UNIVERSIDAD NACIONAL DE COLOMBIA SEDE BOGOTÁ

UNIVERSIDAD NACIONAL DE COLOMBIA SEDE BOGOTÁ

FACULTAD DE CIENCIAS

Proyecto Grupo 6

Presentan:

Juan Pablo Montaño Díaz
Jhon Ayala
José Valdés
David Santiago Garzón Monje.

Profesor

Mario Arrieta

Asignatura

Inferencia Estadística

9 de febrero de 2022

A.Distribución de un estimador máximo verosímil cuando no hay condiciones de regularidad

Considere una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución $U[0, \theta = \kappa]$, es decir, donde el verdadero valor del parámetro es igual al número del grupo ($\kappa = 6$). Se sabe que $\hat{\theta}_{MLE} = X^{(n)}$,sin embargo, como no se tienen condiciones de regularidad, este estimador no tiene una distribución asintótica normal (cuando es debidamente normalizado). Vamos a estudiar también la convergencia en distribución de la variable aleatoria $R_n = n(\theta - X^{(n)})$.

a. Genere $m = 1000$ simulaciones de una muestra de tamaño $n = 10$ de la distribución $U[0, \theta = \kappa]$, usando la función `runif` de R o su equivalente en otro lenguaje de programación. Para cada una de las m muestras, calcule y almacene la estimación máximo-verosímil del parámetro. Haga un histograma de las m estimaciones y superponga sobre este histograma la función de densidad teórica de $X^{(n)}$.

b. Repita el procedimiento descrito en **a.**, variando los tamaños de muestra por los valores $n = 50, 100, 200, 500, 1000$.

c. ¿Qué observa a medida que el tamaño de muestra aumenta ?¿ En qué valor parece concentrarse las realizaciones del estimador máximo-verosímil a medida que el tamaño de muestra aumenta?¿Qué resultado de los vistos en clase explica este fenómeno?

d. Genere nuevamente m muestras de tamaños $10, 50, 100, 200, 500, 1000$; y calcule y almacene las m realizaciones de R_n para cada valor de n. Haga histogramas de los m valores obtenidos para cada n. ¿Observa que, a medida que n aumenta, dicho histograma se asemeja a alguna distribución conocida?

e. Pruebe formalmente su conjetura del apartado anterior, calculando el límite de la distribución cuando $n \longrightarrow \infty$:

$$R_n = n(\theta - X^{(n)}) \xrightarrow{d} ?$$

Nota: El estimador ML de este modelo es súper consistente y por eso su factor de escala no es $\sqrt{n}$ si no n.

Solución A

a.

Junto a este documento se adjuntan los códigos de R y Python utilizados en este proyecto. Sabemos que en una distribución uniforme, el estimador máximo verosímil es $\theta_{MLE} = X^{(n)}$. Así, utilizamos la función `max` sobre cada muestra aleatoria para hallar $X^{(n)}$ y lo almacenamos en cada simulación.
La función de densidad teórica de $X^{(n)}$ está dada por

$$F_{X^{(n)}}(x) := [F_X(x)]^n$$

En este caso, sabemos que la función de distribución de X es: 0, si $x < 0$; x/θ si $0 \leq x \leq \theta$ y 1 si $x > \theta$. Por lo que nuestra función de distribución para $X^{(n)}$ es:

$$F_{X^{(n)}} = \begin{cases} 0 & x < 0 \\ \left(\frac{x}{\theta}\right)^n & 0 \leq x \leq \theta \\ 1 & \theta < x \end{cases}$$

Finalmente, derivamos nuestra función de distribución para así obtener la función de densidad teórica de $X^{(n)}$.

$$f_{X^{(n)}}(x) = \frac{nx^{n-1}}{\theta^n} I_{[0,\theta]}(x)$$

Ahora, comparamos los MLE obtenidos en nuestras simulaciones con muestra aleatoria de tamaño $n = 10$ y la función de densidad de $X^{(n)}$.

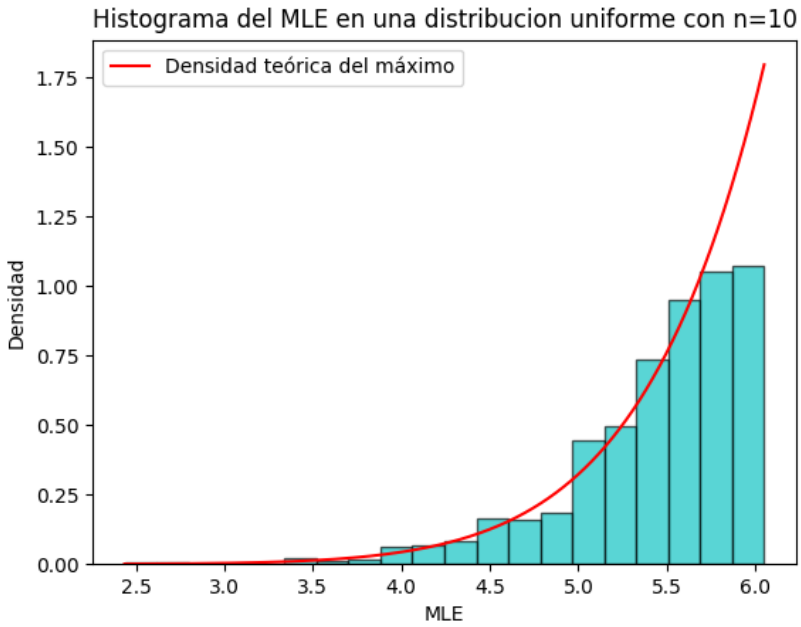
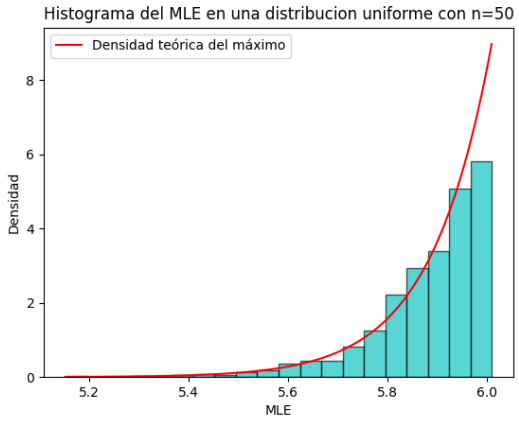


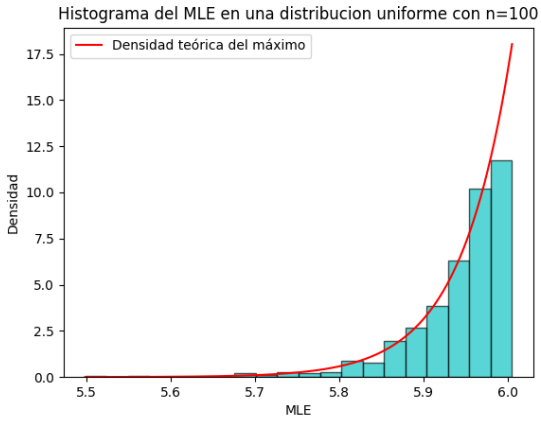
Figura 1: Histograma de los MLE obtenidos en una muestra aleatoria de tamaño 10.

b.

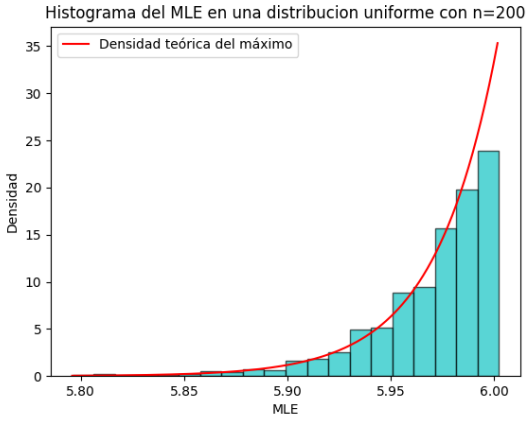
Ahora observemos el comportamiento asintótico de nuestro estimador MLE, para esto, incrementemos el tamaño de nuestra muestra aleatoria y observemos en qué valor se concentra $X^{(n)}$



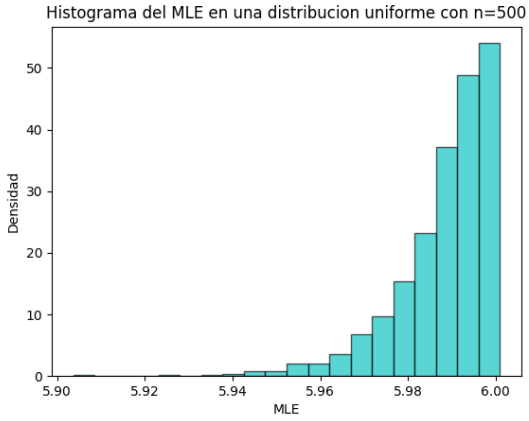
(a) n=50



(b) n=100



(c) n=200



(d) n=500

Figura 2: Histogramas de $X^{(n)}$ para tamaños de muestra $n = 50, 100, 200, 500$.

Notemos que el eje horizontal del histograma cada vez se concentra en un intervalo más pequeño. Si observamos el histograma en el intervalo $[0, \theta = 6]$, obtenemos la siguiente representación.

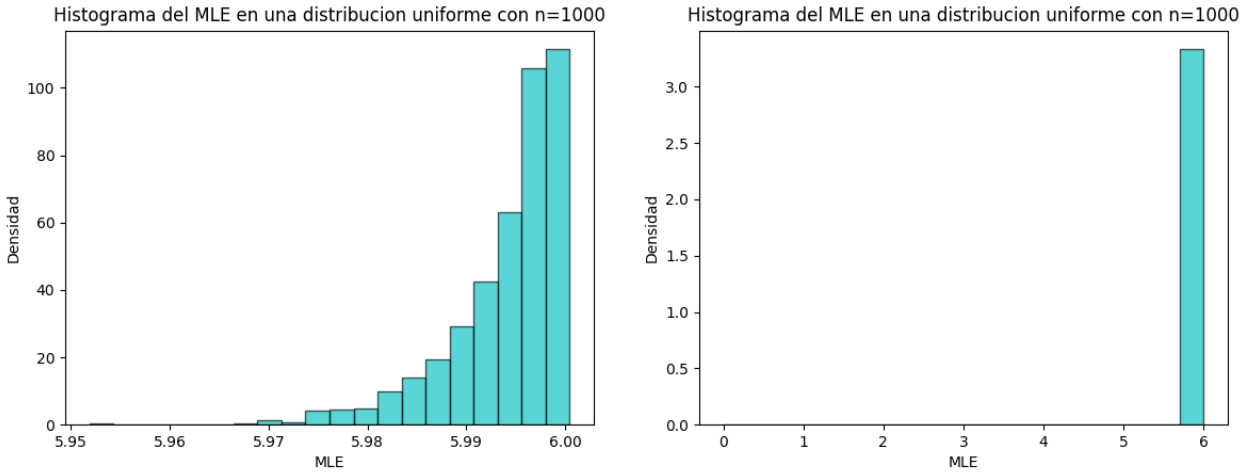


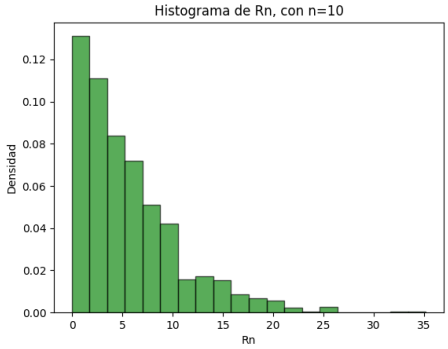
Figura 3: Histogramas de $X^{(n)}$ para $n = 1000$.

c.

