Jackknife & Bootstrap

Una Alternativa para Estimación y Precisión.

Samuel Jacobo Garavito Segura
Estudiante de Matemáticas de la Universidad Nacional de Colombia
sgaravito@unal.edu.co

Julian David Pulido Castañeda
Estudiante de Ingeniería Electrónica de la Universidad Nacional de Colombia
jdpulidoca@unal.edu.co

Publicado: 29-06-2022

Abstract

El siguiente artículo describe las técnicas de remuestreo **Jackknifing** y **Bootstraping** acompañadas de dos casos de uso implementados en Python y R correspondientemente para su ejemplificación.

Keywords: $Jackknife \cdot Bootstrap \cdot Estimador \cdot Remuestreo \cdot Estadístico \cdot Muestra \cdot Media \cdot Varianza$

Introducción

Los métodos de remuestreo estadístico son procedimientos que describen cómo utilizar económicamente los datos disponibles para estimar un parámetro de la población. El resultado puede ser tanto una estimación más precisa del parámetro *(como tomar la media de las estimaciones)* como una cuantificación de la incertidumbre de la estimación *(como añadir un intervalo de confianza)*.

Los métodos de remuestreo son muy fáciles de utilizar y requieren pocos conocimientos matemáticos. Son métodos fáciles de entender y aplicar en comparación con los métodos estadísticos especializados que pueden requerir una profunda habilidad técnica para seleccionar e interpretar. Por esta misma razón, su implementación computacional es una muy buena práctica que evita los cálculos complejos de la teoría estadística tradicional [1].

Entre las técnicas de remuestreo se encuentran el **Jackknifing** y el **Bootstrapping** las cuales son las más populares. A continuación, se estudiará el funcionamiento y la utilidad de cada una.

Jackknife

El método o técnica **Jackknife** es especialmente utilizado para estimar el sesgo y varianza de una población extensa. De igual manera, es útil para la corrección del sesgo de una estimación, obteniendo así un estimador nuevo y de mejor calidad.

Este fue el primer método de remuestreo, introducido por *Maurice Quenouille* en 1949 y nombrado por *John Tukey* en 1958 haciendo alusión a una *"navaja multiuso"*, herramienta de uso *"rápido y sencillo"* en situaciones donde no se dispone de otras.

Este método de remuestreo se encuentra dentro del marco de la estimación puntual. Dada $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria de una variable aleatoria unidimensional X con distribución F desconocida, $\theta \in \mathbb{R}$ un parámetro de la distribución y T_n un estimador de θ, en la mayoría de los casos de estimación puntual, estimadores no insesgados cumplen la siguiente propiedad:

$$E[T_n] \;=\; \theta + \frac{\alpha}{n} + O\left(\frac{1}{n^2}\right) \quad \text{para todo } n \in \mathbb{N}$$

donde claramente el sesgo corresponde al término $\frac{\alpha}{n} + O\left(\frac{1}{n^2}\right)$. Hecha la mención de esta propiedad, se procederá a introducir la siguiente definición.

Definición 1.0: *Dada una muestra aleatoria $X_1, X_2, X_3, \ldots, X_n$ denotada por $\mathcal{X}$, se define* ***la i-esima muestra de Jackknife*** *como la muestra formada eliminando la observación i-esima de la muestra original, es decir*:

$$J_i \;:=\; \{X_j \in \mathcal{X} \quad | \quad j \neq i\}$$

y sea $T_n := T_n(X_1, X_2, X_3, \ldots, X_n)$ *un estimador de un parámetro conocido* θ*, se define:*

$$T_{n-1,i} \;:=\; T_{n-1}(J_i)$$

como ***el i-esimo estadístico de Jackknife.*** [2]

Presentada esta definición, ahora se introducirá el concepto de *pseudo-valores de Jackknife.*

Definición 2.0: *Sea* $\mathcal{X}$ *una muestra aleatoria y* T_n *un estadístico. Sea* $T_{n-1,i}$ *el i-esimo estadístico de Jackknife. Se define el* ***i-esimo pseudo-valor de Jackknife*** *como:*

$$J_{n,i} \;=\; nT_n - (n-1)T_{n-1,i} \quad [2]$$

De este modo, al evaluar la media muestral de estos pseudo-valores se tiene que:

$$\begin{aligned}
\overline{J_n} &= \frac{1}{n}\sum_{i=1}^{n} J_{n,i} \\
&= \frac{1}{n}\sum_{i=1}^{n} nT_n - \frac{1}{n}\sum_{i=1}^{n}(n-1)T_{n-1,i} \\
&= \sum_{i=1}^{n} T_n - \frac{(n-1)}{n}\sum_{i=1}^{n} T_{n-1,i} \\
&= nT_n - \frac{(n-1)}{n}\sum_{i=1}^{n} T_{n-1,i}
\end{aligned}$$

