UNIVERSIDAD DE VALPARAISO
CHILE

Trabajo práctico 2

IECD 424: Estadística no Paramétrica

Bastián Barraza Morales

December 12, 2022

---

1 Introducción

Al trabajar con datos de la vida cotidiana, muchas veces no se tiene el valor real de los parámetros, así como la distribución original de la población. En este sentido, el método de re-muestreo Bootstrap permite realizar estimaciones acerca de los estimadores de una muestra.

La idea de este método es tomar múltiples muestras con reemplazo de una misma muestra aleatoria (generalmente el conjunto de datos original). Así, las muestras repetidas del conjunto de datos original, representan lo que obtendríamos si tomamos B muestras de la población de interés.

De esta manera se puede aproximar el sesgo o la varianza de los parámetros de interés y realizar inferencia acerca de la muestra a partir de los datos obtenidos. Una vez realizado este procedimiento, se puede inferir acerca de la población con mayor certeza, dado que se tienen estimaciones de la muestra con errores mas bajos que al inicio.

De este modo, el presente informe tiene como objetivo realizar estimaciones de una muestra aleatoria usando métodos Bootstrap y Jackknife. Específicamente, se busca calcular la desviación estándar y el sesgo de los estadísticos ρ_1 y $T = artanh(\rho_1)$ y donde ρ es la coeficiente de correlación muestral a partir de la muestra 1 que posee los datos de la puntuación media en dos pruebas diferentes para nuevas admisiones de 15 facultades de derecho.

Asímismo, se busca realizar estimaciones a partir de la muestra 2, la cual posee los puntajes de 15 examinados en las secciones de lectura y comprensión auditiva de una prueba de inglés, en donde el número de preguntas es 50. Específicamente, se busca estimar el sesgo y la desviación estándar del coeficiente de correlación de la muestra (ρ) y del coeficiente de regresión (R^2), a partir de métodos Jackknife y Bootstrap.

Finalmente, se comparan las estimaciones de Jackknife y Bootstrap para cada uno de los estimadores propuestos en ambas muestras.

Para este trabajo se utilizará el software R para aplicar el método bootstrap y realizar las tablas y gráficos necesarios.

2 Metodología

Para realizar la investigación, se necesita calcular el sesgo, la desviación estándar y los intérvalos de confianza para los estadísticos propuestos (ρ_1, $arctanh(\rho_1)$, ρ_2, ρ_2^2) mediante los métodos Jackknife y Bootstrap. De esta manera, para realizar las estimaciones por Jackknife desde una muestra aleatoria X, se define el vector:

$$I_j = (n-1)\left(T - T_{(-j)}\right)$$

.
Donde $T = T(x_1, x_2, ..., x_n)$ es el estadístico propuesto y $T_{(-j)} = T\left(x_i | i = (1, .., n) \neq j\right)$ es el estadístico calculado sin la j-ésima observación. De esta manera, se define el sesgo y desviación estándar de T:

$$\hat{Bias}(T)_{Jackknife} = \bar{I} = \frac{1}{n}\sum_{i=1}^{n} I_{(-j)}$$

$$\hat{Var}(T)_{Jackknife} = S_I^2 = \sum_{i=1}^{n} \frac{I_{(-j)} - \bar{I}}{n(n-1)}$$

Asimísmo, el estimador ajustado se define de la siguiente manera: $\widetilde{T} = T + \bar{I}$. Además, el intérvalo de confianza para T está dado por: $\left[T \pm Z_{1-\frac{\alpha}{2}}\hat{S}_{Jackknife}\right]$, donde $\alpha = 0.05$.

Por otro lado, para realizar las estimaciones mediante Bootstrap, se deben seleccionar B muestras Bootstrap independientes (es decir, B 'remuestreados') de los n datos.

Luego, se busca evaluar el estimador T en cada una de las muestras Bootstrap, de este modo se obtienen B estimaciones de T, denotadas por $\hat{T}^*(b) = T(x^{*b})$, donde $b = 1, 2, ..., B$ son las muestras obtenidas por el remuestreo y x^{*b} es el vector con los valores de X de la b-ésima muestra Bootstrap.

En el paso 3, se realiza la estimación del sesgo y el error de estimación, los cuales están definidos por:

$$\hat{Bias}_{Bootstrap}(T) = \hat{T}^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B} \hat{T}^*(b)$$

$$\hat{SD}_{Bootstrap}(T) = \hat{S_B} = \sqrt{\frac{\sum_{b=1}^{B}(\hat{T}^*(b) - \hat{T}^*(\cdot))^2}{B-1}}$$

Finalmente se puede definir el estimador ajustado $\widetilde{T} = T + \hat{T}^*(\cdot)$ y el intérvalo de confianza $\left[T(x) \pm t_{1-\frac{\alpha}{2},n-1}\hat{S}_B\right]$ donde $\alpha = 0.05$.