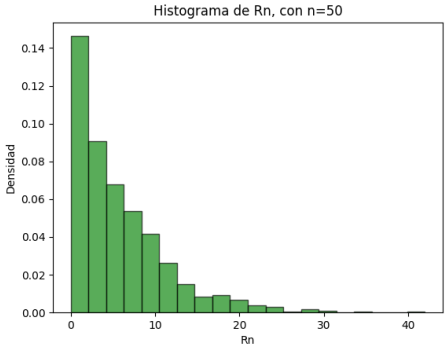
Como podemos observar en los histogramas anteriores, a medida que el tamaño de muestra aumenta, el valor de $X^{(n)}$ suele concentrarse cada vez más en $\theta = 6$.
La razón por la que esto ocurre es por la distribución asintótica de la estadística de orden $X^{(n)}$, pues al tenerse que $n/n \xrightarrow{n\to\infty} 1$, se tiene que $X^{(n)} \xrightarrow{c.s} \xi_1$, donde ξ_1 es el percentil asociado a la probabilidad $p = 1$, dicho percentil es $\xi_1 = \theta$.

d

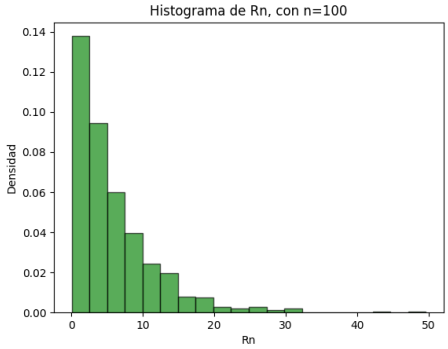
Consideremos ahora la variable aleatoria definida por $R_n := n(\theta - X^{(n)})$. Durante las simulaciones realizadas anteriormente, se obtuvieron los siguientes histogramas de R_n con $n = 10, 50, 100, 200, 500$ y 1000.



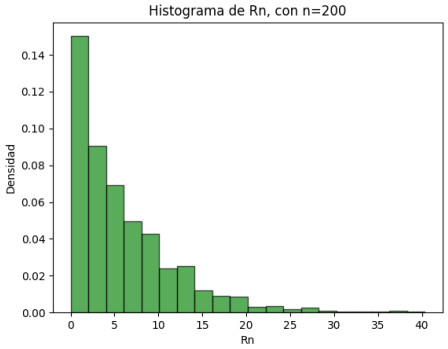
(a) Histograma de Rn con $n = 10$.



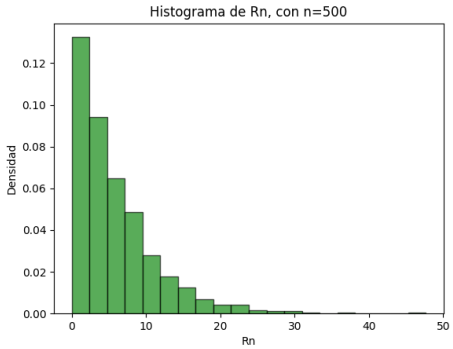
(b) Histograma de Rn con $n = 50$.



(c) Histograma de Rn con $n = 100$.



(d) Histograma de Rn con $n = 200$.



(e) Histograma de Rn con $n = 500$.

Podemos observar que la densidad de R_n se asemeja a la densidad de una distribución exponencial con parámetro $\frac{1}{6} = \frac{1}{\theta}$. Esto se puede observar con mayor claridad al sobreponer la función de densidad exponencial $f_R(x) = \frac{1}{6}e^{-\frac{x}{6}}I_{[0,6]}(x)$, sobre nuestro histograma con $n = 1000$.

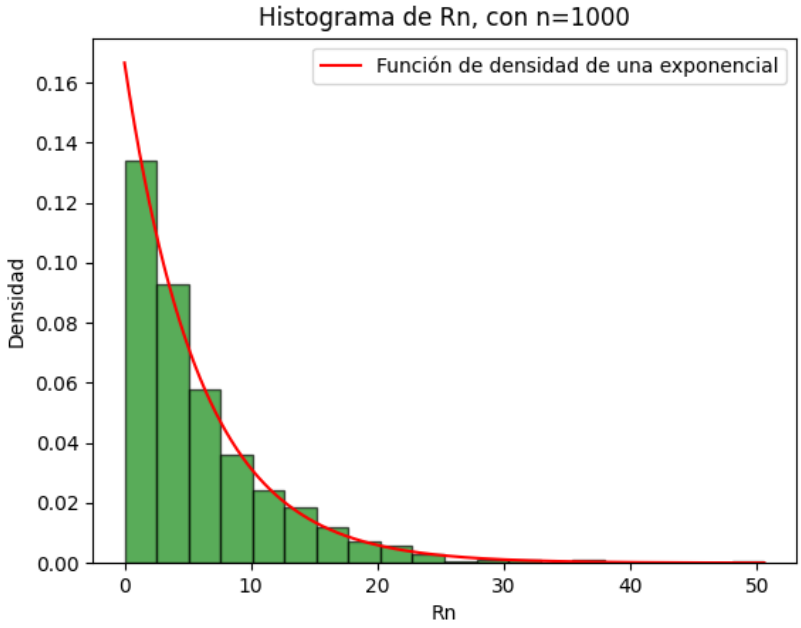


Figura 5: Histograma de Rn con $n = 1000$.

e.

Demostremos nuestra conjetura sobre la convergencia de R_n. Sea $R \sim exp(\frac{1}{\theta})$. Notemos que

$$R_n = h(X^{(n)}) = n(\theta - X^{(n)})$$

Es una función de $X^{(n)}$ estrictamente decreciente para $0 \leq X^{(n)} \leq \theta$, por el teorema de transformación, la función de densidad de R_n es:

$$f_{R_n}(r) := \begin{cases} f_{X^{(n)}}(h^{-1}(r)) \left| \dfrac{d}{dr} h^{-1}(r) \right| & \text{Si } r = h(x) \text{ para algún } x \in [0, \theta(6)] \\ 0 & \text{En otro caso} \end{cases}$$

En este caso, obtenemos que $h^{-1}(r) = \theta - \frac{r}{n}$, por lo que la función de densidad de R_n es

$$f_{R_n}(r) := \frac{1}{\theta^n} \left(\theta - \frac{r}{n}\right)^{n-1} I_{[0,n\theta]}(r)$$

Integramos la función de densidad a lo largo de $(-\infty, x)$ y obtenemos la función de distribución de R_n.

$$F_{R_n}(x) := \begin{cases} 0 & \text{Si } x < 0 \\ 1 - \left(1 - \frac{x}{n\theta}\right)^n & \text{Si } 0 \leq x \leq n\theta \\ 1 & \text{Si } x > n\theta \end{cases}$$

Cuando $n \longrightarrow \infty$, obtenemos que

$$\lim_{n\longrightarrow\infty} F_{R_n}(x) = \begin{cases} 0 & \text{Si } x < 0 \\ \lim\limits_{n\longrightarrow\infty} 1 - \left(1 - \frac{x}{n\theta}\right)^n & \text{Si } 0 \leq x \end{cases}$$

Es decir

$$F_R(x) := \begin{cases} 0 & \text{Si } x < 0 \\ 1 - e^{-\frac{x}{\theta}} & \text{Si } x \geq 0 \end{cases}$$

La cual es la función de distribución de $R \sim exp(\frac{1}{\theta})$. Por lo tanto

$$R_n \xrightarrow{d} R \sim exp\left(\frac{1}{\theta}\right)$$

Con $\theta = 6$. ■

B. Cálculo de una estimación máximo-verosímil cuando no hay una solución analítica

Considere una muestra aleatoria $X_1, X_2, \ldots, X_n$ de distribución $Cauchy(\theta = \kappa, 1)$, es decir, donde el verdadero valor del parámetro de localización es igual al número de grupo ($\kappa = 6$). Se sabe que en este caso, el estimador MLE no tiene expresión analítica, pero es posible estudiar las propiedades del estimador máximo-verosímil haciendo uso de la simulación y de métodos numéricos como la función `optim` de R.

a) Genere $m = 1000$ simulaciones de una muestra aleatoria de tamaño $n = 10$ de la distribución $Cauchy(\theta = 6, 1)$, usando la función `rcauchy` de R o su equivalente en otro lenguaje de programación.Para cada una de las m muestras, calcule y almacene la estimación máximo-verosímil del parámetro y la mediana (como una estimación por analogía). Haga un histograma de las m estimaciones ML y otro de las m estimaciones por analogía. Añada una linea vertical en cada histograma para indicar el verdadero valor del parámetro y otra linea (de diferente color) para indicar el promedio las m estimaciones en cada caso.Reporte en una tabla el promedio y la varianza obtenidos de las m estimaciones para este tamaño de muestra. ¿Hay indicios de que ambos estimador sean insesgados? Establezca qué estimador es mejor en términos de error cuadrático medio (estimado).

b) Repita el procedimiento descrito en **a.** variando los tamaños de muestra por los valores $n = 50, 100, 200, 500, 1000$. ¿Se mantiene la misma conclusión acerca de qué estimador es mejor en términos de error cuadrático medio (estimado)?

c) ¿A qué convergen en distribución las siguientes secuencias?

$$V_n = \sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} ?$$

$$W_n = \sqrt{n}(Me - \theta) \xrightarrow{d} ?$$

Nota: Va a requerir calcular la información de Fisher, $I(\theta)$, así que muestre cómo la obtuvo.

d) Repitiendo el procedimiento de generar m muestras de tamaño $n = 1000$, calcule las m realizaciones de V_{1000} y de W_{1000}, realice un histograma de cada una de ellas y superponga las distribuciones que encontró en el punto anterior. ¿Se ve que el resultado de la convergencia es correcto?¿Cuál de los dos estimadores es más eficiente?

e) **Bonus [+0.2]** Incluya la moda como estimador por analogía y compárelo con los otros dos estimador en los apartados **a** y **b** La moda no puede ser estudiada como variable aleatoria de manera analítica ya que no tiene una dependencia funcional clara de la muestra aleatoria. Además, como las variables son continuas, no se puede calcular la moda como "el dato que más se repite", y es necesario usar la fórmula de la moda para datos agrupados. Explore la función `mlv` del paquete `modeest` en R y explore los métodos que allí proponen (use el método '`naive`' como uno de ellos).

Solución B

Consideremos una muestra aleatoria $X_1, X_2, \ldots, X_n$ de distribución $Cauchy(6, 1)$

a.

Generamos 1000 simulaciones de una muestra aleatoria de tamaño 10 de nuestra distribución. Luego calculamos y almacenamos la estimación máximo-verosímil del parámetro y la mediana.
Para ello, después de generar las 1000 simulaciones, definimos el negativo de la log-verosimilitud de la función de densidad de Cauchy, y usando la función `optim` de R encontramos el MLE de cada una de las simulaciones de nuestras muestras de tamaño 10. Luego, vamos almacenando con cada simulación el MLE y la mediana.

