Inferencia Estadística

Marisol García Peña

Departamento de Matemáticas
Pontificia Universidad Javeriana

Bogotá, 2022

Bootstrap

- Pasamos de la probabilidad a la estadística.
- Situaciones donde se conoce la población $\Longrightarrow$ se desconoce.
- Si todo lo que tenemos son datos y una estadística estimada a partir de los datos $\Longrightarrow$ estimar la distribución muestral de la estadística.
- Bootstrap.

Bootstrap (cont.)

- Datos de un estado en particular, peso medio de 1009 bebés en la muestra es 3448.26 g.
- Se está interesado en μ, verdadero peso medio de todos los bebés nacidos en 2004 en el estado.
- Diferentes muestras del mismo tamaño $\Longrightarrow$ medias muestrales diferentes.
- ¿Exactitud de la estimación 3448.26 de μ?

Bootstrap (cont.)

- Bootstrap $\Longrightarrow$ procedimiento que usa la muestra dada para crear una nueva distribución $\Longrightarrow$ distribución bootstrap.
- Esta distribución, aproxima la distribución muestral para la media muestral (otras estadísticas).

Bootstrap (cont.)

- Considerando un subconjunto de los pesos de nacimiento, tres observaciones, 3969, 3204 y 2892.
- Encontrar la distribución bootstrap de la media.
- Seleccionar muestras de tamaño n (remuestras o muestras bootstrap) con reemplazo de la muestra original y luego calcular la media de cada remuestra.
- La muestra original se trata como la población.

Bootstrap (cont.)

- En el caso de las 3 observaciones $\Longrightarrow 3^3 = 27$ muestras de tamaño 3.
- $x^* \Longrightarrow$ observación remuestreada, $\overline{x}^*$ o $\widehat{\theta}^* \Longrightarrow$ estadística para la muestra bootstrap.
- La distribución bootstrap de la media será aproximadamente como la distribución muestral de la media (aprox. misma forma y dispersión).
- La media de la distribución bootstrap será la misma que la media de muestra original, no necesariamente la de la población original.

Bootstrap (cont.)

Idea Bootstrap

La muestra original se aproxima a la población de la cual es seleccionada. Remuestras a partir de la muestra inicial se aproxima a lo que se obtendría si pudieran tomarse muchas muestras de la población. La distribución bootstrap de una estadística basada en muchas muestras se aproxima a la distribución de muestreo de la estadística basada en muchas muestras.

Bootstrap (cont.)

Todas las posibles muestras de tamaño 3 de 3969, 3204 y 2892.

x_1^*	x_2^*	x_3^*	$\overline{x}^*$
3969	3969	3969	3969
3969	3969	3204	3714
3969	3969	2892	3610
3969	3204	3969	3714
3969	3204	3204	3459
3969	3204	2892	3355
3969	2892	3969	3610
3969	2892	3204	3355
3969	2892	2892	3251
3204	3969	3969	3714
⋮	⋮	⋮	⋮
2892	2892	2892	2892

Bootstrap (cont.)

- La desviación estándar de todas las medias de las remuestras es 266.
- Este valor es una estimación del error estándar (desviación estándar de la verdadera distribución muestral).
- Para 3 observaciones es difícil aproximarse con precisión a la población.

Bootstrap (cont.)

- Si se toman remuestras de tamaño $n = 1009$ de los 1009 pesos de nacimiento y se calcula la media de cada una.
- 1009^{1009} muestras posibles $\Longrightarrow$ exhaustivo.
- Mejor opción $\Longrightarrow$ Seleccionar muestras de tamaño 1009 con reemplzamiento de los datos y calcular la media en cada una.
- El proceso se repite, por ejemplo, 10000 para crear la distribución bootstrap.

Bootstrap (cont.)

- En este caso, la media es 3448.206 aprox. la misma que la media original 3448.26.
- La desviación estándar de la distribución bootstrap es 15.379 (error estándar bootstrap).
- La desviación estándar de los datos es 487.736.
- Error estándar bootstrap menor $\Longrightarrow$ Refleja el hecho de que un promedio de 1009 observaciones es más preciso (menos variable) que una simple observación.

Bootstrap (cont.)

Error estándar bootstrap

El error estándar bootstrap de una estadística es la desviación estándar de la distribución bootstrap de la estadística.

Bootstrap (cont.)

Bootstrap para una población

Dada una muestra de tamaño n de una población:

1. Seleccionar una remuestra de tamaño n con reemplazo a partir de la muestra original. Calcule la estadística que describe la muestra (media, proporción, varianza, etc.)

2. Repetir el proceso de remuestreo muchas veces, por ejemplo, 10000.

3. Construir la distribución bootstrap de la estadística. Inspeccionar su centro, dispersión y forma.

Bootstrap (cont.)

