Método Bootstrap

Rodrigo Barrera

Contenidos

Contenidos

Antecedentes históricos

Descripción: técnica en la clase de métodos que remuestrean a partir del set de datos original. Algunos procedimientos similares:

- **Jackknife:** Quenouille (1949).
- **Métodos de permutaciones:** Fisher y Pitman (1930).

Nombre: "to pull oneself up by one's (own) bootstraps". Conecta con la idea de realizar "lo imposible" con información limitada (tamaño de la muestra, desconocimiento de la distribución).

Línea de tiempo:

- **Efron (1979):** unifica y conecta ideas sobre Bootstrap no paramétrico, para observaciones independientes e idénticamente distribuidas (remuestreo con reposición).
- **1980's:** Avances teóricos y estudios de simulaciones que compararon el Bootstrap y sus variantes con estimadores de otros procedimientos.
- **1990's:** cambio de foco a un desarrollo más práctico: aplicación a diversos contextos y cómo se comportan los resultados de distintas variantes.

Características

- **Computacionalmente intensivos:** en problemas prácticos, cuando la estimación es compleja, se requieren múltiples muestras bootstrap (Chernick (2008) sugiere extraer como mínimo 1000 muestras bootstrap).
- **Versión clásica o No Paramétrica de Efron:** no asume una distribución específica, solamente que la muestra es iid.
- **Versión Paramétrica:** asume un modelo paramétrico (Normal, Exponencial, etc.) y se utiliza para conocer información sobre algún parámetro.

De acuerdo con Chernick (2008):

- **Muestra muy pequeña:** con pocos valores desde donde remuestrear, Bootstrap subrepresentará la verdadera variabilidad del estadístico bajo estudio.
- **Varianza no finita:** Athreya (1987) estudia casos específicos en donde Bootstrap podría tener buen desempeño para la media.
- **Estimación de valores extremos:** método falla cuando el estadístico a describir es el mínimo o máximo.
- **Encuestas:** si la relación entre tamaño muestral y tamaño poblacional no es pequeña, la variabilidad estimada por Bootstrap se ve reducida.

Contenidos

Bootstrap No Paramétrico

Suponer una muestra iid: $X_1, X_2, ..., X_n \sim F$

Se busca estudiar (ejemplo) la Mediana de la distribución. Calculamos: $M_n = \text{median}(X_1, ..., X_n)$

B muestras bootstrap (muestreo con reposición): $\begin{matrix} X_1^{*(1)}, & \cdots, & X_n^{*(1)} \\ \vdots & \ddots & \vdots \\ X_1^{*(B)}, & \cdots, & X_n^{*(B)} \end{matrix}$

B medianas: $M_n^{*(1)}, ..., M_n^{*(B)}$

1. Estimador de la varianza: $\widehat{Var_B(M_n)} = \frac{1}{B-1}\sum_{l=1}^{B}(M_n^{*(l)} - \bar{M}_B^*)^2$
2. Estimador del sesgo: $\widehat{Bias_B(M_n)} = \frac{1}{B}\sum_{l=1}^{B}(M_n^{*(l)}) - M_n$
3. Estimador de MSE: $\widehat{MSE(M_n)} = \frac{1}{B}\sum_{l=1}^{B}(M_n^{*(l)} - M_n)^2$
4. IC $(1-\alpha)$: $M_n \pm z_{1-\alpha/2}\sqrt{\widehat{Var_B(M_n)}}$ (Normalidad asintótica)
5. IC $(1-\alpha)$: $[M^*_{n,\alpha/2}, M^*_{n,1-\alpha/2}]$ (Percentiles)

Definiciones

Siguiendo la notación de Dikta y Scheer (2021):

- Muestra iid: $X_1, ..., X_n \sim F$
- Estadístico de interés: $T_n(F) = T_n(X_1, ..., X_n; F)$
- Se busca aproximar su distribución, por ejemplo para estudiar IC.
- Bootstrap clásico: $T_n(\hat{F}) = T_n(F_n) = T_n(X_1^*, ..., X_n^*; F_n)$, en donde $X_1^*, ..., X_n^* \sim F_n$ es muestra iid (muestra bootstrap).
- $E(X) = \mu$ y $Var(X) = \sigma^2$

Ley Débil de los Grandes Números (Bootstrap)