Al ser la media muestral un estadístico, se ha obtenido un nuevo estimador para el cual se elimina un término del sesgo, lo que se estaba buscando, esto se debe a que:

$$\begin{aligned} E[\overline{J_n}] &= E[nT_n] - E\left[\frac{(n-1)}{n}\sum_{i=1}^{n} T_{n-1,i}\right] \\ &= n\left(\theta + \frac{\alpha}{n} + O\left(\frac{1}{n^2}\right)\right) - (n-1)\left(\theta + \frac{\alpha}{n-1} + O\left(\frac{1}{n^2}\right)\right) \\ &= \alpha - \frac{n\alpha}{n-1} + \theta + \frac{\alpha}{n-1} + O\left(\frac{1}{n^2}\right) \\ &= \frac{(n-1)\alpha - (n-1)\alpha}{n-1} + \theta + O\left(\frac{1}{n^2}\right) \\ &= \theta + O\left(\frac{1}{n^2}\right) \end{aligned}$$

De esta manera, el estimador del sesgo del parámetro θ corresponde a $T_n - \overline{J_n}$, que indica que:

$$\begin{aligned} T_n - \overline{J_n} &= \frac{n-1}{n}\sum_{i=1}^{n} T_{n-1,i} - (n-1)T_n \\ &= (n-1)(\overline{T_n} - T_n) \end{aligned}$$

donde $\overline{T_n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} T_{n-1,i}$.

Ya sabiendo esto, se procederá a definir *el estimador Jackknife del sesgo de* T_n.

Definición 3.0: *Sea* $\mathcal{X}$ *una muestra aleatoria y* T_n *un estadístico. Se define* ***el estimador Jackknife del sesgo de*** $\boldsymbol{T_n}$ *como:*

$$J^b = (n-1)(\overline{T_n} - T_n)$$

Una vez obtenido este estimador, es posible dar con el *estimador Jackniffe de* θ. *[2]*

Definición 4.0: *Sea* $\mathcal{X}$ *una muestra aleatoria y* T_n *un estadístico. Se define* ***el estimador Jackknife de*** θ *como:*

$$J^\theta = nT_n - (n-1)\overline{T_n} \quad [2]$$

Para terminar, se hará la mención del resultado correspondiente a la estimación de la varianza de T_n para este método de remuestreo.

Teorema 1.0: *Sea $\mathcal{X}$ una muestra aleatoria y T_n un estadístico. El* ***estimador Jackknife de la varianza de*** *T_n esta dado por:*

$$v_J = \frac{n-1}{n}\sum_{i=1}^{n}\left(T_{n-1,i} - \frac{1}{n}\sum_{j=1}^{n}T_{n-1,j}\right)^2$$

Como se ha podido observar, el uso de esta técnica provee recursos muy simples para estimar el sesgo y la varianza del estimador de un parámetro. En vista de ello, definido **los estimadores de Jackknife se presentará el siguiente caso de uso,** para ilustrar de una mejor manera este procedimiento.

Enunciado: *Dados los datos de ocupación de camas UCI proporcionados por la Secretaría Distrital de Salud en el crudo "UCI.csv", encontrar una estimación insesgada puntual y de intervalo del parámetro de la media muestral.*

Solución: La solución de este ejercicio puede encontrarla en la referencia [6] .

Bootstrap

El método **Bootstrap** o método computacionalmente intensivo, es un método de estimación de remuestreo con reemplazo, mediante el cual se busca obtener la varianza de un estimador y/o alguna otra característica referente, y de ser necesario la distribución o la aproximación de la distribución que posee el estimador.

El método fue introducido en 1979 por *Bradley Efron* en su artículo *"Bootstrap methods: another look at the jackknife"* en la revista *The Annals of Statistics, Vol. 7, No. 1* inspirado en trabajos anteriores sobre el método *Jackknife* [3]. Su nombre hace referencia a la frase *to pull oneself up by one's bootstrap (Levantarse por uno mismo)* del libro *las aventuras del barón Munchausen* de Rudolph E [5]. Raspe, aludiendo a una de las aventuras del barón, el cual al caer a un lago escapa *tirando* de los cordones de sus propias botas, esto en referencia a como el método hace uso de sus datos para generar más datos. [2]

Este método coincide con el método *Jackknife* en cuanto al uso del remuestreo de datos para la simplificación y reemplazo de cálculos complejos teóricos, con la diferencia de que el *Bootstrap* parte de muestras aleatorias con reemplazo generadas a partir de una muestra original de la población.