Por otro lado, las variables de interés y los datos desde los cuales se realizan las estimaciones Jackknife y Bootstrap en la muestra 1 están dados están dados por:

$X =$ Puntuación media en la prueba A.
$Y =$ Puntuación media en la prueba B.

j	X	Y
1	576	3.39
2	635	3.30
3	558	2.81
4	578	3.03
5	666	3.44
6	580	3.07
7	555	3.00
8	661	3.43
9	651	3.36
10	605	3.13
11	653	3.12
12	575	2.74
13	545	2.76
14	572	2.88
15	594	2.96

Luego, las variables de interés y el conjunto de datos de la segunda muestra se definen de la siguiente forma:

$A =$ Puntajes de la sección de comprensión auditiva.
$B =$ Puntajes de la sección de lectura.

j	A	B
1	35	31
2	46	35
3	36	30
4	36	22
5	49	37
6	41	30
7	40	21
8	41	33
9	45	34
10	44	31
11	33	21
12	39	15
13	49	42
14	35	25
15	36	22

Al ingresar estos datos y realizar los cálculos necesarios para obtener las estimaciones Jackknife y Bootstrap en R, se obtienen resultados, los cuales serán presentados en la siguiente sección.

3 Resultados

Muestra 1:

Al calcular el coeficiente de correlación (ρ_1) en la muestra 1 se obtiene: 0.776, mientras que T definido por $T = arctanh(\rho) = 1.0352$.

Luego, se tiene el sesgo del coeficiente de correlación de ρ_1 calculado con el método Jackknife y Bootstrap, el cual es 0.008 y 0.003 respectivamente. Del mismo modo, la desviación estándar mediante Jackknife para ρ_1 es 0.137 mientras que con Bootstrap es de 0123. Finalmente, el intérvalo de confianza con un $\alpha = 0.05$ para ρ_1 mediante Jackknife y Bootstrap son [0.505, 1] y [0.530, 1] respectivamente (Apéndice A.1).

Ahora bien, el estadístico $T = arctanh(\rho_1)$ posee un sesgo de -0.098 con desviación estándar de 0.438 al realizar el método Jackknife. Asimísmo, el intérvalo de confianza con $\alpha = 0.05$ para T está definido por [0.175, 0.956]. Por otro lado, al calcular las métricas mediante Bootstrap (Apéndice A.2) se obtiene un sesgo de 0.102 con su respectiva desviación estándar de 0.370 y un intérvalo de confianza (95%) comprendido entre [0.207, 1]. Cabe destacar que el límite superior de los intérvalos de confianza de ρ_1 están acotados hasta 1, dado que el coeficiente de correlación está definido entre [-1, 1].

Los resultados están resumidos en la siguiente tabla:

Método	Estadístico (θ)	Sesgo	D.E	I.C(θ)
Jackknife	ρ_1	0.008	0.137	[0.505, 1]
Jackknife	$arctanh(\rho_1)$	-0.098	0.438	[0.175, 0.956]
Bootstrap	ρ_1	0.003	0.123	[0.530, 1]
Bootstrap	$arctanh(\rho_1)$	0.102	0.370	[0.207, 1]

Además, en la figura 1 se observan las simulaciones Bootstrap de ρ_1 según las iteraciones. Notamos una mayor cantidad de valores entre 0.6 y 0.9. Este rango es esperado, dado que el coeficiente de correlación de la muestra aleatoria es 0.77 y el sesgo mediante jackknife y bootstrap es muy bajo, por lo que, la mayor cantidad de ρ_1 simulados deberían estar cercano a 0.77.

Por otro lado, en la figura 2 se observa un histograma con los valores simulados del coeficiente de correlación. Se nota una semejanza a la distirbución normal con una leve asimetría negativa. Además, en el gráfico tambien se añade la media (denotada con una recta azul) y el intérvalo de confianza con $\alpha = 0.05$ (denotado con rectas entrecortadas de color azul), los cuales pueden ser métricas de interés.