Estos son los resultados:

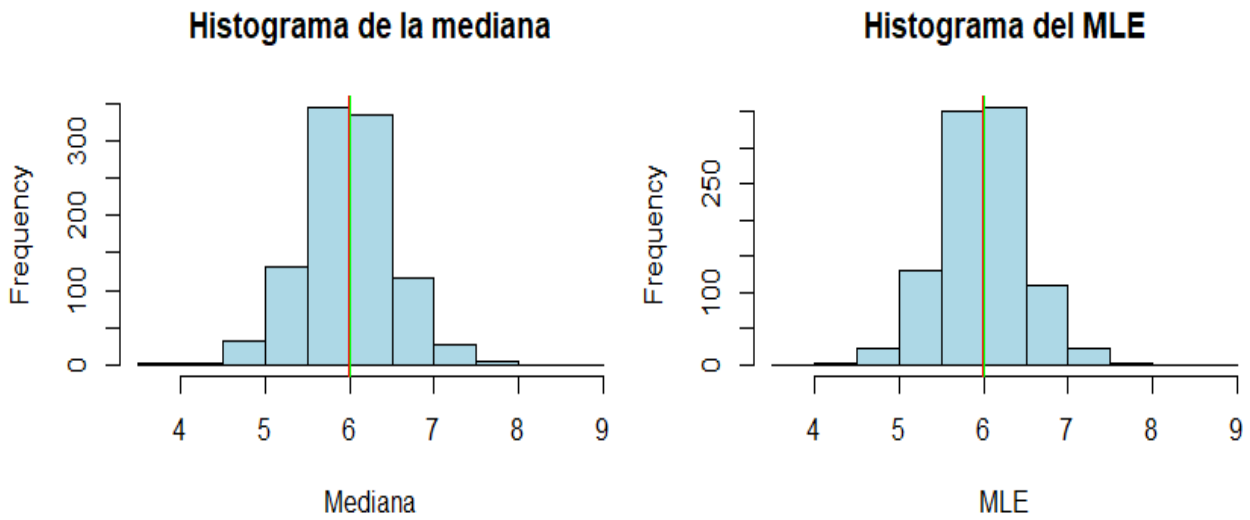


Figura 6: Histograma de las 1000 estimaciones del MLE y la mediana

De los histogramas nos podemos dar cuenta de el promedio de las 1000 estimaciones del MLE y de la mediana están muy cerca del verdadero valor del parámetro ($\theta = 6$). Tal infor-

mación la podemos ver en la siguiente tabla:

$N = 10$	Promedio	Varianza
Mediana	5.9888	0.3201
MLE	5.9944	0.2972

Cuadro 1: Promedio y varianza de las 100 estimaciones para N=10

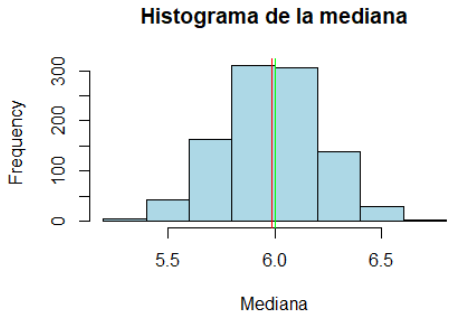
Vemos que efectivamente hay indicios para afirmar que ambos estimadores son insesgados, ya que están muy próximos al verdadero valor del parámetro, y esto considerando que tan solo estamos tomando una muestra de tamaño 10.
Ahora con estos datos obtengamos una estimación del error cuadrático medio. Usando la siguiente formula en R:

```
(mean(data)-6)^2+var(data)
```

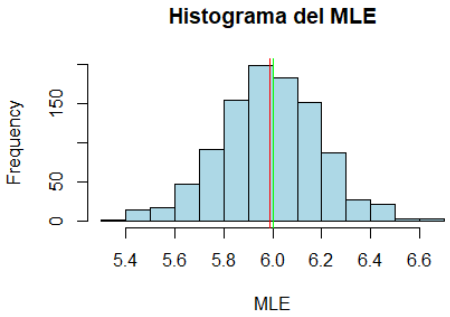
obtenemos que el error cuadrático medio de a mediana es 0.32022 y el error cuadrático medio del MLE es 0.262416. Por lo tanto, vemos que en términos del error cuadrático medio, el estimador de máxima verosimilitud es mas eficiente.

b.

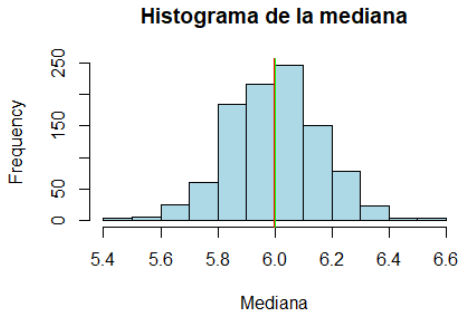
Repetimos este proceso para tamaños de muestra mas grande ($n = 50, 100, 200, 500, 1000$). Los histogramas del MLE y la mediana son



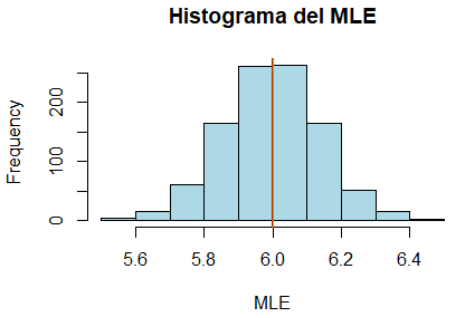
(a) Histograma de la mediana con $n = 50$.



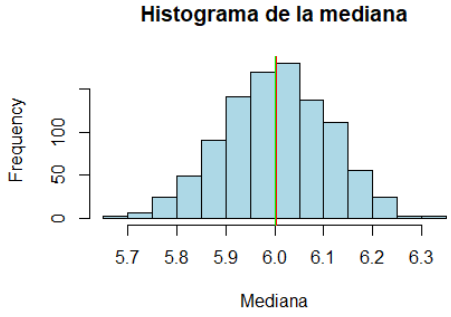
(b) Histograma del MLE con $n = 50$.



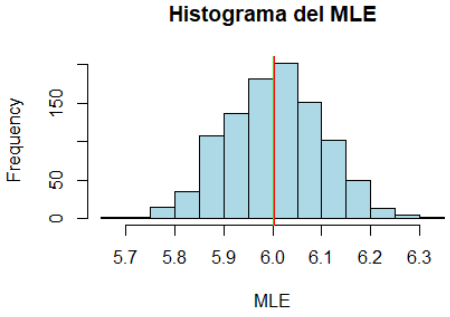
(c) Histograma de la mediana con $n = 100$.



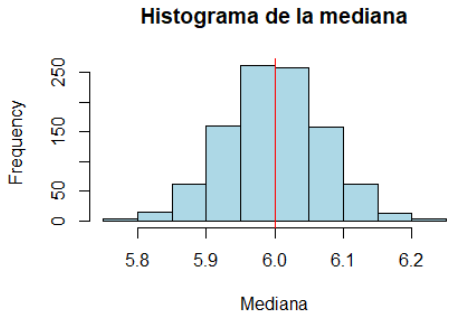
(d) Histograma del MLE con $n = 100$.



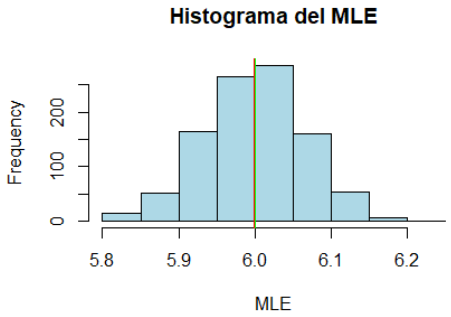
(e) Histograma de la mediana con $n = 200$.



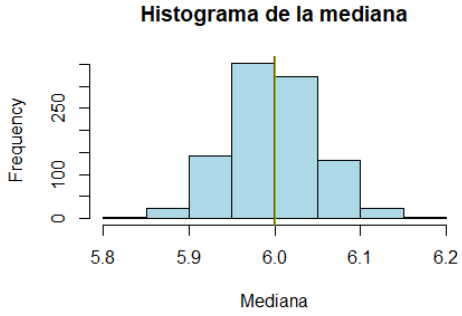
(f) Histograma del MLE con $n = 200$.



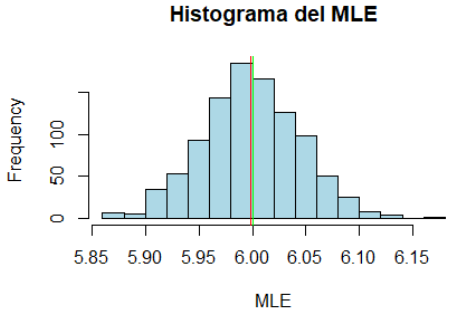
(a) Histograma de la mediana con $n = 500$.



(b) Histograma del MLE con $n = 500$.



(c) Histograma de la mediana con $n = 1000$.



(d) Histograma del MLE con $n = 1000$.

Y la tabla de los datos es

Población	Estimador	Promedio	Varianza	Sesgo	MSE
n=50	Mediana	5.9996	0.05058	-0.0004	0.05058
	MLE	5.9996	0.04163	-0.00044	0.04163
n=100	Mediana	6.00266	0.02350	0.00266	0.02351
	MLE	6.00266	0.01953	-0.0012	0.01953
n=200	Mediana	6.0014	0.0132	0.0014	0.0132
	MLE	5.9999	0.01037	-0.00011	0.01037
n=500	Mediana	6.0011	0.0044	0.0011	0.0044
	MLE	6.0012	0.00375	0.0012	0.0038
n=1000	Mediana	6.0012	0.00236	0.00124	0.0024
	MLE	6.0012	0.0019	0.0012	0.0019

Cuadro 2: Información de cada simulación con diferentes tamaños de muestra

De estas simulaciones podemos deducir lo siguiente, viendo las gráficas podemos observar que los datos obtenidos están cada vez mas cerca del valor del parámetro. También que la linea roja (promedio de las estimaciones) y la linea verde (verdadero valor del parámetro) están cada vez mas cerca pero siempre se mantienen cercanas. Además, viendo los datos podemos ver que con una muestra mas grande el error cuadrático medio se hace mas pequeño, pero esto principalmente ya que con un mayor tamaño de muestra, la varianza disminuye, y no por

el sesgo que por lo visto se mantiene en un valor muy cercano a 0. Y vemos que en términos de error cuadrático medio, el estimador MLE es siempre mas eficiente.

c.

