INFERENCIA ESTADÍSTICA II

Simulación y Bootstrap

Author
Victor Elvira Fernández, Tomás Ruiz Rojo, Juan Horrillo Crespo
Universidad de Valladolid

10 de noviembre de 2024

AVISO

Estos apuntes fueron creados de forma voluntaria por un grupo de estudiantes, invirtiendo tiempo, dedicación y esfuerzo para ofrecer información útil a la comunidad. Apreciamos cualquier apoyo que se nos quiera brindar, ya que nos ayuda a continuar con futuros proyectos de este tipo.

Si deseas colaborar en esta clase de proyectos puedes contactarnos y unirte o invitarnos a unas ricas patatas 5 salsas por el siguiente enlace:

Buy Me a Patatas 5 Salsas

https://www.buymeacoffee.com/ApuntesINdat

- Mail Juan Horrillo
- Mail Victor Elvira
- Mail Tomás Rojo

Si has colaborado de cualquier forma te agradecemos enormemente.

Índice

1. Introducción al Bootstrap

El bootstrap es un mecanismo generador de datos. Hasta ahora hemos trabajado en una situación en la que tenemos una muestra $X = (X_1, \ldots, X_n)$ v.a.i.i.d. de una distribución P_θ, $\theta = (\theta_1, \ldots, \theta_s)$ con el interés de obtener un estimador $T(\theta)$ razonable para θ o $g(\theta)$.

Todo ello en el concepto de **inferencia frecuentista**; se tiene un estimador del parámetro en base al que queremos hacer inferencia respecto a θ. Para esto es necesario conocer la distribución del estadístico (distribución exacta o asintótica). Supongamos que:

- No conocemos la distribución de los datos
- No se cumplen las condiciones de regularidad de Cramer-Rao

Cuando se da uno de los casos anteriores, el bootstrap puede ser una buena opción. Puede resultar interesante poder repetir un mecanismo generador de datos con el que se obtuvo la muestra original de forma que podamos obtener tales muestras como se quiera, y cada una de ellas obtendrá sus estadísticos correspondientes. Es decir, a partir de P_θ obtendremos:

$$\left.\begin{array}{ccccccc} X_{11} & X_{12} & \cdots & X_{1n} & \longrightarrow & T_n^1(X) \\ X_{21} & X_{22} & \cdots & X_{2n} & \longrightarrow & T_n^2(X) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ X_{5000,1} & X_{5000,2} & \cdots & X_{5000,n} & \longrightarrow & T_n^{5000}(X) \end{array}\right\} \text{Simulaciones si } P_\theta \text{ es conocida}$$

Podemos usar bootstrap para **aproximar cualquier característica** de la distribución y hacer inferencia a partir de los datos simulados.

Sin embargo, no siempre se conoce P_θ, si no que solo se dispone de los datos observados. En estos casos no es posible simpular a partir de P_θ. Podremos simularlos si somos capaces de estimar $F_\theta(\cdot)$, la verdadera función de distribución, La estimaremos a partir de la distribución empírica:

$$\hat{F}(X) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(x_i \leq x)}$$

Donde $\mathbb{1}$ se refiere a la función indicadora. *En los apuntes del tema 3 de probabilidad (2º curso) se ve otra forma de definir la función de distribución empírica.*

Definición: **Principio plug-in:** cualquier característica de una distribución pueede ser aproximada. El principio plug-in está apoyado por el **Teorema de Glivenko-Cantelli**:

$$\underset{x\in\mathbb{R}}{sup}\, |\hat{F}(X) - F_0(X)| \xrightarrow{c.s.} 0$$

La idea es simular por remuestreo el experimento y, a continuación, reajustar el modelo y recalcular estimadores con los datos simulados. Estos serían los pasos:

1. Estimar $F_0(X)$ a partir de la muestra

2. Simular $\hat{F}(X)$

Con el bootstrap podemos obtener también estimadores sobre el sesgo, intervalos de confianza y contrastes de hipótesis.

1.1. Aproximación bootstrap de la distribución EMV

Sean $X_1, \ldots, X_n$ con $F(\dot{)}$, $\hat{\theta}$ es el EMV de θ. El bootstrapsimula la distribución de $\hat{\theta}$.

1. Se estima $\hat{F}(X)$ $\begin{cases} \text{En el caso no paramétrico, a partir de la función de distribución empírica} \\ \text{En el caso paramétrico, estimando los parámetros necesarios} \end{cases}$
2. Generamos datos artificiales: las muestras bootstrap:
   - $X_1^*, \ldots, X_n^*$ con función de densidad $\hat{F}$ estimada de F
   - Se obtiene el EMV $\hat{\theta}^*$ basado en la muestra bootstrap

La idea del procedimiento es la siguiente: la distribución $\hat{\theta}^* - \hat{\theta}$ aproxima la distribución de $\hat{\theta} - \theta$. Al repetir los pasos anteriores en un proceso B veces se obtiene una versión bootstrap del EMV.

