Ejercicio 1:

Sea X_1, ..., X_n una muestra de tamaño n obtenida de una población desconocida. Considérese el procedimiento de bootstrap por remuestreo con reemplazo de tamaño n.

a) Muestre que existen $\binom{2n-1}{n}$ distintas muestras de bootstrap de tamaño n.

b) ¿Cuál es la probabilidad de que una muestra de bootstrap sea idéntica a la muestra original?

c) ¿Cuál es la muestra de bootstrap más probable de ser seleccionada?

d) ¿Cuál es la cantidad promedio de veces que X_i es seleccionada en una muestra de bootstrap de tamaño n?

Solución:

Inciso a)

Sea $\mathcal{X} = \{X_1, \ldots, X_n\}$ la muestra original. Una muestra bootstrap de tamaño n obtenida con reemplazo corresponde a elegir n elementos de $\mathcal{X}$ permitiendo repeticiones y sin importar el orden en que se generaron.

Equivale a contar el número de soluciones enteras no negativas del sistema

$$N_1 + N_2 + \cdots + N_n = n, \qquad N_i \in \mathbb{Z}_{\geq 0},$$

donde N_i es la cantidad de veces que X_i aparece en la muestra bootstrap. Esta redefinición del problema nos permite utilizar el teorema de *estrellas y barras*, cuyo enunciado es el siguiente:

Teorema 1 (Estrellas y barras)**.** *Sea $n, k \in \mathbb{N}$ positivos. El número de soluciones en enteros no negativos de*

$$x_1 + x_2 + \cdots + x_k = n$$

es

$$\binom{n+k-1}{k-1} = \binom{n+k-1}{n}.$$

Equivalente: es el número de multiconjuntos de cardinalidad n seleccionados de un conjunto de tamaño k.

En nuestro caso, $k = n$ y el número de muestras bootstrap distintas es

$$\binom{n+n-1}{n} = \binom{2n-1}{n} \qquad \square$$

Inciso b)

La muestra realizada por el bootstrap se considera sin importar el orden: dos réplicas que sólo difieren por permutar posiciones son el mismo resultado. La réplica bootstrap será "idéntica" a la original si cada observación X_j aparece exactamente una vez: $N_j = 1$ para todo $j = 1, \ldots, n$.

Podemos afirmar que $(N_1, \ldots, N_n) \sim \text{Multinomial}(n; 1/n, \ldots, 1/n)$, obteniendo

$$\mathbb{P}(N_1 = 1, \ldots, N_n = 1) = \frac{n!}{1! \cdots 1!} \left(\frac{1}{n}\right)^n = \frac{n!}{n^n}.$$

Esta es la probabilidad buscada de que la muestra bootstrap coincida con la original. Usando la aproximación de Stirling,

$$\frac{n!}{n^n} \sim \sqrt{2\pi n}\, e^{-n}, \qquad n \to \infty,$$

lo cual, por el término e^{-n}, muestra que el evento es extremadamente improbable para n moderado/grande. □

Inciso c)
Buscamos el multiconjunto más probable bajo

$$(N_1, \ldots, N_n) \sim \text{Multinomial}\big(n; 1/n, \ldots, 1/n\big).$$

Para un vector $\mathbf{n} = (n_1, \ldots, n_n)$ con $\sum n_i = n$,

$$\mathbb{P}(\mathbf{N} = \mathbf{n}) = \frac{n!}{n_1! \cdots n_n!} \left(\frac{1}{n}\right)^n.$$

El factor $(1/n)^n$ es constante; maximizar la probabilidad equivale a maximizar $\dfrac{n!}{n_1! \cdots n_n!}$ o, equivalentemente, minimizar el producto de factoriales $n_1! \cdots n_n!$ sujeto a $\sum n_i = n$.

Siendo nuestro objetivo minimizar el producto de factoriales, manteniendo la restricción de que la suma de los n_i es n, podemos razonar que la mejor estrategia es distribuir los n elementos de manera uniforme entre las n categorías. Cualquier desviación de esta uniformidad (por ejemplo, asignar un valor mayor a un n_i y un valor menor a otro n_j) aumentaría el producto de factoriales debido a la naturaleza creciente y convexa de la función factorial. Por ejemplo: $3!1! > 2!2!$, así como $2!0! > 1!1!$.