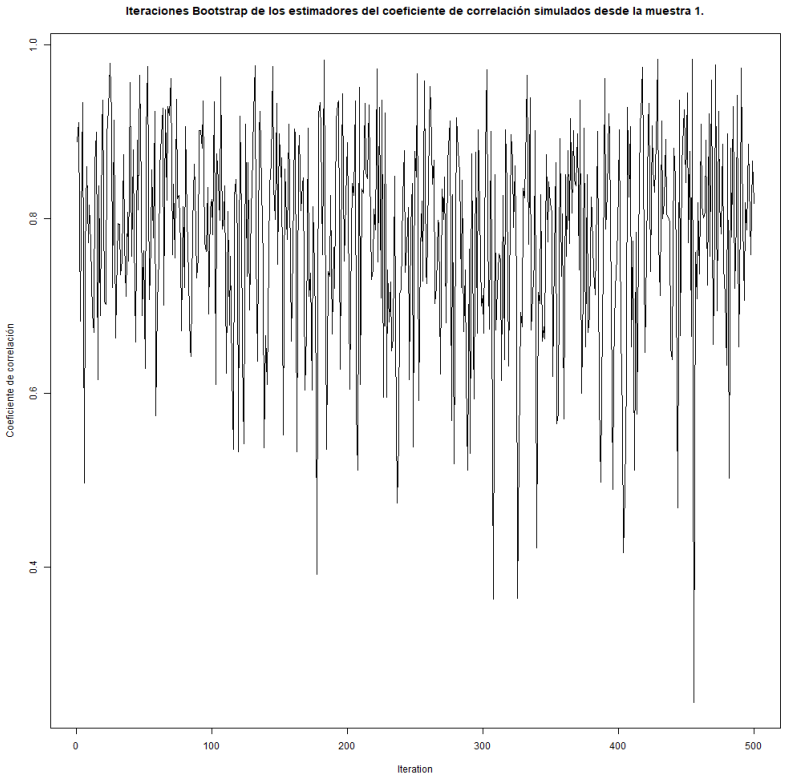

Figure 1: Valores de ρ_1 según su número de iteración.

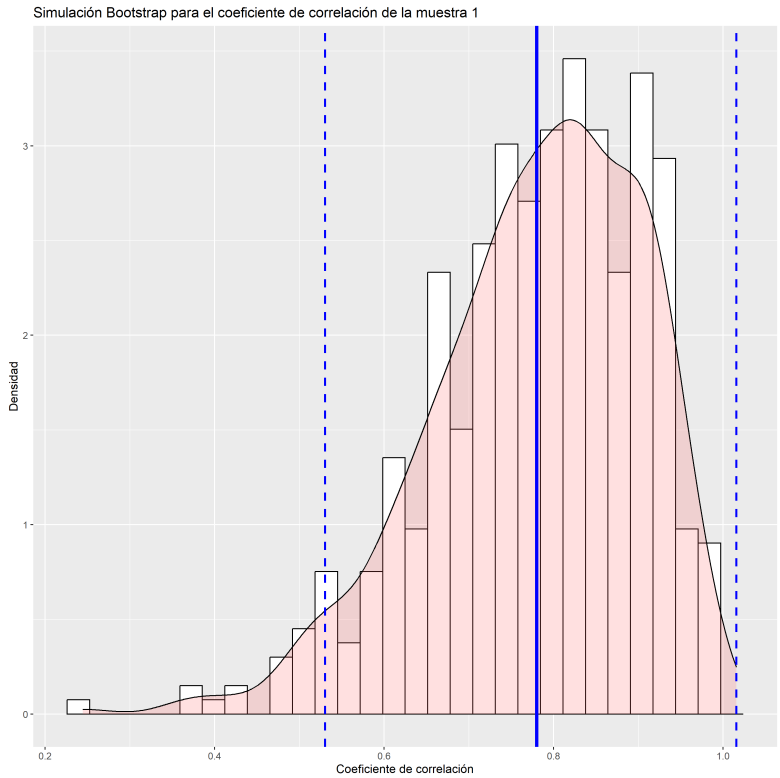

Figure 2: Distribución del coeficiente de correlación de la muestra 1.

Muestra 2:

Al calcular el coeficiente de correlación ρ_2 en la muestra 2 se obtiene un valor de 0.733, mientras que el estadístico $\theta = R^2 = \rho_2^2 = 0.538$

Ahora bien, al calcular el sesgo, la desviación estándar y el intérvalo de confizanza con $\alpha = 0.05$ de ρ_2 mediante el método Jackknife (Apéndice B.1) obtenemos los resultados -0.003, 0.137 y [0.505, 1] respectivamente. Asimismo, con la estimación Boopstrap (Apéndice B.2) se tienen los valores de -0.005, 0.142 y [0.460, 1] para el sesgo, desviación estándar e intérvalo de confianza de ρ_2 respectivamente.

