Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Cómputo Estadístico
Tarea 3

Marcelo Alberto Sanchez Zaragoza

3 de noviembre de 2021

1. Problema 1

a) Muestra 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 original?

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

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

Solución

Inciso a)

Para demostrar lo solicitado consideremos el caso de hacer k extracciones de una urna de n objetos con las condiciones que cada objeto extraído es regresado a la urna (y entonces puede ser elegido nuevamente), y en donde el orden de la muestra no es relevante. Para encontrar una fórmula para el total de muestras que pueden obtenerse con estas características, usaremos una modelación distinta pero equivalente. Consideremos el arreglo de n casillas, supongamos que la primera casilla tiene dos cruces, y eso indica que la bola uno fue seleccionada dos veces, la segunda casilla está vacía, y ello significa que la bola dos no fue seleccionada, etc.
El número de cruces en la casilla i indica entonces el número de veces que la bola i fue seleccionada. En total debe haber k cruces pues es el total de extracciones. Deseamos entonces conocer el número de posibles arreglos que pueden obtenerse con estas características, y debe ser claro, después de algunos momentos de reflexión, que éste es el número de muestras de tamaño k, con reemplazo y sin orden, que se pueden obtener de un conjunto de n elementos distinguibles. Consideremos que las dos paredes en los extremos de este arreglo son fijas, estas paredes se encuentran ligeramente remarcadas. Consideremos además, que las posiciones intermedias, cruz o línea vertical, pueden moverse. En total hay $n + k$,1 objetos movibles y cambiar de posición estos objetos produce las distintas configuraciones posibles que nos interesan.

El número total de estos arreglos es: $\binom{n+k-1}{k}$

Pero para nuestro caso k es igual a n por lo que finalmente tenemos: $\binom{2n-1}{n}$

Inciso b)

La probabilidad de que una muestra de bootstrap sea idéntica a la original es igual a tomar una de las distintas selecciones del conjunto original pero con el detalle que no podemos repetir algún x_i, por lo que vamos a tener: $(n)(n-1)...(2)(1) = n!$. Lo que resta es dividir por el total de combinaciones que es igual n^n. Así:

$$P = \frac{n!}{n^n}$$

Inciso c)

El conjunto que es más probable es la que contenga todos los elementos de la muestra original ya que cualquier otra muestra nos limitaria las posibles opciones que tenemos de elegir un conjunto.

Inciso d)

Ya que nos piden obtener la cantidad promedio de veces que salga X_i en una muestra de bootstrap se entiende que basados en como se planteo las distintas muestras del inciso a) entonces se entiende que hay que tener fijo un lugar con al menos una casilla con una bola por lo que realmente vamos a tener:

$$\frac{\binom{2n-2}{n-1}}{\binom{2n-1}{n-1}} = \frac{n}{2n-1}$$

2. Problema 2

Sea $x_1, x_2, ..., x_n$ una muestra aleatoria. Entonces $\hat{\mu} = \bar{x}$ es un estimador del parámetro desconocido μ. Considere el estimador dejando uno fuera de una muestra de Jackknife. Sea $\widetilde{\mu}_i$ el promedio de los restantes (n-1) observaciones cuando x_i es excluido:

a) Muestra que $x_i = n\hat{\mu} - (n-1)\widetilde{\mu}_i$

b) Ahora supongamos que necesitamos estimar un parámetro θ y seleccionamos que sea un estimador de la muestra. Imitando el procedimiento de Jackknife para estimar μ notemos que $\theta_i^* = n\hat{\theta} - (n-1)\widetilde{\mu}_i$. ¿Cuál es el estimador de Jackknife de θ? ¿Cuál es el estimador Jackknife de sesgo estimate de $\hat{\theta}$ y su varianza?

Solución

Inciso a)

Para desmotrar podemos partir de lo siguiente:

$$x_i = x_i$$

donde vamos a ir sumando y restando los valores de x_j, tal que se mantenga la igualdad, observe que $j = 1, 2, ..., n$ con $j \neq i$, así:

$$x_i = x_i$$
$$x_i = x_1 - x_1 + x_2 - x_2 + x_3 - x_3 + \ldots + x_{i-1} - x_{i-1} + x_i + x_{i+1} - x_{i+1} + \ldots + x_n - x_n$$

Ahora vamos a agrupar los valores en dos grupos, uno que tenga todos aquellos valores negativos y otro donde sean positivos. Así:

$$x_i = (x_1 + x_2 + \ldots + x_{i-1} + x_i + x_{i+1} + \ldots + x_n) - (x_1 + x_2 + \ldots + x_{i-1} + x_{i+1} + \ldots + x_n)$$