Así, el multiconjunto más probable es precisamente el que contiene cada observación exactamente una vez. Su probabilidad ya la calculamos en (b): $\dfrac{n!}{n^n}$. Ningún otro multiconjunto alcanza un valor mayor. □

Inciso d)

Fijado un índice i, cada extracción del bootstrap selecciona X_i con probabilidad $1/n$ de manera independiente. Por tanto

$$N_i \sim \text{Binomial}\big(n, 1/n\big).$$

Luego

$$\mathbb{E}[N_i] = n \cdot \frac{1}{n} = 1, \qquad \text{Var}(N_i) = n\frac{1}{n}\left(1 - \frac{1}{n}\right) = 1 - \frac{1}{n}.$$

Es decir, en promedio cada observación original aparece exactamente una vez en una muestra bootstrap (aunque aleatoriamente algunas se repiten y otras se omiten). De hecho,

$$\mathbb{P}(N_i = 0) = \left(1 - \frac{1}{n}\right)^n$$

$$\Rightarrow \lim_{n\to\infty} \mathbb{P}(N_i = 0) = e^{-1},$$

lo que muestra el fenómeno clásico: aproximadamente una fracción $e^{-1} \approx 0.368$ de las observaciones no aparece en una réplica bootstrap grande, mientras que alrededor de $1 - e^{-1} \approx 0.632$ sí aparece al menos una vez. □

Ejercicio 2:

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de una normal $N(\theta, 1)$ y suponga que $\bar{x}$ es un estimador de θ. Sean $X_1^*, X_2^*, \ldots, X_n^*$ una muestra bootstrap de $N(\theta, 1)$. Muestre que $\bar{X} - \theta$ y $\bar{X}^* - \bar{x}$ tienen la misma distribución $N(0, 1/n)$.

Solución:

Comenzamos determinando la distribución de la cantidad $\bar{X} - \theta$.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria (m.a.) tal que $X_i \sim N(\theta, 1)$ para $i = 1, \ldots, n$.

La media muestral se define como $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Sabemos que una combinación lineal de variables aleatorias (v.a.) normales independientes también sigue una distribución normal. Por lo tanto, $\bar{X}$ es normal.

Calculamos la esperanza y la varianza de $\bar{X}$:

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \theta = \frac{1}{n}(n\theta) = \theta$$

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{\text{indep.}}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} 1 = \frac{1}{n^2}(n) = \frac{1}{n}$$

Por lo tanto, la distribución de la media muestral es exacta:

$$\bar{X} \sim N\left(\theta, \frac{1}{n}\right)$$

Al centrar esta variable restando su media θ, obtenemos la distribución del pivote de muestreo:

$$\bar{X} - \theta \sim N\left(0, \frac{1}{n}\right)$$

Distribución del Pivote Bootstrap (Mundo Bootstrap)

No conocemos θ, por lo que estimamos la distribución de la población $N(\theta, 1)$ usando la muestra original.

El estimador de máxima verosimilitud (y de momentos) para θ es $\hat{\theta} = \bar{X}$. Sea $\bar{x}$ la media *observada* de nuestra muestra original.

El bootstrap paramétrico asume que el "mundo real" es la distribución que hemos estimado. Por lo tanto, generamos la muestra bootstrap $X_1^*, X_2^*, \ldots, X_n^*$ extrayéndola de $N(\hat{\theta}, 1)$, es decir:

$$X_i^* \sim N(\bar{x}, 1)$$

Definimos la media muestral bootstrap como $\bar{X}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$.