Distribución bootstrap y distribución muestral

Para la mayoría de las estadísticas, las distribuciones bootstrap aproximan la dispersión, el sesgo y la forma de la distribución muestral verdadera.

Si al comparar el centro de la distribución bootstrap con la estadística observada, estas son diferentes $\Longrightarrow$ sesgo.

Bootstrap (cont.)

Ejemplo

Considerando un ejemplo, donde ni la población ni la distribución muestral es normal.

$X \sim Gamma(r, \lambda)$, $E[X] = \frac{r}{\lambda}$ y $V[X] = \frac{r}{\lambda^2}$.

Sea $X_1, X_2, \ldots, X_n$ una m.a. entonces $X_1, X_2, \ldots, X_n \sim Gamma(r, \lambda)$ y $\overline{X} \sim Gamma(nr, n\lambda)$.

Tomando una m.a. de tamaño $n = 16$ de una distribución gamma, $Gamma(1, 1/2)$

Bootstrap (cont.)

	Media	Desv.est.
Población	2	2
Distribución muestral de $\overline{X}$	2	$0.5 = \sqrt{16/8^2}$
Muestra	2.07	1.58
Distribución bootstrap	2.07	0.38

Bootstrap (cont.)

- La distribución de la muestra no coincide exactamente con la distribución de la población.
- La distribución bootstrap es similar a la distribución muestral en forma y la dispersión es más **pequeña** (depende de la muestra aleatoria).
- La media de la distribución bootstrap coincide con la de la distribución empírica y no con la de la población.

Bootstrap (cont.)

Principio plug-in

Para estimar un parámetro, cantidad que describe la población, use la estadística correspondiente para la muestra.

- Este principio se usa bastante en estadística.
- Error estándar de $\overline{X}$ es $\sigma/\sqrt{n}$, σ desconocido $\Longrightarrow$ plug-in la estimación S para $S/\sqrt{n}$.
- La diferencia en bootstrap es que plug-in una estimación para toda la población, no solo para un resumen numérico de la población.

Bootstrap (cont.)

- Objetivo $\Longrightarrow$ Estimar la distribución muestral de una estadística.
- La distribución muestral depende de:
  - La población de estudio.
  - El procedimiento de muestreo.
  - La estadística.
- Distribución muestral de una estadística es el resultado de tomar muchas muestras de la población y calcular la estadística en cada una.
- El problema es que la mayoría de las veces la población es desconocida.

Bootstrap (cont.)

- El principio bootstrap es plug-in una estimación para la población y luego imitar el procedimiento de muestreo de la vida real y del cálculo de la estadística.
- La distribución bootstrap depende de:
  - Una estimación para la población.
  - El procedimiento de muestreo.
  - La estadística.
- Puede usarse la distribución empírica como una estimación para la poblacion.

Bootstrap (cont.)

Estimación de la distribución de la población

- F y f denotan las funciones de distribución acumulada y de densidad de una distribución desconocida.
- $X_1, X_2, \ldots, X_n$ una muestra aleatoria.
- Podrían hacerse supuestos sobre la población, por ejemplo, asumir que sigue una distr. exponencial, estimar λ con los datos y tomar muestras bootstrap de una exponencial con el $\widehat{\lambda}$.
- Bootstrap paramétrico.

Bootstrap (cont.)

- En bootstrarp $\Longrightarrow$ hacer pocos supuestos sobre la población.
- La idea es que los datos "digan" lo que puedan.
- No introducir sesgos al hacer suposiciones que pueden ser incorrectas.
- Recurre a la distribución empírica.

Bootstrap (cont.)

- $\widehat{F}(s) = \frac{1}{n}\{\text{número de puntos} \leq s\}$.
- Distribución discreta con probabilidad $\frac{1}{n}$ en cada uno de los puntos observados.
- $\widehat{f}(s) = \frac{1}{n}\{\text{número de puntos} = s\}$.
- Por ejemplo, si la muestra es 5,5,6,10,11,11,11,12, entonces $\widehat{f}(5) = \frac{2}{8}$, $\widehat{f}(6) = \frac{1}{8}$, $\widehat{f}(10) = \frac{1}{8}$, $\widehat{f}(11) = \frac{3}{8}$, $\widehat{f}(12) = \frac{1}{8}$ y $\widehat{f}(s) = 0$ e.o.c.

Bootstrap (cont.)

- En bootstrap raramente se escribe $\widehat{F}$ y $\widehat{f}$.
- Solo es necesario saber como seleccionar las muestras $\Longrightarrow$ muestreo a partir de las observaciones originales con igual probabilidad para cada una $\frac{1}{n}$.
- Algunas veces se necesita la media y la varianza de $\widehat{F}$ y esto depende de si la población es discreta o continua.
- Si es discreta, se tiene:

$$E_{\widehat{F}}[X] = \mu_{\widehat{F}} = \sum_{x} x\widehat{f}(x) = \sum_{i=1}^{n} x_i \frac{1}{n} = \overline{x}$$

$$V_{\widehat{F}}[X] = \sigma^2_{\widehat{F}} = E_{\widehat{F}}[(X - \mu_{\widehat{F}})^2] = \sum_{i=1}^{n} (x_i - \overline{x})^2 \frac{1}{n}$$

Bootstrap (cont.)