Observe que el primer grupo de valores contiene un total de n elementos mientras que el segundo contiene $n-1$ elementos, ahora vamos a multiplicar por 1 ambas expresiones, es decir, por un valor que no afecte dicha expresión:

$$
\begin{aligned}
x_i = \frac{n}{n}\left(\sum_{j=1}^{n} x_j\right) - \frac{n-1}{n-1}\left(\sum_{j\neq i}^{n} x_j\right) \\
= n\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right) - (n-1)\left(\frac{1}{n-1}\sum_{j\neq i}^{n} x_j\right)
\end{aligned}
$$

así finalmente tenemos:

$$x_i = n\hat{\mu} - (n-1)\widetilde{\mu}_i$$

Inciso b)

Ya que nos mencionan que se desea estimar el parámetro θ y se selecciona a $\hat{\theta}$ como el estimador e imitando el procedimiento del iniciso anterior tenemos que $\theta_i^* = n\hat{\theta} - (n-1)\widetilde{\mu}_i$, el estimador Jackknife es el promedio de estos pseudovalores θ_i^*, así nuestra expresión es la siguiente:

$$
\begin{aligned}
\theta_J = \frac{1}{n}\sum_{i=1}^{n} \theta_i^* \\
\theta_{(*)} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\theta}_i \\
\theta_J = n\hat{\theta} - (n-1)\theta_{(*)}
\end{aligned}
$$

Ahora el estimador Jackknife de la desviación estándar de $\hat{\theta}$ se define como:

$$b_{Jack} = (n-1)(\theta_{(*)} - \hat{\theta})$$

Ahora el estimador Jackknife de la varianza de $\hat{\theta}$ se define como:

$$V_{Jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\widetilde{\theta}_i - \theta_{(*)})^2$$

3. Problema 3

Sea $x_1, x_2, ..., x_n$ una muestra aleatoria de una normal $N(\theta, 1)$ y suponga que $\bar{x}$ es un estimador de θ.

Sean $X_1^*, X_2^*, ..., X_n^*$ una muestra bootstrap de $N(\bar{x}, 1)$. Muestre que $\bar{X} - \theta$ y $\bar{X}^* - \bar{x}^*$ tienen la misma distribución $N(0, 1/n)$.

Solución

Observemos que $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{1}{n} x_i$ y $\bar{X}^* = \frac{1}{n}\sum_{i=1}^{n} x_i^* = \sum_{i=1}^{n} \frac{1}{n} x_i^*$, recordando que son distribuciones normales tenemos que $\bar{X} \sim N(\theta, \frac{1}{n})$ y $\bar{X}^* \sim N(\bar{X}, \frac{1}{n})$, así tenemos para el primer caso:

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

para el segundo caso:

$$\bar{X}^* - \bar{X} \sim N(\bar{X}, \frac{1}{n}) - N(\theta, \frac{1}{n})$$

Observe que al hacer la resta la varianza se hace cero y solo nos queda como distribución $\bar{X} - \theta$ y ya conocemos su distribución $\bar{X} - \theta \sim N(0, \frac{1}{n})$.

Por lo que finalmente vamos a tener que:

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

4. Problema 4

Considere el conjunto de datos 2,5,3,9. Sea $x_1^*, x_2^*, ..., x_n^*$ 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 boostrap sea igual a 9.

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

Solución

Inciso a)
Sabemos que solo hay 35 posibles promedios con esos 4 números y las posibles combinaciones que podemos obtener de realizar selección con remplazo es igual a 4^4. Se puede observar que dado que son muestras bootstrap debemos buscar cuantos casos nos resultan que al realizar el promedio sea igual 2, solo encontramos una por lo que su probabilidad es $\frac{1}{35}$.

Inciso b)
Lo mismo ocurre para el número 9 ya que solo encontramos una forma de tener este promedio por lo que su probabilidad es igual a $\frac{1}{35}$.

Inciso c)
Al buscar para el número 4 encontramos dos posibles opciones por lo que

su probabilidad es igual a $\frac{2}{35}$.

5. Problema 5

Maximice las siguientes 2 funciones utilizando el algoritmo de recocido simulado.

a) Función 1

$$f_a(x,y) = \frac{sin^2(\sqrt{(x+\alpha)^2} + (y+\beta)^2 - 0{,}5)}{[1 + 0{,}001 + ((x+\alpha)^2 + (y+\beta)^2)]^2}$$
$$-100 \leq x \geq 100$$
$$-100 \leq y \geq 100$$

b) Función 2

$$f_a(x,y) = 21{,}5 + x * sin(4 * \pi * x) + sin(20 * \pi y)$$
$$-3 \leq x \geq 12{,}1$$
$$4{,}1 \leq y \geq 5{,}8$$