Veamos a qué convergen en distribución las siguientes secuencias. Para

$$V_n = \sqrt{n}(\hat{\theta}_{MLE} - \theta)$$

Por el teorema de la distribución asintótica del MLE tenemos que

$$V_n = \sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right)$$

Por lo tanto, vamos a calcular $I(\theta)$. Tenemos que la función de distribución de nuestra muestra aleatoria es

$$f(x) = \frac{1}{\pi(1 + (x-\theta)^2)}$$

Entonces tenemos

$$\ln(f(x)) = -\ln(\pi) - \ln(1 + (x-\theta)^2)$$

Y luego,

$$\left(\frac{\partial}{\partial\theta}\ln(f(x))\right)^2 = \frac{4(x-\theta)^2}{(1 + (x-\theta)^2)^2}$$

Y por ultimo,

$$I(\theta) = E\left(\left(\frac{\partial}{\partial\theta}\ln(f(x))\right)^2\right) = \int_{-\infty}^{\infty} \frac{4(x-\theta)^2}{(1 + (x-\theta)^2)^2} \cdot \frac{1}{\pi(1 + (x-\theta)^2)} dx$$

Calculando esta integral nos queda que $I(\theta) = \frac{1}{2}$. Por lo que

$$V_n = \sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} N(0, 2)$$

Ahora vamos a encontrar la distribución asintótica de $W_n = \sqrt{n}(Me - \theta)$. Para ello vamos a usar el teorema de la distribución asintótica de una estadística de orden. esto es que

$$W_n = \sqrt{n}(Me - \theta) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f^2(\theta)}\right)$$

Donde p se asocia a la probabilidad de la mediana, es decir $p = 0.5$, y también sabemos que como la distribución es $f(x) = \frac{1}{\pi(1+(x-\theta)^2)}$. Entonces

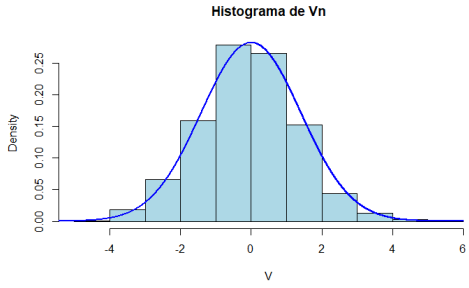
$$f(\theta) = \frac{1}{\pi}.$$

Por lo tanto

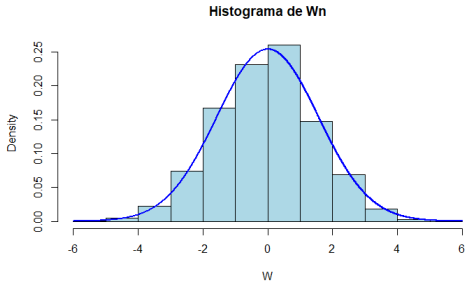
$$W_n = \sqrt{n}(Me - \theta) \xrightarrow{d} N\left(0, \frac{\pi^2}{4}\right)$$

d

Repitiendo el proceso de hacer las 1000 realizaciones de V_{1000} y de W_{1000} tenemos el siguiente histograma junto con las distribuciones encontradas anteriormente



(a) Histograma de V_{1000}.



(b) Histograma de W_{1000}.

Vemos que el resultado de la convergencia de V_n W_n concuerdan con los resultado del punto anterior, ya que la linea azul representa $N(0, 2)$ y $N(0, \frac{\pi}{4})$ los cuales se ajustan correctamente al histograma de la secuencia V_n y W_n respectivamente.
Por la distribución asintótica del MLE y resultados dados en el taller 2, tenemos que el MLE es más eficiente asintóticamente que Me. ■

C. Comparación de estimador ML y de momentos en un modelo doble exponencial

Considere una muestra aleatoria $X_1, X_2, ..., X_n$ de una distribución *Double_Exp*$(\theta = 6, 1)$. Recuerde que , en este caso, $\hat{\theta}_{MLE} = Me$ y $\hat{\theta}_{MOM} = \overline{X}_n$.

a) Genere $m = 100$ simulaciones de una muestra aleatoria de tamaño $n = 10$ de la distribución *Double_Exp*$(\theta = 6, 1)$, usando la función rdexp del paquete de nimble de R o su equivalente en otro lenguaje de programación. Para cada una de las m muestras, calcule y almacene la estimación máximo-verosímil del parámetro y la estimación de momentos. Haga un histograma de las m estimaciones ML y otro de las m estimaciones por momentos. Añada una línea vertical en cada histograma para indicar el verdadero valor del parámetro y otra línea (de diferente color) para indicar el promedio de las m estimaciones en cada caso. ¿Hay indicios de que ambos estimadores sean insesgados? Reporte en una tabla el promedio y la varianza obtenidos de las m estimaciones para este tamaño de muestra. ¿Hay indicios de que ambos estimadores sean insesgados? Establezca qué estimador es mejor en términos de los valores de error cuadrático medio (estimado).

b) Repita el procedimiento descrito en a) , variando los tamaños de muestra por los valores $n = 50, 100, 200, 500, 1000$. ¿Se mantiene la misma conclusión acerca de qué estimador es mejor en términos de error cuadrático medio (estimado)?

c) ¿A qué convergen en distribución las siguientes secuencias?

$$Z_n = \sqrt{n}(\overline{X}_n - \theta) \xrightarrow{d} ?$$

$$M_n = \sqrt{n}(Me - \theta) \xrightarrow{d} ?$$

d) Repitiendo el procedimiento de generar m muestras de tamaño $n = 1000$, calcule m realizaciones de Z_{1000} y de M_{1000}, realice un histograma de cada una de ellas y superponga las distribuciones que encontró en el punto anterior. ¿Se ve que el resultado de la convergencia es correcto? ¿Cuál de los dos estimadores es asintóticamente más eficiente?

e) **Bonus [+0.2].** Incluya la Moda como estimador por analogía y compárelo con los otros dos estimadores en los apartados a) y b) La moda no puede ser estudiada como variable aleatoria de manera analítica ya que no tiene una dependencia funcional clara de la muestra aleatoria. Además, como las variables son continuas, no se puede calcular

la moda como "el dato que más se repite", y es necesario usar la fórmula de la moda para datos agrupados. Explore la función mlv del paquete modeest en R y explore los métodos que allí se proponen (use el método 'naive' como uno de ellos).

Solución C

Consideremos una muestra aleatoria $X_1, X_2, \ldots, X_n$ de distribución *Doble Exponencial*$(6, 1)$. Teniendo en cuenta que $\widehat{\theta}_{MLE} = Me$ y $\widehat{\theta}_{MOM} = \overline{X}_n$.

a.

Se genera 1000 simulaciones de una muestra aleatoria de tamaño $n = 10$ de la distribución. las almacenamos y tomamos una de las mil para ver un histograma al cual le sobrepondremos un linea representando la distribución

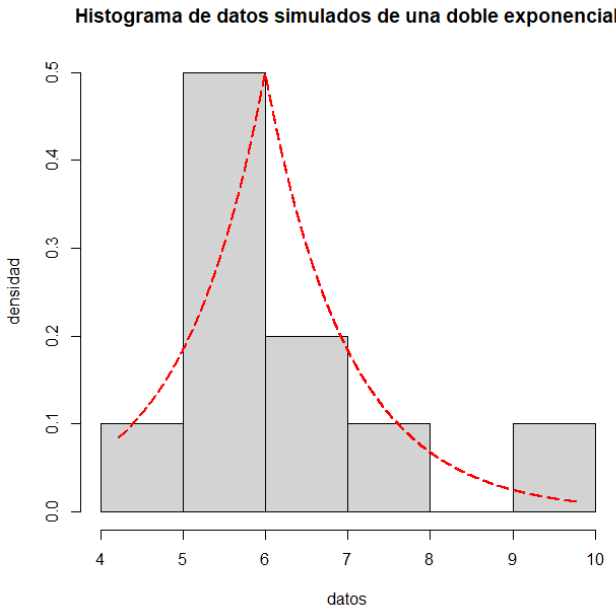
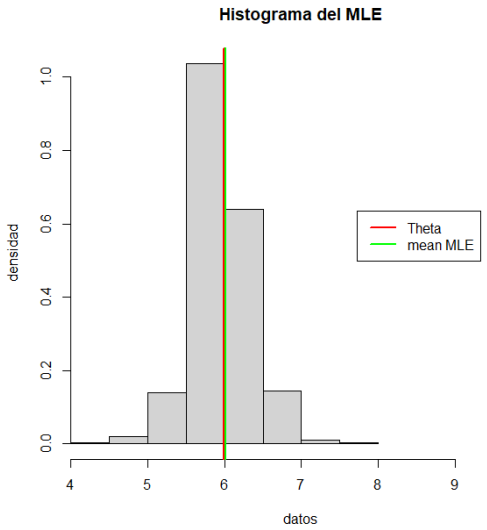
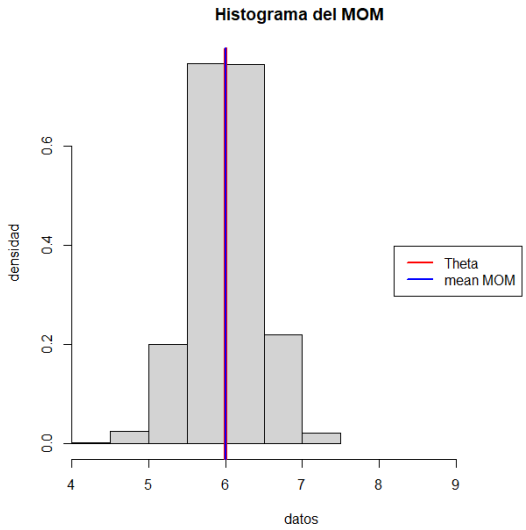


Figura 10: Histograma de una simulación de distribución doble exponencial.

Luego para calcular el estimador `MLE` y el `MOM` creamos una matriz para almacenarnos. Para el `MLE` implementamos una función que genera el negativo de la log-verosimilitud, después haciendo uso de la función `optim` de R con ellos encontramos el `MLE` para cada una de las simulaciones de tamaño 10.Para el `MOM` usaremos la función `colMeans`.
Realizamos unos histogramas para las estimaciones del `MLE` y para el `MOM`.



(a) Histograma de la mediana con $n = 10$.



(b) Histograma del media muestral con $n = 10$.

Estos histogramas nos dan indicios de que las medias de ambos el `MLE` y el `MOM` para las 1000 simulaciones están bastante cerca al valor real. En la siguiente tabla se presenta la media y la varianza para ambos.

	MLE	MOM
Mean	6.011	6.006
Var	0.156	0.192

Cuadro 3: Promedio y varianza de las 1000 estimaciones para n=10

Con toda esta información podemos decir que para las 1000 simulaciones de tamaño 10 ambos estimadores son aproximadamente insesgado por ende el `MSE` estimado queda solo en términos de la varianza y como se ve en el Cuadro 3 el `MLE` reporta la menor varianza; por ende, podemos decir que $\widehat{\theta}_{MLE} = Me$ es el mejor estimador en términos del error cuadrático medio.

b.