¿Qué tan útil es la distribución bootstrap?

- ¿Qué tan bien se aproxima la distribución bootstrap a la distribución muestral?
- Las estadísticas que se usan en bootstrap generalmente son estimadores.
- Para los estimadores más comunes y bajo supuestos de distribución bastante generales, se espera lo siguiente respecto al centro, dispersión, sesgo y simetría.

Bootstrap (cont.)

Centro El centro de la distribución bootstrap no es una aproximación precisa para el centro de la distribución muestral. La distribución bootstrap de $\overline{X}$ es centrada en $\overline{x} = \mu_{\widehat{F}}$, la media de la muestra, mientras que la distribución muestral es centrada en μ.

Dispersión La dispersión de la distribución bootstrap refleja la dispersión de la distribución muestral.

Sesgo El sesgo estimado bootstrap refleja el sesgo de la distribución muestral. El sesgo ocurre si la distribución muestral no está centrada en el parámetro.

Asimetría La asimetría de la distribución bootstrap refleja la asimetría de la distribución muestral.

Bootstrap (cont.)

Bootstrap no se usa para obtener mejores estimaciones de parámetros $\Longrightarrow$ las distribuciones bootstrap están centradas alrededor de las estadísticas $\widehat{\theta}$ calculadas a partir de los datos en lugar de los valores poblacionales desconocidos.

El muestreo bootstrap es útil para cuantificar el comportamiento de una estimación de parámetro $\Longrightarrow$ errores estándar, asimetría y sesgo o para clacular IC.

Bootstrap (cont.)

Ejemplo

El arsénico es un elemento natural en el agua subterránea de Bangladesh. Sin embargo, gran parte de esta agua subterránea se utiliza para beber en poblaciones rurales, por lo que la intoxicación por arsénico es un problema de salud grave.

Se tiene información de 271 pozos en Bangladesh.

Arsenic contamination of groundwater in Bangladesh 1
Arsenic contamination of groundwater in Bangladesh 2

Bootstrap (cont.)

- Media muestral $\overline{x} = 125.31$ y desviación estándar $s = 297.98$, medidas en microgramos por litro, o en partes por billón (ppb).
- US Environmental Protection Agency (EPA) $\Longrightarrow$ nivel de contaminantes máximo (MCL) de arsénico para suministros públicos de agua en 10 ppb. 57 % de las muestras superan ese nivel.
- Se toman muestras (remuestras) de tamaño 271 con reemplazo a partir de los datos y se calcula la media para cada remuestra.
- Media bootstrap 125.53 y error estándar 18.25.

Bootstrap (cont.)

Percentiles bootstrap para intervalos de confianza

El intervalo entre los percentiles 2.5 y 97.5 de la distribución bootstrap de una estadística, es el intervalo de confianza con percentiles bootstrap del 95 % para el parámetro correspondiente.

Bootstrap (cont.)

- Para los pesos de bebés al nacer $\Longrightarrow$ IC percentiles bootstrap del 95 % (3419,3478).
- Con un 95 % de confianza el verdadero peso medio de los bebés nacidos en un estado en particular in 2004 está entre 3419 y 3478 gr.
- Para los datos de arsénico $\Longrightarrow$ IC 95 % perc.bootstrap (92.9515,164.4418)
- Con un 95 % de confianza el verdadero nivel medio de arsénico está entre 92.95 y 164.44 μg/l. (microgramos por litro).
- Este último $(\overline{x} - 32.37, \overline{x} + 39.12)$ No es simétrico al rededor de la media $\Longrightarrow$ la asimetria de la distribución bootstrap.

Bootstrap (cont.)

- Un buen IC para la media no necesariamente es simétrico: uno de los límites puede estar "lejos" de la media muestral en dirección de cualquier outlier.
- Existen algunos valores extremos en las mediciones de arsénico: de las 271 observaciones, 8 son mayores que 1000 μg/l y 2 mayores que 2200 μg/l. La media muestral es 125.31.
- En la población no se sabe cuantos niveles de arsénico pueden ser tan "grandes" $\Longrightarrow$ Valores grandes pueden estar subrepresentados en la muestra.
- Tener solo 0.025 de probabilidad de perder un valor grande de la media, el intervalo debe extenderse a la derecha.
- Hay menos riesgo de perder un valor de la media al lado izuierdo $\Longrightarrow$ el límite inferior no necesita estar tan lejos de la media muestral.