Dada $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria de una variable aleatoria unidimensional X con distribución F desconocida y $T_n = T_n(X_1, X_2, X_3, \ldots, X_n)$ una estadística, es evidente que al no conocer F, no es posible calcular el estadístico deseado. El método Bootstrap plantea aproximar el cálculo de los estadísticos mediante el uso de una es-

timación de F. Según la estimación de F que se asuma, se presentará alguno de los siguientes casos de Bootstrap:

- **Bootstrap no paramétrico:** También conocido simplemente como Bootstrap, se basa en el remuestreo de muestras aleatorias a partir de una muestra original haciendo uso de la distribución empírica.

  Se emplea cuando no se dispone de ningún comportamiento o característica de la muestra aleatoria. Consiste en obtener B muestras Bootstrap, la cuales son conjunto de datos de tamaño n armados a partir de la muestra original y con su mismo tamaño, con repetición de datos, y cada uno con probabilidad $\frac{1}{n}$.

  Al realizar este procedimiento se garantiza la independencia, ya que cada observación está idénticamente distribuida y es independiente de las otras, de tal forma que se tiene B muestras Bootstrap independientes $x^*_{b,1}, x^*_{b,2}, x^*_{b,3}, \ldots, x^*_{b,n} \sim \hat{F}$ para $b = 1, \ldots, B$.

- **Bootstrap paramétrico:** Este caso es se da cuando se conoce la familia de distribución de F pero no todos sus parámetros, por lo cual se busca estimar a partir de los datos de la muestra.

  Se diferencia del proceso anterior en el sentido de que no se realiza el muestreo con reemplazo, en su lugar las muestras son obtenidas a partir del estimador paramétrico de la función $\hat{F}(\xi) = F(\hat{\xi})$. De tal forma que se tiene B muestras Bootstrap independientes $x^*_{b,1}, x^*_{b,2}, x^*_{b,3}, \ldots, x^*_{b,n} \sim F(\hat{\xi})$ para $b = 1, \ldots, B$.

El caso no paramétrico suele darse con mucha más frecuencia que el parametrico, por lo cual se trabaja más con casos parametricos. En caso de conocer la distribución F se hace uso del método de *MonteCarlo*.

Ahora se abordaran algunas de sus aplicaciones más comunes.

Estimación de la Distribución de un Estadístico

Definición 5.0: *Sea $\boldsymbol{X}$ una muestra aleatoria simple (X_1, X_2, X_3, …, X_n), se le conoce como muestra Bootstrap asociada a una función de distribución F, a la muestra aleatoria simple $\boldsymbol{X}^* = X_1^*, X_2^*, X_3^*, \ldots, X_n^*$ de la distribución empírica F_n asociada a $\boldsymbol{X}$.* [2]

Definición 6.0: *Sea X una variable aleatoria con distribución desconocida F, $\boldsymbol{X}$ una muestra aleatoria simple de X y $T_n = T_n(X_1, X_2, X_3, \ldots, X_n)$ un estadístico. La distribución Bootstrap ideal de T_n es:*

$$H_{BOOT}(x) \quad = \quad P_*(T_n(X_1^*, \ldots, X_n^*) \leq x | \mathbf{X})$$

donde $\boldsymbol{X}^* = X_1^*, \ldots, X_n$ es una muestra de bootstrap y P_* es la probabilidad condicionada dado $\boldsymbol{X}$. [2]

Note que esta probabilidad se corresponde con los n^n remuestreos de la muestra original $\mathbf{X}$, pero solo es válida para tamaños de muestra pequeños. Por lo que para muestras grandes se opta por realizar B veces el remuestreo, es decir, estimar mediante el método de **MonteCarlo.**

Definición 7.0: *Sea X una variable aleatoria con distribución desconocida F,* $\boldsymbol{X}$ *una muestra aleatoria simple de X y* T_n *un estadístico. Sea* B *un número fijo. La distribución Bootstrap de MonteCarlo es:*

$$H_{BOOT}^{(B)}(x) = \frac{1}{B}\sum_{i=1}^{B} I_{(-\infty,x)}(T_n(X_{i,1}^*, \ldots, X_{i,n}^*))$$

En donde $X_{i,1}^*$,...,$X_{i,n}^*$ *con i=1,...,B son las muestras Bootstrap asociadas a* $\boldsymbol{X}$. [2]