Asumiendo que $\int |h(x)|F(dx) < \infty$, entonces con probabilidad 1, para cada $\epsilon > 0$:

$$P_n^*(|\frac{1}{n}\sum_{i=1}^{n} h(X_i^*) - \int h(x)F(dx)| > \epsilon) \to 0$$

cuando $n \to \infty$

En este punto se asume:

- $E(X) = \mu$
- $V(X) = \sigma^2$

Justificación Matemática

Teorema de Límite Central (Bootstrap)

Considerando $E(X^2) < \infty$ y $E(X) = \mu$, entonces con probabilidad 1:

$$\sup_{x \in \mathbb{R}} |P(\frac{1}{n^{-1/2}} \cdot (\bar{X}_n - \mu) \leq x) - P_n^*(\frac{1}{n^{-1/2}} \cdot (\bar{X}_n{}^* - \bar{X}_n) \leq x)| \to 0$$

cuando $n \to \infty$

Corolario

Bajo los supuestos del Teorema anterior, con probabilidad 1:

$$\sup_{x \in \mathbb{R}} |P(\frac{1}{n^{-1/2}} \cdot (\bar{X}_n - \mu)/s_n \leq x) - P_n^*(\frac{1}{n^{-1/2}} \cdot (\bar{X}_n{}^* - \bar{X}_n)/s_n^* \leq x)| \to 0$$

cuando $n \to \infty$

Justificación Matemática

Teorema de Berry-Esseen

Sea $X_1, ..., X_n$ una muestra iid, con media μ y varianza σ^2. Sea $\mu_3 = E[|X_i - \mu|^3] < \infty$. Sea $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ la media muestral y sea Φ la cdf asociada a una v.a. que distribuye $N(0,1)$. Sea $Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. Entonces:

$$\sup_x |P(Z_n \leq x) - \Phi(x)| \leq \frac{C \cdot \mu_3}{\sigma^3 \cdot \sqrt{n}}, \qquad C \geq \frac{\sqrt{10} + 3}{6\sqrt{2\pi}}$$

Contenidos

Se supone en este caso que se quiere recabar información sobre el parámetro de una distribución.

Ejemplo: muestra iid $X_1, X_2, ..., X_n \sim N(0, \sigma^2)$, σ^2 es desconocida.

Estimador insesgado para la varianza poblacional:
$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$

$$B \text{ muestras bootstrap: } \begin{array}{ccc} X_1^{*(1)}, & \cdots, & X_n^{*(1)} \sim N(0, S_n^2) \\ \vdots & \ddots & \vdots \\ X_1^{*(B)}, & \cdots, & X_n^{*(B)} \sim N(0, S_n^2) \end{array}$$

B varianza muestrales: $S_n^{2*(1)}, ..., S_n^{2*(B)}$

1. Estimador de la varianza:
$\widehat{Var_B(S_n^2)} = \frac{1}{B-1}\sum_{l=1}^{B}(S_n^{2*(l)} - \bar{S_B^{2*}})^2$
2. Estimador del sesgo: $\widehat{Bias_B(S_n)} = \frac{1}{B}\sum_{l=1}^{B}(S_n^{2*(l)}) - S_n$
3. Estimador de MSE: $\widehat{MSE(S_n^2)} = \frac{1}{B}\sum_{l=1}^{B}(S_n^{2*(l)} - S_n^2)^2$
4. IC $(1-\alpha) : S_n^2 \pm z_{1-\alpha/2}\sqrt{\widehat{Var_B(S_n^2)}}$ (Normalidad asintótica)
5. IC $(1-\alpha) : [S_{n,\alpha/2}^{2*}, S_{n,1-\alpha/2}^{2*}]$ (Percentiles)

Contenidos

Bibliografía

Chernick, M. (2008). *Bootstrap Methods. A guide for practitioners and Researchers.* Wiley.

Dikta, G., Scheer, M. (2021). *Bootstrap Methods with applications in R.* Springer.

Chernick, M., LaBudde, R. (2011). *An introduction to bootstrap methods with applications to R.* Wiley.

Politis, D., Romano, J., Wolf, M. (1999). *Subsampling.* Springer.

Urban, J.S. (1994). *Computer Intensive Statistical Methods.* Chapman & Hall.