Universitat Autònoma de Barcelona
Facultat de Ciències
Estadística
M. Barcelona

Seminario 7: the Bootstrap

Ejercicio 1

[1]. En un estudio clínico controlado *Physician's Health Study I* que empezó en 1982 y acabó en 1987, participaron más de 22000 médicos. Los participantes fueron asignados aleatoriamente en 2 grupos: (i) Aspirina y (ii) Placebo, donde el grupo Aspirina tomó 325 mg de aspirina cada dos días. Al final del ensayo, se evaluó el número de participantes que sufrieron infarto de miocardio. Los resultados se presentan en la siguiente tabla:

	MyoInf	NoMyoInf	Total
Aspirina	104	10933	11037
Placebo	189	10845	11034

Popularmente, se utiliza como medida para evaluar los resultados en ensayos clínicos el *Risk Ratio (RR)* que se corresponde con el ratio entre las proporciones de casos (riesgos) en los dos grupos/tratamientos. De la tabla,

$$RR = R_a/R_p = \frac{104/11037}{189/11034} = 0{,}55.$$

Del RR se puede extraer que el riesgo de sufrir un infarto de miocardio en el grupo Placebo es aproximadamente $1/0{,}55 = 1{,}82$ veces mayor que en el grupo Aspirina. Usar el bootstrap para calcular el error estándar del RR.

Ejercicio 2

A continuación, veremos un ejemplo demostrando que el bootstrap puede fallar. Consideramos $X_1, \ldots, X_n$ muestras i.i.d. de una distribución uniforme en $[0, \theta]$, $\theta > 0$. Denotamos su distribución como F_θ.

(a) Demuestra que el estimador de máxima verosimilitud para θ es $\hat{\theta}_n = X_{(n)} = max_i X_i$.

(b) Demuestra que la distribución de $\hat{\theta}_n$ es

$$G(t) = \mathbf{P}_{F_\theta}(\hat{\theta}_n \le t) = \begin{cases} 0 & t < 0 \\ (t/\theta)^n & 0 \le t \le \theta \\ 1 & t > \theta \end{cases}.$$

(c) Utilizar el apartado anterior para encontrar la expresión analítica para la varianza de $\hat{\theta}_n$, $\mathbf{V}_{F_\theta}(\hat{\theta}_n)$.

(d) Escribe un script que genere n observaciones a partir de una variable aleatoria distribuida uniformemente en $[0, \theta]$ e implemente *bootstrap no-paramétrico* generando B muestras de bootstrap. Utiliza el código con $n = 25$ y $B = 5000$, calcular $\hat{\theta}_n$ y $\hat{\theta}^{\star}_{n,b}$, $b = 1, \ldots, B$. Utiliza las muestras del bootstrap para aproximar $\mathbf{V}_{F_\theta}(\hat{\theta}_n)$ y compararlo con al respuesta a la pregunta anterior (c). Repite el experimento diversas veces y prueba a modificar los valores de B y n. ¿Qué se observa?

(e) De los experimentos en (d) se debería notar que no es posible usar bootstrap no-paramétrico para estimar correctamente la varianza. Vamos a intentar entender qué es lo que no funciona. Definimos $T_n = n(\theta - \hat{\theta}_n)$ y $T_n^\star = n(\hat{\theta}_n - \hat{\theta}_n^\star)$. Demostrar que para $t \geq 0$,

$$\mathbf{P}_{\hat{F}_n}(T_n^\star \leq t) \geq \mathbf{P}_{\hat{F}_n}(T_n^\star \leq 0) = 1 - (1 - 1/n)^n,$$

siendo la primera desigualdad trivial.

(f) Demuestra que

$$\liminf_{n\to\infty} \sup_t |\mathbf{P}_F(T_n \leq t) - \mathbf{P}_{\hat{F}_n}(T_n^\star \leq t)| \geq 1 - e^{-1}.$$

Esto significa que la distribución de $\hat{\theta}_n$ calculada a partir del bootstrap no-paramétrico es significativamente distinta que la distribución real. Vuelve a la pregunta (d) y dibuja el histograma de $\hat{\theta}_{n,b}^\star$ y comprueba que se cumple el enunciado anterior.

(g) El bootstrap paramétrico genera muestras $X_1^\star, \ldots, X_n^\star$ a partir de una distribución uniforme en $[0, \hat{\theta}_n]$, en lugar de generarse a partir de la función de distribución empírica. Prueba que

$$\sup_t |\mathbf{P}_F(T_n \leq t) - \mathbf{P}_{F_{\hat{\theta}_n}}(T_n^\star \leq t)| \to 0,$$

en distribución para $n \to \infty$ (podéis suponer convergencia uniforme en t). Verifica que se cumple a partir de generar $B = 5000$ simulaciones bootstrap.

Hint: prueba que T_n converge a una distribución exponencial con media θ. Verificarlo experimentalmente también.

Referencias

[1] Kvam, P.H. and Vidakovic, B. (2007) *Nonparametric Statis- tics with Applications to Science and Engineering*, Wiley.