Estimación de la Varianza de un Estadístico

Definición 8.0: *Sea X una variable aleatoria con distribución desconocida F, sea* $\boldsymbol{X}$ *una muestra aleatoria simple de X y* T_n *un estadístico. El estimador ideal Bootstrap para la varianza de* T_n *es:*

$$\begin{aligned} Var(T_n) &= \int_{\mathbf{X}\in Sop(\mathbf{X})} (T_n(\mathbf{X}) - \int_{\mathbf{X}\in Sop(\mathbf{X})} T_n(\mathbf{X}) \prod_{i=1}^{n} dF_n(x_i)^2 \prod_{i=1}^{n} dF_n(x_i) \\ &= Var_*[T_n(X_1^*, \ldots, X_n)|\mathbf{X}] \end{aligned}$$

Donde $\boldsymbol{X}$ *es una muestra Bootstrap.* [2]

Definición 9.0: *Sea X una variable aleatoria con distribución desconocida F, sea* $\boldsymbol{X}$ *una muestra aleatoria simple de X y* T_n *un estadístico. Sea B un entero positivo,* $X_{i,1}^*$,...,$X_{i,n}^*$ *con i= 1,..., B muestras Bootstrap y* T_i^* *el valor del estadístico de cada muestra. El estimador Bootstrap MonteCarlo para la varianza de* T_n *es:*

$$v_{BOOT}^{(B)}(x) = \frac{1}{B}\sum_{b=1}^{B}\left(T_b^* - \frac{1}{B}\sum_{l=1}^{B} T_l^*\right)^2 \quad [2]$$

Intervalos de Confianza

El método Bootstrap permite establecer intervalos de confianza que puedan incluir los parámetros estimados con una alta probabilidad. De los intervalos de confianza los más usados son:

- ***(Bootstrap-t)*** Tiene como idea fundamental usar un pivote de la forma $T_n = \dfrac{\hat{\theta}_n - \theta}{\hat{\sigma}_n}$. [2]

  Definición 9.0:*Sea X una variable aleatoria con distribución desconocida F,* $\boldsymbol{X}$ *una muestra aleatoria simple de* X *y sea* θ *un parámetro de interés de F. Si se posee una muestra Bootstrap y* $T_n = \dfrac{\hat{\theta}_n - \theta}{\hat{\sigma}_n}$*, el cual es función pivote de* θ *para una distribución desconocida* G_n*, en donde* $\hat{\theta}_n$ *es el estimador de* θ *y* $\hat{\sigma}_n^2$ *la varianza de* $\hat{\theta}_n$*.*

  Sea $G_{BOOT} = P_*(T_n \leq x | X_1, \ldots, X_n)$ *la distribución Bootstrap de* T_n*. Se define el intervalo de confianza Bootstrap-t como* (I_{BT}, S_{BT})*, el cual*

$$I_{BT} = \hat{\theta}_n - \hat{\sigma}_n G^{-1}_{BOOT}(1-\alpha)$$
$$S_{BT} = \hat{\theta}_n - \hat{\sigma}_n G^{-1}_{BOOT}(\alpha)$$

- ***(Bootstrap percentil)*** En este caso se asume la siguiente distribución:

$$K_{BOOT}(x) = P_*(\hat{\theta}^*_n \leq x)$$

  Siendo $\hat{\theta}^*_n = \hat{\theta}_n(X^*_1, \ldots, X^*_n)$, y $\hat{\theta}_n$ el estimador de θ. Una vez conseguida la distribución, se toma percentiles en α y en $1-\alpha$ de la distribución como intervalo de confianza. [2]

  Definición 10.0:*Sea X una variable aleatoria con distribución desconocida F, sea* $\boldsymbol{X}$ *una muestra aleatoria simple de X y sea* θ *un parámetro de interés de F. Si se posee una muestra Bootstrap y* $\hat{\theta}_n$ *es un estimador de* θ*. Se define el intervalo Bootstrap percentil como* (I_{BP}, S_{BP})*, el cual:*

$$\begin{aligned} I_{BP} &= K^{-1}_{BOOT}(\alpha) \\ S_{BP} &= K^{-1}_{BOOT}(1-\alpha) \end{aligned}$$

- ***(Bootstrap hıbrido)*** En este caso se asume la distribución como:

$$H_{BOOT}(x) = P_*(n^l(\hat{\theta}^*_n - \hat{\theta}_n) \leq x)$$

En donde l es una constante fija *(por lo general $\frac{1}{2}$)*. [2]