Repetimos este proceso para tamaños de muestra mas grande ($n = 50, 100, 200, 500, 1000$). Los histogramas del `MLE` y el `MOM` son:

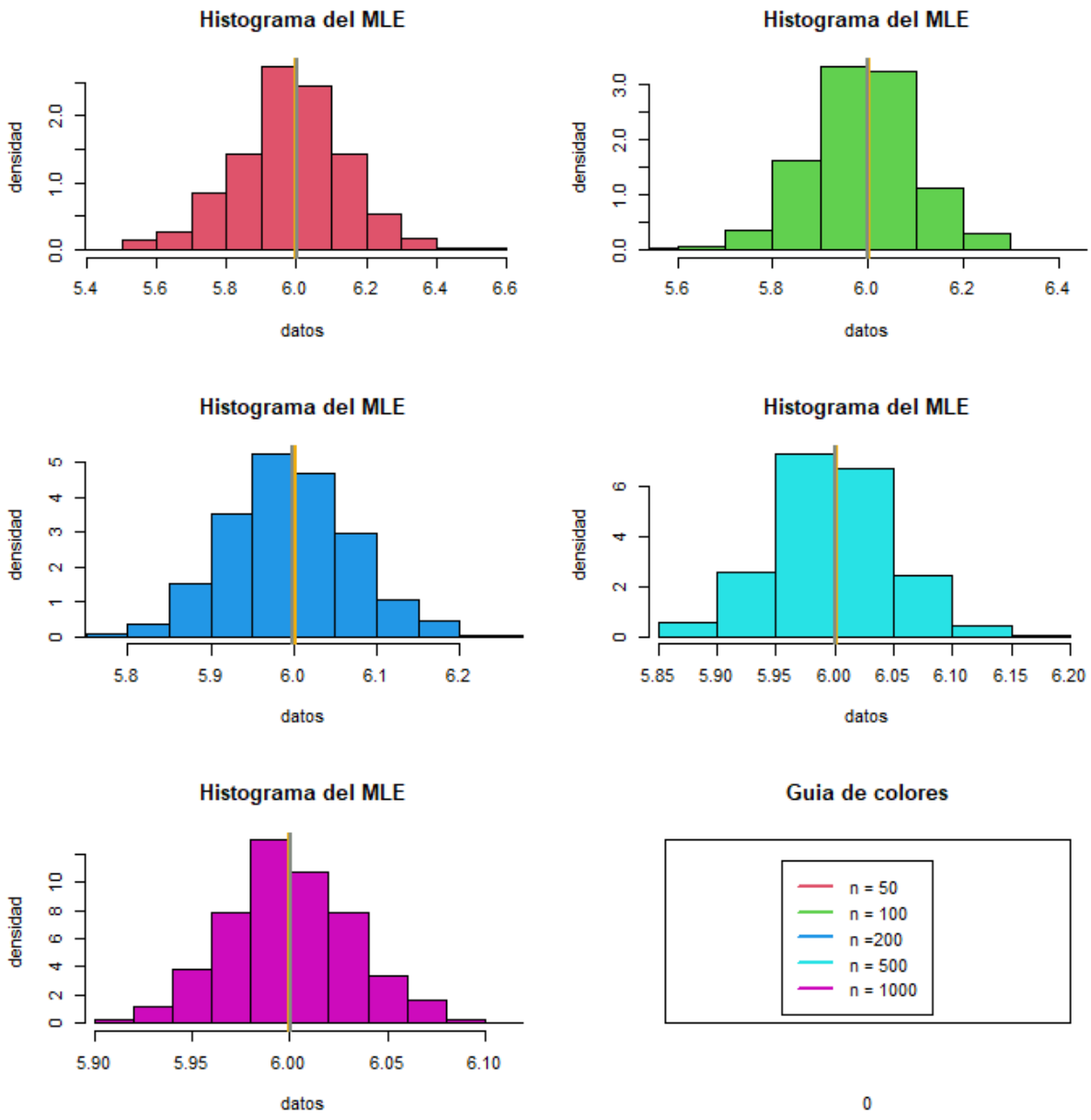


Figura 12: Histogramas para diferentes tamaños de muestras de la mediana

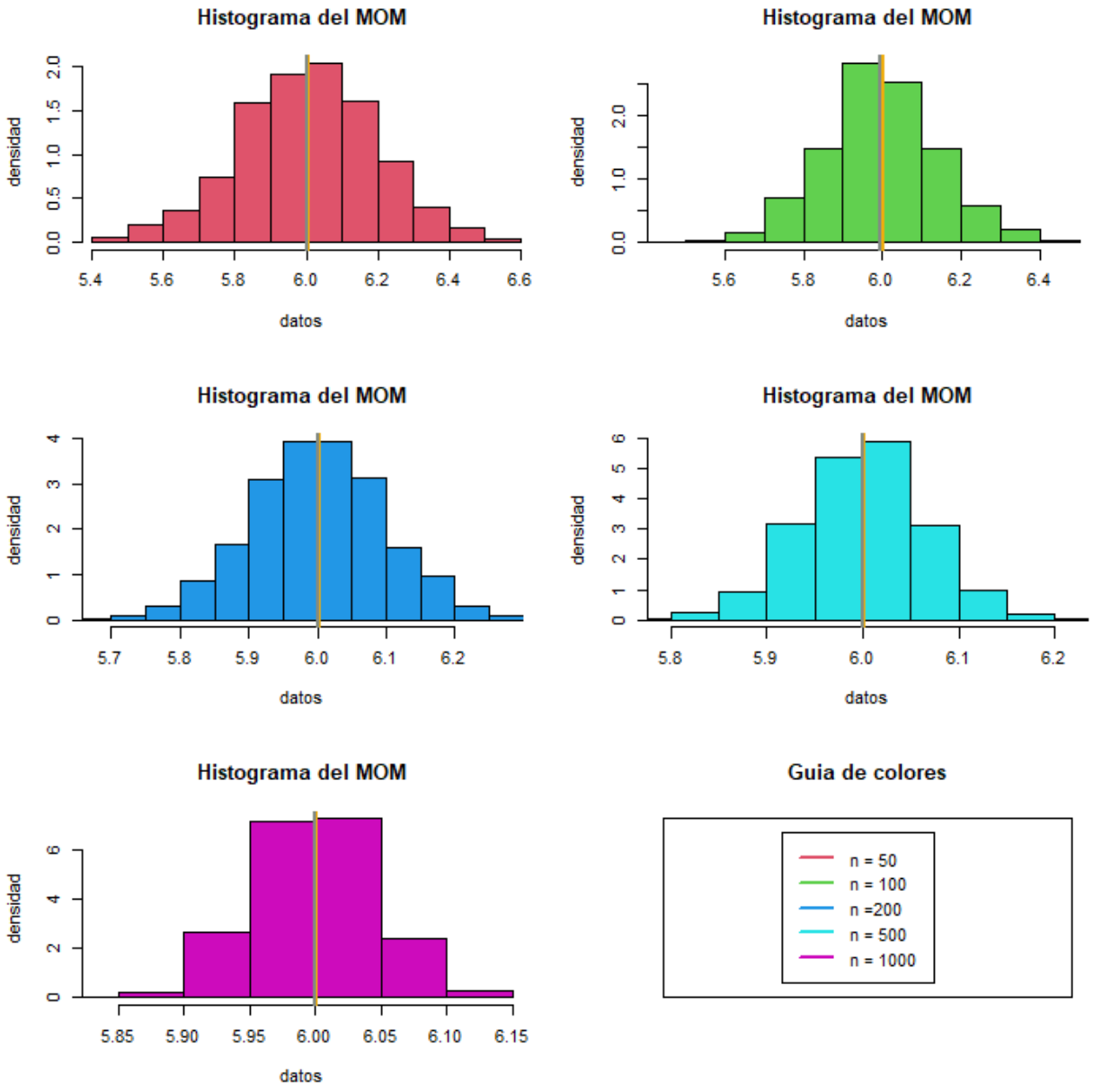


Figura 13: Histogramas para diferentes tamaños de muestras de la media muestral

Las tablas para estos resultados son:

Muestra	n =	50	n =	100	n =	200	n =	500	n =	1000
	MLE	MOM	MLE	MOM	MLE	MOM	MLE	MOM	MLE	MOM
Mean	6.003	6.002	5.999	5.994	5.999	6.000	6.000	6.000	6.000	6.000
Var	0.025	0.037	0.012	0.020	0.006	0.010	0.002	0.004	0.001	0.002

Cuadro 4: Promedio y varianza de las 1000 estimaciones para diferentes tamaños de muestras

Como se puede observar de los histogramas de las figuras 12 y 13, además del cuadro 4 toda esta información nos permite decir que se mantiene la misma conclusión con respecto al punto anterior donde se llaga a que el `MLE` es el mejor en términos del erro cuadrático medio estimado.

c.

Dada una muestra aleatoria $X_1, X_2, \ldots, X_n$ con media θ y varianza $2(1)^2 = 2$, por el TCL tenemos que:

$$Z_n = \sqrt{n}(\overline{X}_n - \theta) \xrightarrow{d} N(0, 2)$$

Por las propiedades asintóticas del `MLE` podemos observar que:

$$f_X(x, \theta) = \frac{1}{2}e^{-|x-\theta|} \qquad\qquad ln(f_X(x, \theta)) = -ln(2) - |x - \theta|$$

Usando resultados de teoría de la medida, es posible obtener la información de Fisher, la cual es:

$$\frac{\partial}{\partial\theta} ln(f_X(x, \theta)) = \frac{\partial}{\partial\theta}(-ln(2) - |x - \theta|) = \frac{x - \theta}{|x - \theta|}$$

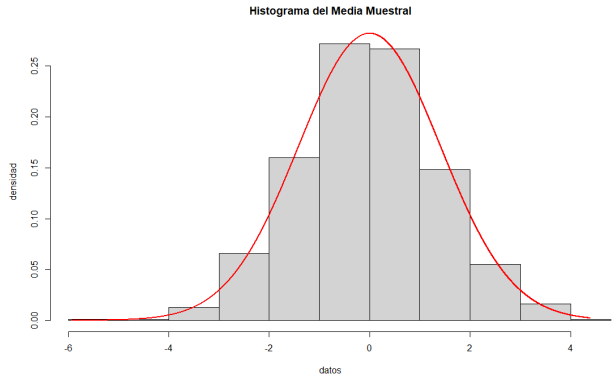
$$I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta} ln(f_X(x, \theta))\right)^2\right] = \left(\frac{x - \theta}{|x - \theta|}\right)^2 = 1$$

Así, tenemos que:

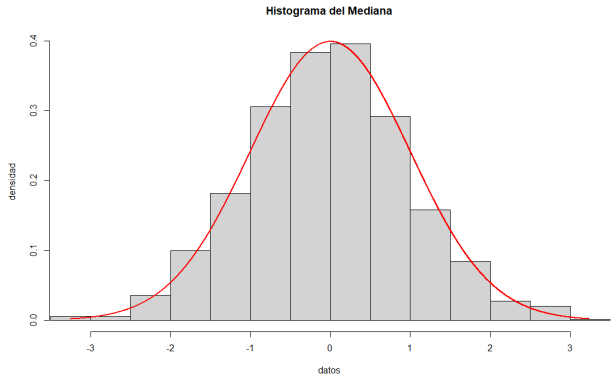
$$M_n = \sqrt{n}(Me - \theta) \xrightarrow{d} N(0, 1)$$

d.

Repetimos el proceso de generar 1000 simulaciones de tamaño $n = 1000$ almacenamos estos datos en un matriz; ahora vamos usar las funciones `colMeans` de R para la media de las simulaciones y `median` de R para las medianas de las simulaciones.
luego calculamos los valores para Z_n y M_n para cada simulación para poder hacer un histograma de cada uno y sobreponer la distribución encontrada en el punto anterior.



(a) Histograma de la Z_N con $n = 1000$.



(b) Histograma del M_n con $n = 1000$.

Se puede observar en los histogramas que la convergencia es correcta; además que M_n es asintóticamente más eficiente pues se ajusta mejor el resultado de la convergencia. ■

D. Técnica de remuestreo (Bootstrap)

En ejercicios de simulación, cuando es necesario calcular algunas propiedades de una variable aleatoria, poder generar muestras de la verdadera distribución de los datos es de gran utilidad para verificar que los resultados teóricos se tienen. Sin embargo, en la vida real, rara vez se conoce la "verdadera" distribución de los datos y solo se cuenta con un conjunto de datos $x_1, x_2..., x_n$ recogidos.