Por otro lado, al estimar el sesgo, la desviación estándar y el intérvalo de confianza para θ mediante Jackknife (Apéndice B.3) obtenemos los resultados -0.014, 0.208 y [0.325, 1] respectivamente. Del mismo modo, mediante Bootstrap (Apéndice B.4) se tienen los valores de 0.012, 0.182 y [0.167, 0.884] para las estimaciones del sesgo, error estándar e intérvalo de confianza respectivamente. Cabe destacar que el límite superior de los intérvalos de confianza de la estimación de ρ_2 y θ están acotados hasta 1 dado que el coeficiente de correlación está definido entre [-1, 1] y el coeficiente de determinación (R^2) se encuentra entre [0, 1].

En la siguiente tabla se resumen los resultados obtenidos:

Método	Estadístico (θ)	Sesgo	D.E	I.C(θ)
Jackknife	ρ_2	-0.003	0.140	[0.457, 1]
Jackknife	ρ_2^2	-0.014	0.208	[0.325, 1]
Bootstrap	ρ_2	-0.005	0.142	[0.460, 1]
Bootstrap	ρ_2^2	0.012	0.182	[0.167, 0.884]

En la figura 3 se observan las simulaciones Bootstrap de ρ^2 según las iteraciones. Se espera que los valores se encuentren cercanos a 0.53, el cual es el valor del coeficiente de regresión en la muestra 2. En nuestro caso, la mayor cantidad de simulaciones de ρ^2 se encuentra entre 0.4 y 0.7.

Además, la figura 4 es un histograma con los valores simulados del coeficiente de regresión en donde se nota una semejanza a una normal centrada cercano a 0.5. Junto a ello se grafica la media (denotada con una recta azul) y el intérvalo de confianza con $\alpha = 0.05$ (denotado con rectas entrecortadas de color azul), los cuales pueden ser métricas de interés.

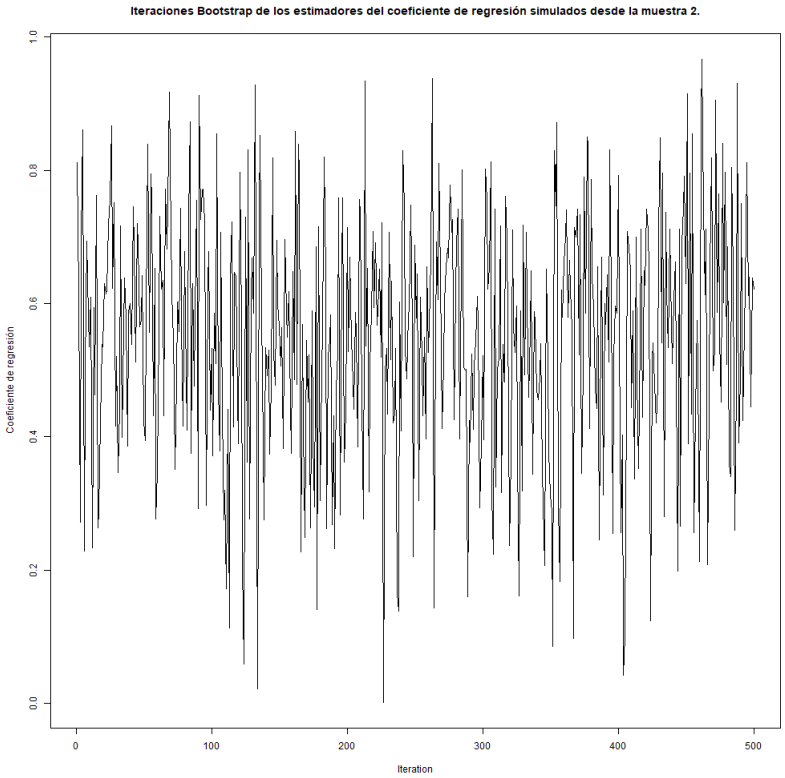

Figure 3: Valores de ρ^2 según su número de iteración.