Definición 11.0:*Sea X una variable aleatoria con distribución desconocida F, sea* $\boldsymbol{X}$ *una muestra aleatoria simple de X y sea θ un parámetro de interés de F. Se posee una muestra Bootstrap y sea H_{BOOT} la distribución Bootstrap de $n^l(\hat{\theta}_n - \theta)$ basada en la muestra Bootstrap, en el cual $\hat{\theta}_n$ es un estimador de θ. El intervalos de confianza híbrido (I_{HB}, S_{HB}) se define,*

$$\begin{aligned} I_{HB} &= \hat{\theta}_n - n^{-l} H^{-1}_{BOOT}(1-\alpha) \\ S_{HB} &= \hat{\theta}_n - n^{-l} H^{-1}_{BOOT}(\alpha) \end{aligned}$$

En este caso, se usa $n^l(\hat{\theta}^*_n - \hat{\theta}_n)$ como función pivote (Bootstrap-t) y se obtienen los percentiles de la función H_{BOOT} (Bootstrap percentil), de ahí que a este método se le conozca como híbrido.

Por último, se expone un ejemplo simple del método.

Ejemplo (Ver referencias [4] y [5]): *Se realizó un estudio sobre si pequeñas dosis de aspirina puede prevenir ataques cardiacos en personas de mediana edad. Los resultados se resumieron en la siguiente tabla:*

	Ataques	NO Ataques
Aspirina	104	11037
Placebo	189	11034

Según el informe, los pacientes que tomaron aspirina tenían cerca de la mitad de riesgo de sufrir ataques. ¿Es verdad?

En primera instancia observe que los datos son muy pocos para poder concluir, por lo que se hace uso del método BootStrap. Para comprobar la afirmación del estudio se hará uso de la razón de odds *(razón de momios (RM), razón de oportunidades o razón de probabilidades)* para comprobar el resultado *(Aspirina sobre Placebo).* Se utiliza el siguiente código en R:

```
n1=11037+104 #Total Aspirina
s1=104 #Exitos Aspirina
n2=11034 + 189 #Total Placebo
s2=189 #Exitos Placebo
p1pre=c(rep(1,s1),rep(0,n1-s1))
p2pre=c(rep(1,s2),rep(0,n2-s2))
p1=sample(p1pre,n1) #Aspirina
p2=sample(p2pre,n2) #Placebo
```

```
n.bs=4000 #Numero muestras bootstrap
#Reservo dos vectores de ceros
bs1=rep(0,n.bs)
for(i in 1:n.bs){
#Proporcion de exitos en muestras bootstrap Aspirina
y Placebo
bs1[i]=sum(sample(p1,n1,replace=TRUE))/n1
bs2[i]=sum(sample(p2,n2,replace=TRUE))/n2
}
#Replicas de la estimacion bootstrap del ratio
ratio=bs1/bs2
#Histograma de las estimaciones del ratio
lattice::histogram(ratio)
mean(ratio)
median(ratio)
#IC bootstrap son los cuantiles del 0 .025 y
#0 .975 de la muestra ordenada
quantile(ratio,c(0.025,0.975))
```

Se obtuvo una media de 0.5577, una mediana de 0.5538 y un intervalo de confianza 0.4352 a 2.5% y 0.7018 a 97.5%, lo que confirma la conclusión de estudio.

References

[1] Nisbet R. (2018) Ken Yale D.D.S., J.D., In Handbook of Statistical Analysis and Data Mining Applications (Second Edition).

[2] Jordá Muñoz, R. (2019). METODOS DE REMUESTREO: JACKKNIFE Y BOOTSTRAP. https://www.um.es/documents/118351/14493552/Jorda+Muñoz+TF_48742912+%28publicar%29.pdf/1bfd-4617-a28d-9c9709a17582.

[3] Belio Miranda, J. (2020), Métodos Bootstrap y sus aplicaciones.

https://zaguan.unizar.es/record/98153/files/TAZ-TFG-2020-1954.pdf

[4] Harold M. Schmeck Jr. (1988), HEART ATTACK RISK FOUND TO BE CUT BY TAKING ASPIRIN.

https://www.nytimes.com/1988/01/27/us/heart-attack-risk-found-to-be-cut-by-taking-aspirin.html

[5] J.M. Marín (2018), Métodos de Remuestreo: Tema 1. Introducción a los Métodos de Remuestreo, Universidad Carlos III de Madrid

http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Boots/tema1BooPres.pdf

[6] Garavito. S (2022) Jackknife Notebook.