podemos calcular cuantiles y derivar regiones de rechazo.

La idea de este estadístico es medir el area que separa \hat{F}_n de F_0 , el estadístico se define como:

$$C_n = \int (F_0(t) - \hat{F}_n(t))^2 dF_0(t)$$

De forma analoga a el estadístico D_n , C_n se puede escribir en función de los estadísticos de orden:

$$C_n = \frac{1}{12n} + \sum_{i=1}^n \left(F_0(X_{\{i\}}) - \frac{2i-1}{2n} \right)^2$$

Con el siguiente resultado encontraremos la distribución asintótica de C_n

Teorema 2

Sea $X_1, X_2, \dots, X_n \sim F_0$ y sea $C_n = \int (F_0(t) - \hat{F}_n(t))^2 dF_0(t)$ entonces, suponiendo que F_0 es continua;

$$nC_n \rightarrow \int_0^1 B^2(t) dt \quad \text{en distribución}$$

Donde $B(t)$ es un puente Browniano.

De forma similar que con D_n se puede calcular la distribución asintótica de C_n

Proposición 2

Bajo la hipótesis H_0 entonces

$$\lim_n \mathbb{P}(nC_n > \lambda) = \frac{1}{\pi} \sum_{j=1}^{\infty} (-1)^{j+1} \int_{(2j-1)^2\pi^2}^{4j^2\pi^2} \sqrt{\frac{-\sqrt{y}}{\sin(\sqrt{y})}} \frac{e^{-\frac{xy}{2}}}{y} dy$$

De igual forma podemos crear regiones de rechazo con esta distribución.

La idea es también medir el área en que separa \hat{F}_n de F_0 , sin embargo, el término $F_0(t)(1 - F_0(t))$ busca tener una mayor ponderación en la región donde la distribución $F_0(t)$ tiene mayor incertidumbre. El estadístico se define como

$$A_n = \int \frac{(\hat{F}_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t)$$

De forma analoga a los estadísticos anteriores A_n se puede escribir como:

$$A_n = -n - \frac{1}{n} \left[\sum_{i=1}^n (2i-1) (\log F_0(X_{\{i\}}) + \log(1 - F_0(X_{\{n-i+1\}}))) \right]$$

Con el siguiente resultado encontraremos la distribución asintótica de A_n

Teorema 3

Sea $X_1, X_2, \dots, X_n \sim F_0$ y sea $A_n = \int \frac{(\hat{F}_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t)$ entonces, suponiendo que F_0 es continua;

$$nA_n \rightarrow \int_0^1 \frac{B^2(t)}{t(1-t)} dt \quad \text{en distribución}$$

Donde $B(t)$ es un puente Browniano.

Supongamos que tenemos dos muestras independientes, $X_1, X_2, \dots, X_n \sim F_0$ y $Y_1, Y_2, \dots, Y_m \sim F$, entonces para comparar las 2 muestras hacemos la prueba de hipótesis $H_0 : F = F_0$, para construir la prueba usamos F_n, G_m la función de distribución empírica de X_i y Y_i respectivamente. De forma análoga a lo desarrollado para las prebas de una muestra, necesitamos una forma de comparar ambas distribuciones. Para ello usamos el estadístico:

$$D_{n,m} = \sup_{0 \leq t \leq 1} |F_n(t) - G_m(t)|$$

La distribución límite de $D_{n,m}$ es la siguiente:

Teorema 4

Sea $X_1, X_2, \dots, X_n \sim F_0$ y $Y_1, Y_2, \dots, Y_m \sim F$ donde F_0 y F son distribuciones continuas. Bajo la hipótesis $H_0 : F = F_0$:

$$\lim_{n,m} \mathbb{P}\left(\sqrt{\frac{nm}{n+m}} D_{n,m} \leq \lambda\right) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2}$$

dado que para algún $0 < \gamma < 1$, $\frac{m}{n+m} \rightarrow \gamma$