Solución

Inciso a)

```
h <- function(x,y,a,b){
  (-1)*(sin((x + a)**2 + (y + b)**2  )**2 -
         0.5 ) / (1.0 + 0.001*( (x+a)**2 + (y+b)**2 )  )**2
  }

h00 <- h(0,0,0,0)
```

```r
set.seed(2021)
x0 <- 9
y0 <- 9
a0 <- 11
b0 <- 11

h0 <- h(x0,y0, a0, b0)
ua <- 0.01
n <- 50000

z <- matrix(0,n,6) ; z[1,] <- c(x0,y0,h0,a0, b0,1)
rem <- c()
for( i in 2:n ){
  ti <- 0.1/(log(1+i))
  xt <- runif(1,x0-ua,x0+ua)
  yt <- runif(1,y0-ua,y0+ua)
  at <- runif(1,a0-ua,a0+ua)
  bt <- runif(1,b0-ua,b0+ua)
  if( abs(xt) > 100 | abs(yt) > 100){
    rem <- c(rem,i)
    #print('22')
    next
  }
  ht <- h(xt, yt, at, bt)
  dh <- h0 - ht
  r <- min(exp(dh/ti),1)
  if(runif(1) < r){
    x0 <- xt
    y0 <- yt
    h0 <- ht
    a0 <- at
    b0 <- bt
    z[i,] <- c(xt,yt,ht,at,bt,r)
```

```
    #print('22')
  }else{
    z[i,] <- c(x0,y0,h0,a0,b0,r)
    #print('p')
  }
}

#z <- z[-rem,]
m <- nrow(z)

cmin <- which.min(z[,3])
round(z[cmin,],6)

[1]  9.112218  9.423899 -0.154401 10.438327 11.004298  1.000000
```

Los valores que obtenemos al final son los siguientes: $x = 9.1122$, $y = 9.4238$, $f(x, y, \alpha, \beta) = 0.1544$, $\alpha = 10.4383$ y $\beta = 11.004$.

Cabe recalcar que los resultados de la función se observan con signo negativo, esto es debido a que el problema nos pide maximizar la función.

Inciso b)

```
h <- function(x,y){
  (-1)*( 21.5 + x*sin(4*pi*x) + y*sin(20*pi*y) )
  }

set.seed(2021)
x0 <- 8
y0 <- 4.9

h0 <- h(x0,y0)
ua <- 0.001
```

```
n <- 50000

z <- matrix(0,n,4) ; z[1,] <- c(x0,y0,h0,1)
rem <- c()
for( i in 2:n ){
  ti <- 0.1/(log(1+i))
  xt <- runif(1,x0-ua,x0+ua)
  yt <- runif(1,y0-ua,y0+ua)
  if( xt < -3.0 | xt > 12.1 | yt < 4.1 | yt > 5.8){
    rem <- c(rem,i)
    #print('vava',i)
    next
  }
  ht <- h(xt, yt)
  dh <- h0 - ht
  r <- min(exp(dh/ti),1)
  if(runif(1) < r){
    x0 <- xt
    y0 <- yt
    h0 <- ht
    z[i,] <- c(xt,yt,ht,r)
    #print('22')
  }else{
    z[i,] <- c(x0,y0,h0,r)
    #print('p')
  }
}

#z <- z[-rem,]
m <- nrow(z)

cmin <- which.min(z[,3])
round(z[cmin,],6)
```

```
[1]   8.125768   4.925057 -34.550415   1.000000
```

Los valores que obtenemos al final son los siguientes: $x = 8.1257$, $y = 4.9250$, $f(x, y) = 34.5504$.

Cabe recalcar que los resultados de la función se observan con signo negativo, esto es debido a que el problema nos pide maximizar la función.

6. Problema 6

Con el algoritmo Metropolis-Hastings(MH), simular lo siguiente:

1) Sean $x_i \sim Ga(\alpha, \beta); i = 1, 2, ..., n$. Simular datos x_i con $\alpha = 3$ y $\beta = 100$ considerando los casos $n = 4$ y 30.

   Con $\alpha \sim U(1, 4)$, $\beta \sim exp(1)$ distribuciones a priori, se tiene la posterior

   $$f(\alpha, \beta|\bar{x}) \propto \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} r_1^{\alpha-1} e^{\beta(r_2-1)} I(1 \leq \alpha \geq 4) I(\beta > 1)$$

   con $r_2 = \sum_{i=1}^{n} x_i$ y $r_1 = \prod_{i=1}^{n} x_i$,
   $I = funcion\ indicadora(1\ si\ se\ cumple\ la\ condicion, 0\ si\ no)$
   En ambos casos, grafica los contornos para visualizar donde esta concentrada la posterior. Utilizar la propuesta:

   $$q\left(\begin{pmatrix}\alpha_p \\ \beta_p\end{pmatrix} | \begin{pmatrix}\alpha \\ \beta\end{pmatrix}\right) = \begin{pmatrix}\alpha \\ \beta\end{pmatrix} + \begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix}$$

   donde

   $$\begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix} \sim \left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & 0 \\ 0 & \sigma_2^2\end{pmatrix}\right)$$

Solución

falta este ejercicio