Ahora bien, uno podría hacer un histograma de los datos, suponer un modelo para ellos a partir de la evidencia del histograma y luego, estimar los correspondientes parámetros. Sin embargo, es posible que la elección de ese modelo "adecuado" no sea tan sencilla y quisiéramos poder contar con un procedimiento que nos permitiera estudiar a una determinada estadística SOLO con la información que tienen los datos, sin necesidad de recurrir a hacer suposiciones sobre su distribución.

a) Genere **una única** muestra simulada de tamaño $n = 10$ de una distribución $N(\mu = \kappa, \sigma^2 = \kappa^2)$ donde $\kappa = 6$, usando la función rnorm de R o su equivalente en otro lenguaje de programación. Calcule para dicha muestra el promedio y la varianza muestrales. ¿Son cercanas las estimaciones muestrales a los verdaderos parámetros?

b) Haga una gráfico de la función de distribución empírica de los datos (explore la función ecdf de R), y superponga la función de distribución real de los datos con los parámetros reales. ¿Se parecen?

c) Ahora bien, vamos a olvidar que sabemos de qué distribución vienen los datos y solo se cuentan con esos $n = 10$ datos tomados que llamaremos data. Calcule $B = 1000$ muestras bootstrap, cada una de tamaña 10, con reemplazo de ese vector *data*. Explore la función sample para tal fin.

d) Para cada una de las muestras de Bootstrap, calcule la media $\{\overline{x}_i\}_{i=1}^B$; la expresión de la varianza

$$\left\{ \frac{(n-1)s_i^2}{\hat{\theta}^2} \right\}_{i=1}^B,$$

donde $\hat{\theta}^2$ es la estimación de la varianza en a), y el coeficiente de variación

$$\left\{ \frac{s_i}{\overline{x}_i} \right\}_{i=1}^B,$$

obteniendo $B = 1000$ realizaciones de cada una. Realice un histograma de cada una y

superponga a estos histogramas (en el caso de la media y de la expresión relacionada con la varianza) la función de densidad teórica de cada una.

e) ¿Qué conclusiones puede sacar de los resultados anteriores? ¿Cree que la similitud de las curvas de distribución del ítem b) está relacionada con sus hallazgos del ítem d)?

f) En relación con la muestra Bootstrap del coeficiente de variación, ¿qué tipo de distribución parece tener el estimador propuesto? ¿Parece ser un estimador insesgado? Si no, ¿de cuánto parece ser su sesgo?

g) Repita los pasos a)-f) pero con un tamaño de muestra $n = 1000$. ¿Cómo cambia sus respuestas a los incisos anteriores? ¿Por qué? Comente.

Solución D

a.

Para la muestra simulada, obtenemos un promedio muestral de $\hat{\mu} = 7.35$, varianza muestral $\hat{\sigma}^2 = 29.81$ y una desviación estándar muestral $\hat{\sigma} = 5.46$. De lo anterior, calculamos los respectivos errores absolutos y relativos con respecto a los parámetros conocidos media y desviación estándar como sigue

$$ErrorAbsProm = |\hat{\mu} - \kappa| = 1.35,$$

$$ErrorRelProm = \frac{ErrorAbsProm}{\kappa} = 0.22,$$

estos dos cálculos nos dan una idea de que con respecto al verdadero parámetro aún estamos lejos de la media; en cuanto a la desviación estándar,

$$ErrorAbsDes = |\hat{\sigma} - \kappa| = 0.54,$$

$$ErrorRelDes = \frac{ErrorAbsProm}{\kappa} = 0.09,$$

lo cual nos da un error menor que en el caso del otro parámetro.

b.

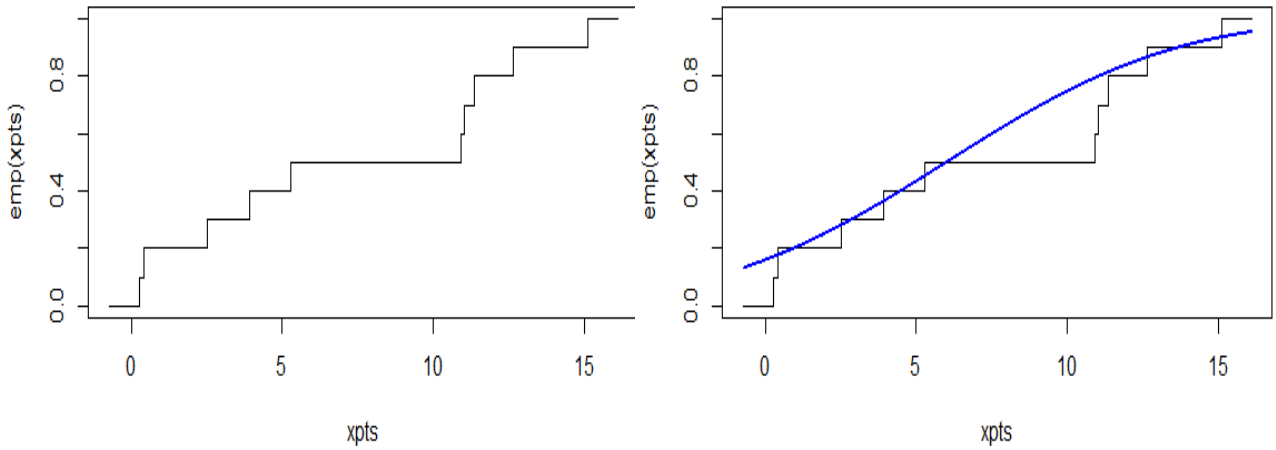


Figura 15: Distribución empírica de la muestra simulada .

De la figura vemos que al superponer la función de distribución real sobre la distribución empírica de los datos simulados en a., estas funciones no son parecidas.

c.

Con ayuda de R, calculamos $B = 1000$ muestras bootstrap, cada una de tamaño 10, con reemplazo de ese vector data, cuyo script se adjunta al proyecto.

d.

Nuevamente, con ayuda de R calculamos las muestras bootstrap de lo propuesto en el enunciado, obteniendo los siguientes histogramas. Para las medias:

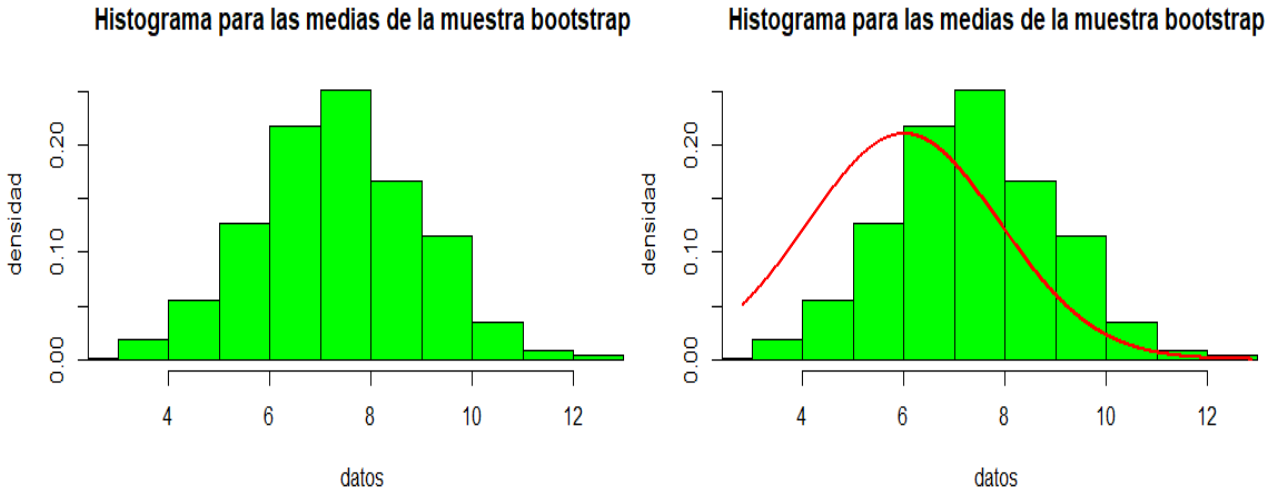


Figura 16: Histograma para las medias de las muestras bootstrap-superposición de f.d.d teórica.

En la figura anterior al lado izquierdo observamos el respectivo histograma para $\{\overline{x}_i\}_{i=1}^{B}$ y al lado derecho superponemos la función de densidad teórica, la cual proviene de una distribución $N(\kappa, \kappa^2/n) = N(6, 6^2/10)$, ello puesto que cada muestra es de tamaño 10 de una distribución $N(\kappa, \kappa^2)$ y hemos visto repetidamente en clase que su promedio muestral tiene una distribución $N(\kappa, \kappa^2/n)$.

Para la expresión de la varianza

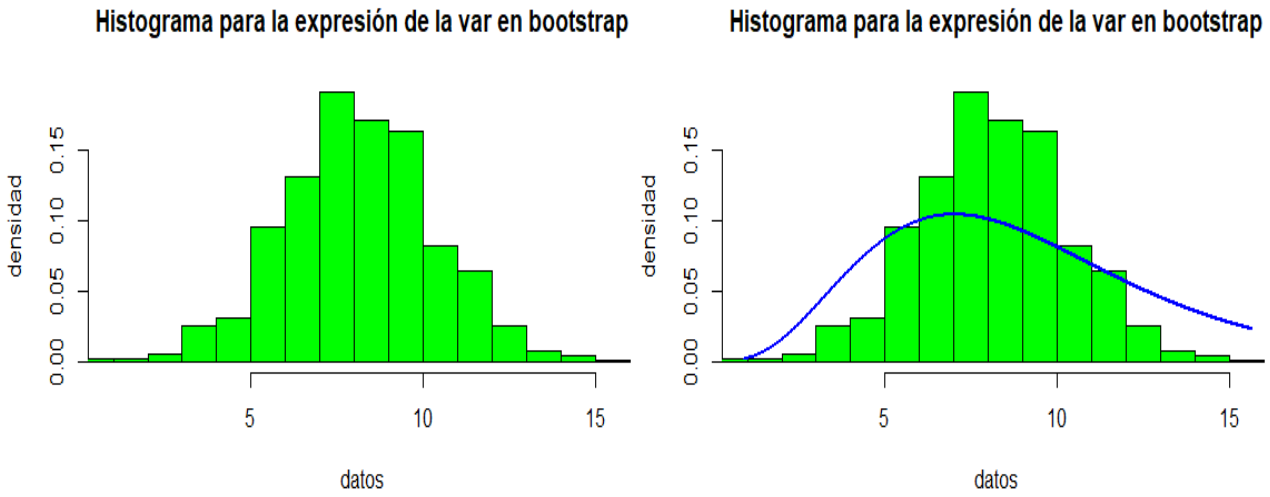


Figura 17: Histograma para las expresiones de varianza de las muestras bootstrap-superposición de f.d.d teórica.

En la figura anterior al lado izquierdo observamos el respectivo histograma para

$$\left\{ \frac{(n-1)s_i^2}{\hat{\theta}^2} \right\}_{i=1}^{B}$$

y al lado derecho superponemos la función de densidad teórica, la cual proviene de una distribución $\chi^2(n-1) = \chi^2(9)$, ello puesto que cada muestra es de tamaño 10 de una distribución $N(\kappa, \kappa^2)$ y hemos visto repetidamente en clase que para tal muestra aleatoria,

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1),$$

en este caso tomamos la variación como la estimada en el inciso a., es decir $\hat{\theta}^2 = 29.81$.

