

Estadística III

Bootstrap y Jackknife

Alejandro López Hernández

FES Acatlán - UNAM

May 2, 2020

1 Bootstrap

2 Jackknife

El bootstrap es un método de remuestreo el cual nos proporciona información acerca de un funcional T de una muestra X_1, \dots, X_n . Una forma en la que se podría tener mas información de T es la situación cuando tenemos muchas muestras aleatorias, y podemos calcular T para cada una de las muestras, de esta manera podrías conocer mas acerca de la distribución de T .

Sin embargo, con el bootstrap no es necesario tener mas muestras ya que se realizan *remuestreos* de la unica muestra con la que contamos.

Definición 1

Sea $X_1, \dots, X_n \sim F$ una muestra aleatoria y T un funcional de F . La distribución bootstrap de T está definida como

$$H_{Boot}(x) = \mathbb{P}_{\hat{F}_n}(T(X_1^*, \dots, X_n^*) \leq x)$$

Donde X_1^*, \dots, X_n^* es una muestra aleatoria proveniente de \hat{F}_n

La distribución de $H_{Boot}(x)$ se puede utilizar para estimar quantiles o la varianza de T .
La forma de calcular la varianza bootstrap es

Varianza Bootstrap

- ① Extrae una muestra aleatoria de $X_1^*, \dots, X_n^* \sim \hat{F}_n$
- ② Calcula $T_n^* = T(X_1^*, \dots, X_n^*)$
- ③ Repite el paso 1 y 2 B veces, para obtener $T_{n,1}^*, \dots, T_{n,B}^*$
- ④ Sea

$$v_{boot} = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^B T_{n,r}^* \right)^2$$

Podemos utilizar v_{boot} para generar intervalos de confianza para T con la siguiente formula

$$T_n \pm z_{\alpha/2} \sqrt{v_{\text{boot}}}$$

Sin embargo, este intervalo solo es bueno si T_n tiene una distribución similar a la normal.

Sea $\theta = T(F)$ y $\hat{\theta}_n = T(\hat{F}_n)$, y definimos $R_n = \hat{\theta}_n - \theta$, sea $H(x)$ la distribución de R_n . Entonces definimos nuestro intervalo $C_n^* = (a, b)$ con

$$a = \hat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \quad \text{y} \quad b = \hat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

Notemos que $\mathbb{P}(a < \theta < b) = 1 - \alpha$ por lo tanto C_n^* es un intervalo de exactamente $1 - \alpha$ de confianza. Sin embargo, a y b dependen de la distribución de H pero se pueden estimar usando bootstrap.

Podemos estimar la distribución H como:

$$\hat{H}(r) = \frac{1}{B} \sum_{b=1}^B 1_{\{R_{n,b}^* \leq r\}}$$

Donde $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$, con la distribución empírica $\hat{H}(r)$ podemos estimar cuantiles de R_n , con $r_\beta^* = \inf\{x : \hat{H}(x) \geq \beta\}$, si θ_β^* es el cuantil β de θ se puede probar que $r_\beta^* = \theta_\beta^* - \hat{\theta}$ por lo tanto

$$\hat{a} = \hat{\theta}_n - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - r_{1-\alpha/2}^* = 2\hat{\theta}_n - \theta_{1-\alpha/2}^*$$

$$\hat{b} = \hat{\theta}_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - r_{\alpha/2}^* = 2\hat{\theta}_n - \theta_{\alpha/2}^*$$

Intervalos de confianza Bootstrap

El intervalo pivotal de $1 - \alpha\%$ confianza de Bootstrap es

$$C_n = \left(2\hat{\theta}_n - \hat{\theta}_{1-\alpha/2,B}^*, 2\hat{\theta}_n - \hat{\theta}_{\alpha/2,B}^* \right)$$

El jackknife es método para aproximar el sesgo y la varianza de los estimadores. Sea T_n un estimador de cierta cantidad θ , entonces definimos $\text{sesgo}(T_n) = \mathbb{E}(T_n) - \theta$ como el sesgo del estimador. Definimos $T_{(-i)}$ como el estadístico calculado excluyendo la i -ésima observación. El estimador del sesgo *jackknife* se define como

$$b_{jack} = (n - 1)(\bar{T}_n - T_n)$$

Donde $\bar{T} = \frac{1}{n} \sum_{i=1}^n T_{(-i)}$ Derivado de este estimado de la varianza podemos corregir nuestro estadístico T_n como $T_{jack} = T_n - b_{jack}$ y se puede probar que $\mathbb{E}(T_{jack}) = O(\frac{1}{n^2})$

Notemos que podemos reescribir $T_{jack} = \frac{1}{n} \sum_{i=1}^n \tilde{T}_i$, donde $\tilde{T}_i = nT_n - (n-1)T_{(-i)}$, estos valores son llamados *pseudo-valores*. Se utilizan para estimar la varianza de T_n

$$v_{jack} = \frac{\tilde{s}^2}{n}$$

donde

$$\tilde{s}^2 = \frac{\sum_{i=1}^n \left(\tilde{T}_i - \frac{1}{n} \sum_{j=1}^n \tilde{T}_j \right)^2}{n-1}$$

Esta estimación bajo ciertas condiciones converge a la varianza real.

Teorema

Sea $\mu = \mathbb{E}(X_i)$ y $\sigma^2 = \text{Var}(X_i) < \infty$ y supongamos que $T_n = g(\bar{X}_n)$ donde g es una función continua y diferenciable en μ . Entonces

$$\frac{T_n - g(\mu)}{\sigma_n^2} \rightarrow N(0, 1)$$

donde $\sigma_n^2 = \frac{1}{n}(g'(\mu))^2\sigma^2$ y la estimación de la varianza jackknife es consistente, es decir

$$\frac{v_{jack}}{\sigma_n^2} \rightarrow 1$$