Existen dos tipos de bootstrap:

- **Bootstrap paramétrico:** si el estimador de F en el paso 1 es un estimador paramétrico
- **Bootstrap no paramétrico:** si usamos la función de distribución empírica para estimar F en el paso 1

Ejemplo

Sean $X_1, \ldots, X_n$ v.a.i.i.d de una $N(\mu, \sigma^2)$. Según lo visto en temas anteriores, sabemos que el EMV para $\theta = (\mu, \sigma)$ sigue la siguiente distribución:

$$\sqrt{(n)} \begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma} - \sigma \end{pmatrix} \xrightarrow{\mathcal{L}} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 4\sigma^2 \end{pmatrix} \right)$$

En este ejercicio supondremos que sabemos que los datos vienen de una distribución normal (aunque desconocemos los valores de μ y σ). Para conseguir una muestra bootstrap tenemos que seguir los pasos ya mencionados anteriormente:

1. Estimar μ y σ a partir del EMV
2. Simular desde nuestra nueva $\hat{F}$ (que sabemos que sigue una distribución normal)

El algoritmo se ejecuta de forma iterativa, y es un proceso muy laborioso a mano. Por ello, se deja como ejercicio propuesto al lector hacer un script que siga dichos pasos (se ve en clases prácticas).

La distribución de $\begin{pmatrix} \mu_i^* \\ \sigma_i^{2*} \end{pmatrix} - \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} \forall i = 1, 2, \ldots, B$ aproxima la distribución de $\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$

De igual forma, el histograma de $\mu_1^*, \ldots, \mu_B^*$ aproxima el de $\hat{\mu}$. Pasa lo mismo para σ^2

En el caso no paramétrico estimaríamos F a partir de la muestra original y su función de distribución empírica.

1.1.1. Estimador bootstrap de la varianza del EMV

Estamos estudiando la distribución de θ, por lo que tenemos que poder estudiar cualquier característica que dependa de θ. Estimaremos la varianza del EMV usando las B muestras bootstrap.

Justificación $\Longrightarrow$ La distribución de $\hat{\theta}^* - \hat{\theta}$ aproxima la de $\hat{\theta} - \theta$; por lo que la distribución de $Var(\hat{\theta}^* - \hat{\theta})$ aproxima la de $Var(\hat{\theta} - \theta)$, y, por tanto:

$$Var^*(\hat{\theta}^*) \approx Var^*(\hat{\theta})$$

1.2. Intervalos de confianza bootstrap (Método percentil)

Consideremos $X_1, \ldots, X_n$, con función de distribución $F_0(\dot{)}$, dependiente del parámetro s-dimensional $\theta = (\theta_1, .s, \theta_s)$. Para obtener un intervalo de confianza de nivel 1-α para la k-ésima componente de θ (θ_k) utilizaremos el **método percentil**.

Sean $\hat{\theta}^*_{k1}, \ldots, \hat{\theta}^*_{kB}$ las B versiones bootstrap del estimador $\hat{\theta}_k$, y sean $\hat{\theta}^*_{k,\frac{\alpha}{2}}$ y $\hat{\theta}^*_{k,1-\frac{\alpha}{2}}$ los cuantiles $\alpha/2$ y $1-\alpha/2$ respectivamente. El intervalo de confianza bootstrap percentil para θ_k será $\left(\hat{\theta}^*_{k,\frac{\alpha}{2}}, \hat{\theta}^*_{k,1-\frac{\alpha}{2}}\right)$. Es decir, el método percentil consiste en sustituir los extremos del intervalo por los percentiles correspondientes para nuestro nivel α.

$$P_\theta\left(\hat{\theta}^*_{k,\frac{\alpha}{2}} \leq \theta \leq \hat{\theta}^*_{k,1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

La justificación viene dada por la suposición de que, bajo las condiciones de regularidad apropiadas, el comportamiento de $\hat{\theta}$ como estimador de θ sea parecido al comportamiento de $\hat{\theta}^*$ como estimador de $\hat{\theta}$. En otras palabras, un intervalo que contenga a $\hat{\theta}^*_k$ con probabilidad aproximada $1-\alpha$ es también un intervalo que contiene a θ_k con probabilidad aproximada $1-\alpha$.

En cuanto a una transformación $g(.)$, si la función g es monótona creciente, el método percentil es **invariante a transformaciones**.

1.3. Contrastes de hipótesis bootstrap

Sean $X_1, \ldots, X_n$ i.i.d. con $f(., \theta), \theta \in \Theta$. Vamos a contrastar $H_0 : \theta \in \Theta_0$; $H_1 : \theta \notin \Theta_0$.

Sea T el estadístico de contraste y $\{T \geq C_\alpha\}$ la región crítica de nivel α. Existen situaciones en las que es posible calcular la distribución asintótica o exacta del estadístico T bajo la hipótesis nula. En ese caso, podemos determinar directamente C_α y calcular el p-valor del test. En el caso en el que calcular la distribución de T no sea posible podemos aproximarla mediante bootstrap.

- H_0 **es simple**: el bootstrap no es necesario, basta con simulación. La idea es simular un número grande de muestras (B) de tamaño n de la distribución; sean $T_1, \ldots, T_B$ los valores del estadístico T observados en las B muestras, podemos aproximar C_α por el cuantil $1-\alpha$ de estos B valores y el p-valor del test por la proporción de estos valores mayores que el observado para los datos originales.

- H_0 **es compuesta**: bootstrap paramétrico. Si $\hat{\theta}_0$ es el EMV de θ bajo H_0, se generan B muestras **bootstrap** de la distribución. Al igual que antes se calcula el p-valor del test a partir de la proporción de los valores mayores que el valor observado para los datos; pero esta vez con muestreo bootstrap.

Referencias

[1] Juan Camilo Yepes Borrero,
Apuntes Manuscritos Tema 2.
Universidad de Valladolid 2024.

[2] Yolanda Larriba González,
Apuntes INFE2 Tema 2.
Universidad de Valladolid 2023.