Para los coeficientes de variación tenemos el siguiente histograma

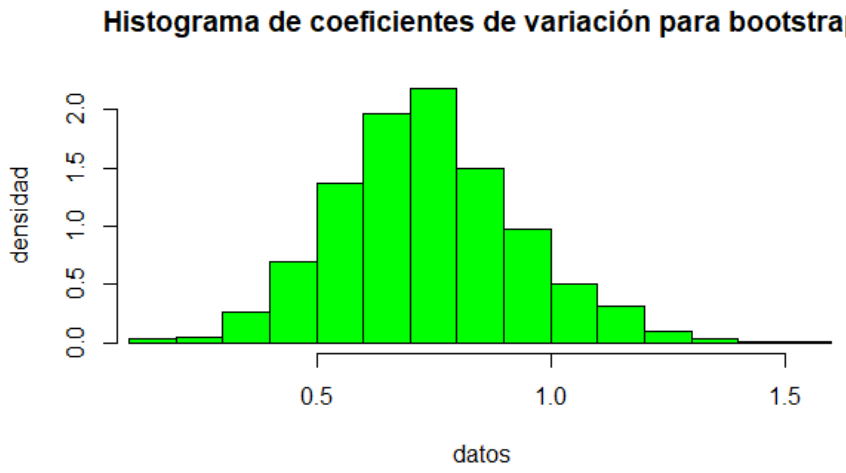


Figura 18: Histograma de coeficientes de variación para las muestras bootstrap calculadas.

e.

Vemos, así como en el inciso b., que la muestras no se están pareciendo o aproximando a las teóricas, esto era de esperarse en cuanto se hace un análisis de los resultados de a. y b., note por ejemplo que para el histograma de las medias, considerando las muestras bootstrap, el achatamiento de los datos es similar a su densidad teórica, puesto que la varianza estimada en a. no da tan alejada del parámetro real, mientras que el eje de simetría del histograma está bastante alejado de la realidad, que se puede ver en la gráfica de la Figura 15 y en los errores obtenidos para el promedio en a.

f.

Viendo a la Figura 18, intuimos según la simetría, que el estimador propuesto para

$$\left\{\frac{s_i}{\overline{x}_i}\right\}_{i=1}^{B},$$

tiene una distribución normal.

No parece ser un estimador insesgado por el siguiente motivo. Haciendo una aproximación por la ley débil de los grandes números concluimos que el promedio se parece al valor esperado del estimador; con ayuda de R obtenemos que tal promedio es 0.74 y lo superponemos al histograma de la Figura 18 como la línea roja punteada de a continuación.

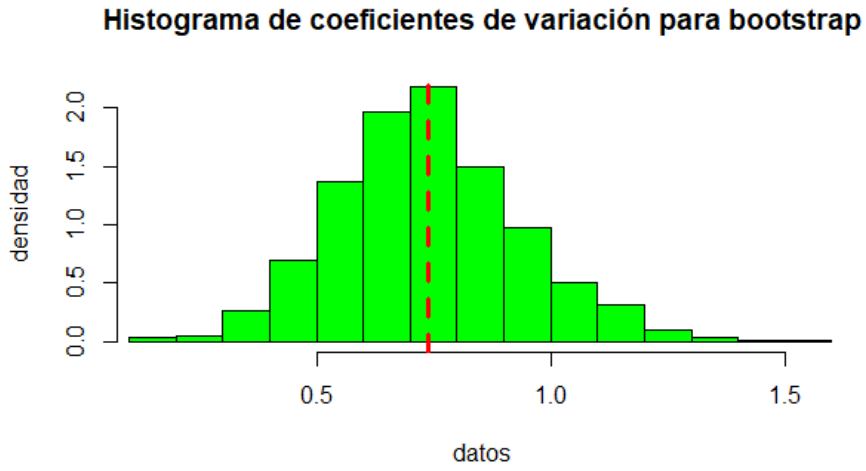


Figura 19: Histograma de los coeficientes de variación obtenidos de la muestra Bootstrap.

Sabemos además el verdadero valor del parámetro $\kappa/\kappa = 1$, puesto que la media muestral y la desviación estándar poblacionales son $\mu = \kappa$ y $\sigma = \kappa$ respectivamente. Graficamos también este parámetro a continuación como la recta vertical de color azul.

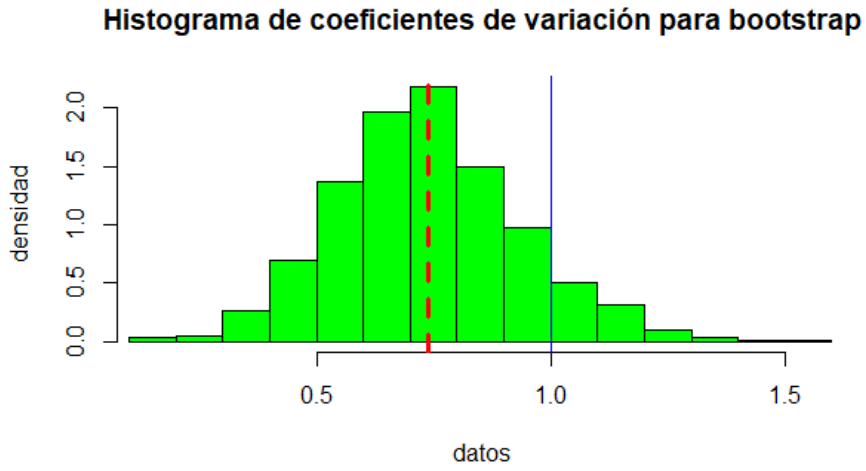


Figura 20: Histograma de los coeficientes de variación obtenidos de la muestra Bootstrap.

El sesgo es aproximadamente de $0.74 - 1 = -0.26$.

g.

Repetimos los cálculos para $n = 1000$

a.

Para la muestra simulada, obtenemos un promedio muestral de $\hat{\mu} = 6.13$, varianza muestral $\hat{\sigma}^2 = 36.72u^2$ y una desviación estándar muestral $\hat{\sigma} = 6.06$. De lo anterior, calculamos los respectivos errores absolutos y relativos con respecto a los parámetros conocidos media y desviación estándar como sigue

$$ErrorAbsProm = |\hat{\mu} - \kappa| = 0.13,$$

$$ErrorRelProm = \frac{ErrorAbsProm}{\kappa} = 0.02,$$

en cuanto a la desviación estándar,

$$ErrorAbsDes = |\hat{\sigma} - \kappa| = 0.06,$$

$$ErrorRelDes = \frac{ErrorAbsProm}{\kappa} = 0.01,$$

obteniendo en ambos casos estimaciones muestrales cercanas a los verdaderos parámetros, debido a que el tamaño de la muestra es grande.

b.

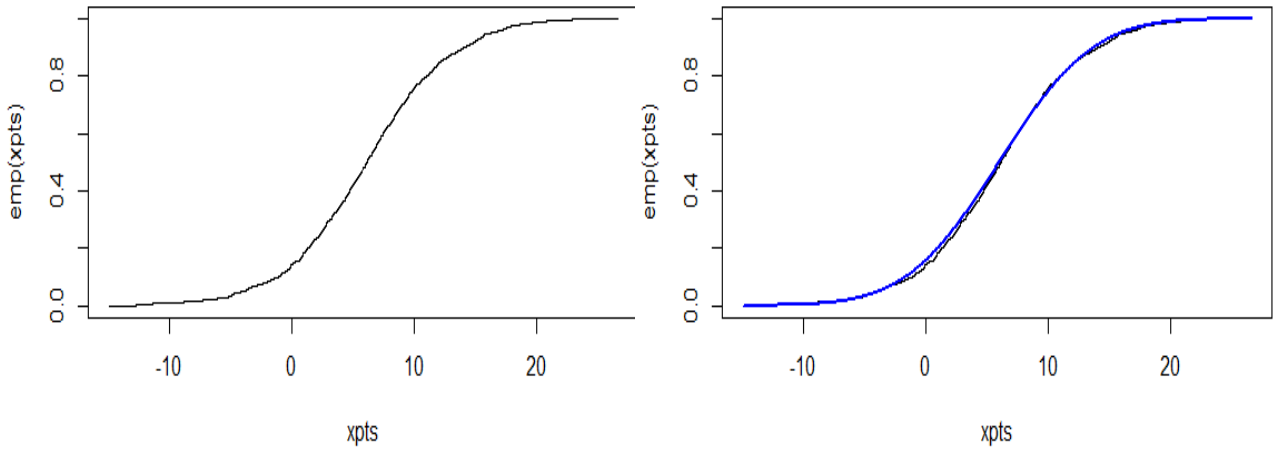


Figura 21: Distribución empírica de la muestra simulada .

Como se esperaba, según el Teorema de Glivenko-Canteli, como el tamaño de la muestra es grande, la distribución empírica es una aproximación de la verdadera distribución, es decir una distribución $N(6, 6^2)$.

c.

Con ayuda de R, calculamos $B = 1000$ muestras bootstrap, cada una de tamaño 1000, con reemplazo de ese vector data, cuyo script se adjunta al proyecto.

d.

Nuevamente, con ayuda de R calculamos las muestras bootstrap de lo propuesto en el enunciado, obteniendo los siguientes histogramas.

Para las medias

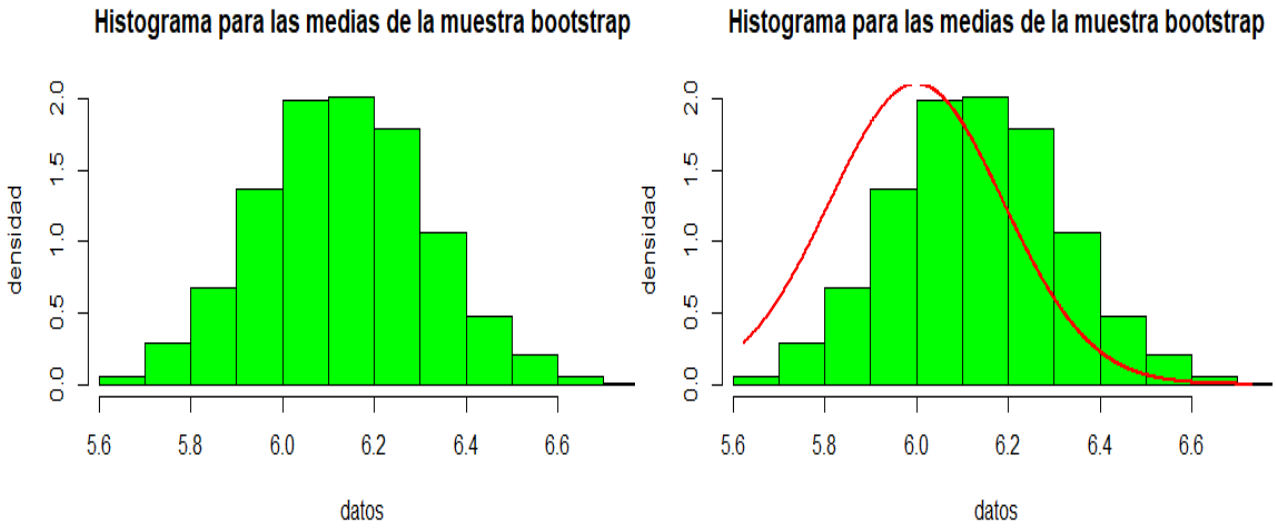


Figura 22: Histograma para las medias de las muestras bootstrap-superposición de f.d.d teórica.

Como calculamos anteriormente en el inciso d., sabemos que la distribución teórica es $N(\kappa, \kappa^2/N) = N(6, 6^2/1000)$.