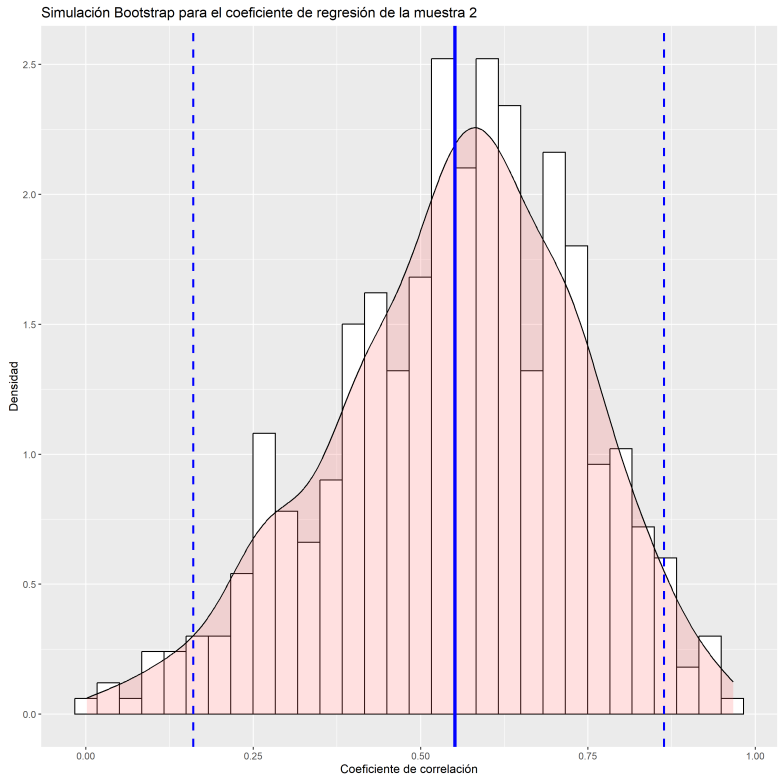

Figure 4: Distribución del coeficiente de regresión de la muestra 2.

4 Conclusión

Los objetivos planteados se cumplen, dado que se realizaron las estimaciones del sesgo, desviación estándar e intérvalo de confianza para los estimadores de cada muestra mediante métodos Jackknife y Bootstrap.

Al realizar la comparaciones acerca de las aplicaciones de los métodos Jackknife y Bootstrap de las estimaciones sobre ρ en la muestra 1 y 2, se obtiene que tanto los sesgos, como la desviación estándar e intérvalos de confianza son muy semejantes, de hecho las diferencias son mínimas. Sin embargo, al comparar los métodos de Jackknife y Bootstrap en las métricas estimadas de $arctanh(\rho)$ en la muestra 1 y ^{2} en la muestra 2, se nota una diferencia mayor que las estimaciones sobre ρ. Esto puede suponer que uno de los métodos no trabaja óptimamente al realizar cálculos sobre la función de un estimador. También se puede plantear la idea de que el tamaño de muestra no es el óptimo para aplicar los métodos de Jackknife y Bootstrap sobre la función de un estimador.

Dado esto, se puede plantear realizar estimaciones con una cantidad mas grande de datos para verificar si existe alguna diferencia significativa a medida que n (total de datos) aumenta.

Por otro lado, vale la pena graficar la distribución de los estimadores simulados, además de realizar tests para corroborar el supuesto de normalidad, y así, obtener intérvalos de confianza más precisos.

5 Apéndice A

Muestra 1

Apéndice A.1

```
# BOOTSTRAP RHO
set.seed(1)
b3 <- boot(data,
   statistic = function(data, i) {
     cor(data[i, "x"], data[i, "y"], method='pearson')
   },
   R = 500
)
boot.ci(b3, type = c("norm")) #bootstrapped CI
```

Apéndice A.2

```
# BOOTSTRAP ATANH(RHO)
set.seed(1)
b2 <- boot(data,
   statistic = function(data, i) {
     atanh(cor(data[i, "x"], data[i, "y"], method='pearson'))
   },
   R = 500
)
b2
boot.ci(b2, type = c("norm", "perc")) #bootstrapped CI.
```

6 Apéndice B

Muestra 2

Apéndice B.1

```
# Jack RHO
jack = numeric(15)
for (i in 1:15){
  jack[i] = cor(a[-i], b[-i])
  jack[i] = 14*(cor(a,b) - jack[i])
}
```

Apéndice B.2

```
# Jack R2
jack = numeric(15)
for (i in 1:15){
  jack[i] = cor(a[-i], b[-i])**2
  jack[i] = 14*(cor(a,b)**2 - jack[i])
}
```

Apéndice B.3

```
# Bootstrap RHO
set.seed(1)
b_m2 <- boot(df,
   statistic = function(df, i) {
     cor(df[i, "a"], df[i, "b"], method='pearson')
   },
  R = 500
)
b_m2
boot.ci(b_m2, type = c("norm", "perc")) #bootstrapped CI.
```

Apéndice B.4

```
#### Bootstrap R2
set.seed(1)
b_m2R2 <- boot(df,
   statistic = function(df, i) {
     cor(df[i, "a"], df[i, "b"], method='pearson')**2
   },
  R = 500
)
b_m2R2
boot.ci(b_m2R2, type = c("norm", "perc")) #bootstrapped CI.
```