Esta es, nuevamente, una combinación lineal de v.a. normales, por lo que $\bar{X}^*$ sigue una distribución normal. Calculamos su esperanza y varianza (condicionales a la muestra original, $\bar{x}$):

$$E[\bar{X}^*|\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^* \middle| \bar{x}\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^*|\bar{x}] = \frac{1}{n}\sum_{i=1}^{n} \bar{x} = \bar{x}$$

$$Var(\bar{X}^*|\bar{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i^* \middle| \bar{x}\right) \overset{\text{indep.}}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i^*|\bar{x}) = \frac{1}{n^2}\sum_{i=1}^{n} 1 = \frac{1}{n}$$

La distribución de la media bootstrap es:

$$\bar{X}^* \sim N\left(\bar{x}, \frac{1}{n}\right)$$

El pivote bootstrap se construye restando el parámetro del "mundo bootstrap"(que es $\bar{x}$):

$$\bar{X}^* - \bar{x} \sim N\left(0, \frac{1}{n}\right)$$

Al comparar las distribuciones del pivote de muestreo y del pivote bootstrap, encontramos que son idénticas:

$$\bar{X} - \theta \sim N\left(0, \frac{1}{n}\right) \quad \text{y} \quad \bar{X}^* - \bar{x} \sim N\left(0, \frac{1}{n}\right)$$

Esto demuestra que, en este caso paramétrico normal, la distribución del pivote bootstrap replica exactamente la distribución del pivote de muestreo. □

Ejercicio 3:

Considere el conjunto de datos $\{2, 5, 3, 9\}$. Sean $x_1^*, x_2^*, x_3^*, x_4^*$ una muestra bootstrap de este conjunto de datos.

a) Encuentre la probabilidad de que el promedio de la muestra bootstrap sea igual a 2.

b) Encuentre la probabilidad de que el promedio de la muestra bootstrap sea igual a 9.

c) Encuentre la probabilidad de que el promedio de la muestra bootstrap sea igual a 4.

Solución:

Inciso a)

La única forma de que el promedio de la muestra bootstrap sea igual a 2 es que todos los elementos de la muestra bootstrap sean 2. Dado que estamos muestreando con reemplazo, la probabilidad de seleccionar 2 en cada uno de los 4 intentos es:

$$P(\bar{x}^* = 2) = P(x_1^* = 2) \cdot P(x_2^* = 2) \cdot P(x_3^* = 2) \cdot P(x_4^* = 2) = \left(\frac{1}{4}\right)^4 = \frac{1}{256} \approx 0.004.$$

Inciso b)

La única forma de que el promedio de la muestra bootstrap sea igual a 9 es que todos los elementos de la muestra bootstrap sean 9. Dado que estamos muestreando con reemplazo, la probabilidad de seleccionar 9 en cada uno de los 4 intentos es la misma que en el inciso anterior:

$$P(\bar{x}^* = 9) = P(x_1^* = 9) \cdot P(x_2^* = 9) \cdot P(x_3^* = 9) \cdot P(x_4^* = 9) = \left(\frac{1}{4}\right)^4 = \frac{1}{256} \approx 0.004.$$

Inciso c)

Para que el promedio de la muestra bootstrap sea igual a 4, necesitamos encontrar todas las combinaciones de los elementos $\{2, 5, 3, 9\}$ que sumen 16, ya que $4 \times 4 = 16$.

Las únicas combinaciones posibles que suman 16 son:

- $\{2, 2, 3, 9\}$
- $\{3, 3, 5, 5\}$

Cada combinación puede ocurrir en diferentes órdenes. La cantidad de permutaciones de cada combinación es:

- Para $\{2, 2, 3, 9\}$: $\frac{4!}{2!1!1!} = 12$ permutaciones.
- Para $\{3, 3, 5, 5\}$: $\frac{4!}{2!2!} = 6$ permutaciones.

La probabilidad de que el promedio de la muestra bootstrap sea igual a 4 es la suma de las probabilidades de cada una de estas combinaciones, multiplicadas por la probabilidad de

seleccionar cada combinación específica. Dado que cada elemento tiene una probabilidad de $\frac{1}{4}$ de ser seleccionado, la probabilidad total es:

$$P(\bar{x}^* = 4) = P(\{2, 2, 3, 9\}) + P(\{3, 3, 5, 5\}) = \frac{12}{256} + \frac{6}{256} = \frac{18}{256} = \frac{9}{128} \approx 0.070.$$

Por lo tanto, la probabilidad de que el promedio de la muestra bootstrap sea igual a 4 es aproximadamente 0.070. □