Para la expresión de la varianza

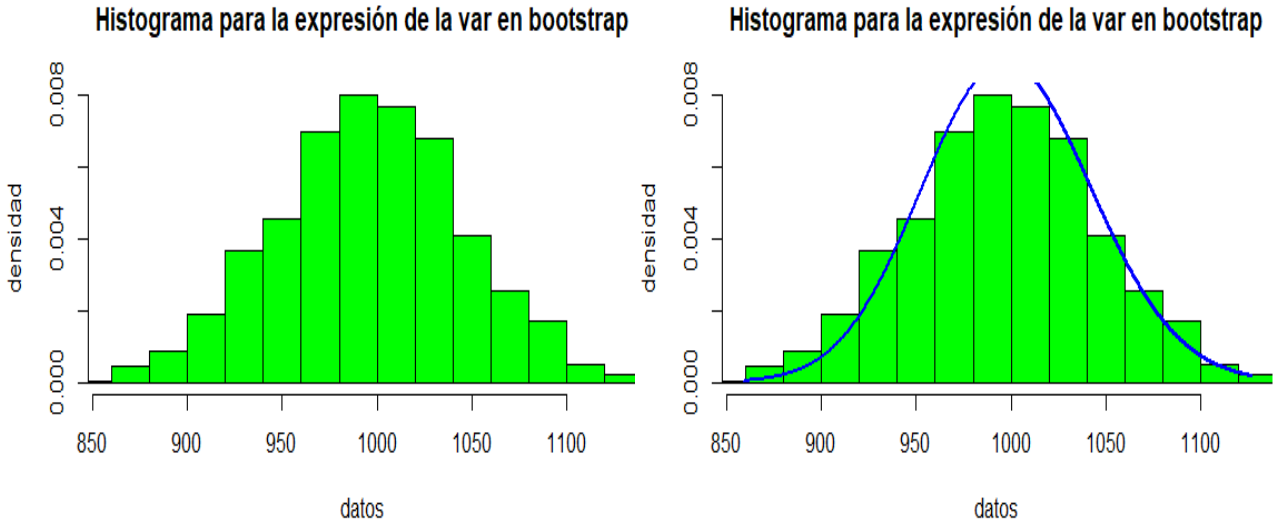


Figura 23: Histograma para las expresiones de varianza de las muestras bootstrap-superposición de f.d.d teórica.

Como calculamos anteriormente en el inciso d., sabemos que la distribución teórica es

$\chi^2(N-1) = \chi^2(999)$.

Para los coeficientes de variación tenemos el siguiente histograma

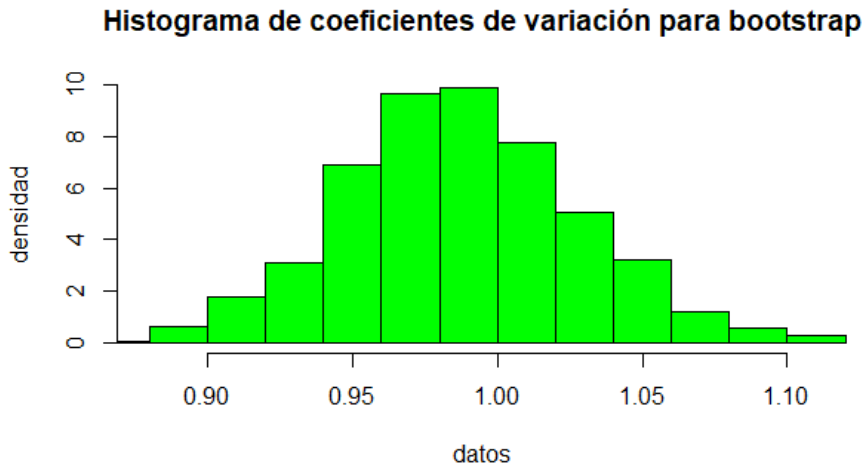


Figura 24: Histograma de coeficientes de variación para las muestras bootstrap calculadas.

e.

En este caso, el remuestreo , parece según el histograma para las medias, aún un poco alejado del eje real de simetría, pero sin duda alguna mejor que el calculado en el análisis de el inciso e. anterior. Similarmente vemos como la distribución para la expresión de la varianza, en el remuestreo, es más parecida a la teórica que cuando N era tan solo 10. Corroborando lo anterior con la distribución empírica para el caso $N = 1000$.

f.

Viendo a la Figura 24, intuimos según la simetría, que el estimador propuesto para

$$\left\{ \frac{s_i}{\overline{x}_i} \right\}_{i=1}^{B},$$

tiene una distribución normal.

Para el tamaño de la muestra trabajado aquí, el estimador propuesto parece ser insesgado por el siguiente motivo. Haciendo una aproximación por la ley débil de los grandes números para concluimos que el promedio se parece al valor esperado del estimador; con ayuda de R obtenemos que tal promedio es 0.99 y lo superponemos al histograma de la Figura 24 como la línea roja punteada de a continuación.

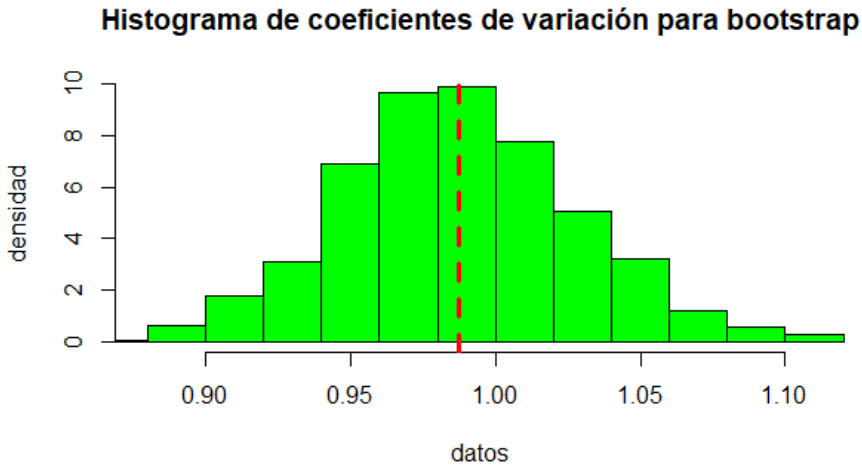


Figura 25: Histograma de los coeficientes de variación obtenidos de la muestra Bootstrap.

Sabemos además el verdadero valor del parámetro $\kappa/\kappa = 1$, puesto que la media muestral y la desviación estándar poblacionales son $\mu = \kappa$ y $\sigma = \kappa$ respectivamente. Graficamos también este parámetro a continuación como la recta vertical de color azul.

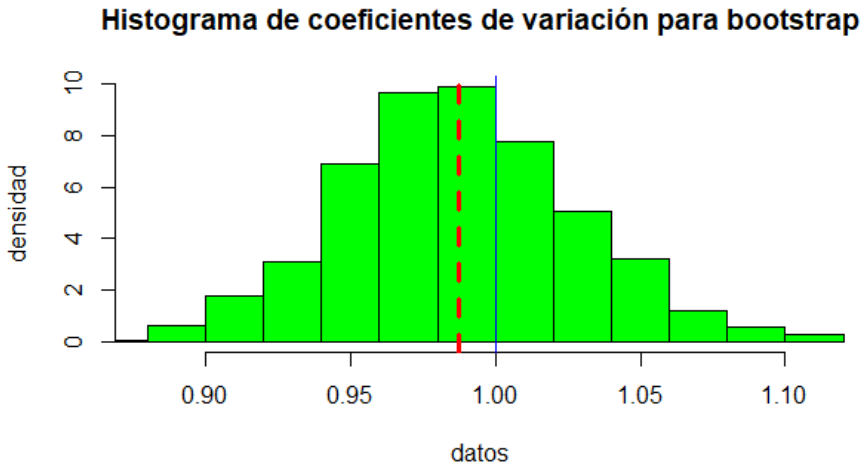


Figura 26: Histograma de los coeficientes de variación obtenidos de la muestra Bootstrap.

El sesgo es aproximadamente de $0.99 - 1 = -0.01$. ■

A. Comparación de varios intervalos de confianza para una proporción en una muestra aleatoria Bernoulli.

Vimos en clase que para estimar por intervalo la proporción, p, en el modelo Bernoulli; usando una cantidad pivote asintótica hay tres posibilidades:

Posibilidad 1: A partir del TCL se obtiene

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha.$$

siendo $\hat{p}_n = \overline{X}_n$, la media muestral. Aunque la variable aleatoria no es monótona en p, es posible obtener un intervalo de confianza despejando el parámetro de allí.

Posibilidad 2: A partir del TCL y reemplazando la varianza del modelo por un estimador consistente se obtuvo que el intervalo podía ser calculado como:

$$ICA_{100(1-\alpha)\,\%}(p) = \hat{p}_n \mp z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}}.$$

Posibilidad 3: A partir del TCL A partir del teorema central del límite y usando el método Delta con una transformación estabilizadora de varianza, se obtuvo que:

$$ICA_{100(1-\alpha)\,\%}(p) = \left[\sin^2\left\{\arcsin(\sqrt{\hat{p}_n}) - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right\}, \sin^2\left\{\arcsin(\sqrt{\hat{p}_n}) + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right\}\right].$$

Adicionalmente, se cuenta con la posibilidad de hacer muestreo por bootstrapping para obtener un intervalo de confianza para la proporción.

Posibilidad 4: A partir de la única muestra de datos que se pueda obtener, $x_1, ..., x_n$, calcule la estimación puntual $\hat{p}_{n,orig} = \overline{x}_n$. Ahora, genere $B = 1000$ muestras Bootstrap, cada una del tamaño de la muestra original y calcule las estimaciones Bootstrap de la proporción:

$$\{\hat{p}_{n,boot_i} = \overline{x}_{n.i}\}_{i=1}^{B}.$$

El bootstrap basado en percentiles simplemente definiría como límites del intervalo de confianza a:

- Límite interior = percentil $\frac{\alpha}{2}$ de la secuencia de valores $\{\hat{p}_{n,boot_i} = \overline{x}_{n.i}\}_{i=1}^{B}$.
- Límite superior = percentil $1 - \frac{\alpha}{2}$ de la secuencia de valores $\{\hat{p}_{n,boot_i} = \overline{x}_{n.i}\}_{i=1}^{B}$.

Exploren el uso de la función quantile para extraer percentiles empíricos de un arreglo de datos.

Posibilidad 5: El método anterior es generalmente criticado por no generar un intervalo centrado en $\hat{p}_{n,orig} = \overline{x}_n$, la estimación puntual obtenida con la muestra original. Para ello, se usa el método de *Bootstrap empírico*. Calcule la secuencia de diferencias de la estimación puntual original con cada muestra bootstrap:

$$\{\hat{\delta}_{boot_i} = \hat{p}_{n,boot_i} - \hat{p}_{n,orig}\}_{i=1}^{B},$$

luego, calcule

- $\hat{\delta}_{\frac{\alpha}{2}}$ = percentil $\frac{\alpha}{2}$ de la secuencia de valores $\{\hat{\delta}_{boot_i}\}_{i=1}^{B}$,
- $\hat{\delta}_{1-\frac{\alpha}{2}}$ = percentil $1 - \frac{\alpha}{2}$ de la secuencia de valores $\{\hat{\delta}_{boot_i}\}_{i=1}^{B}$,

para finalmente calcular el intervalo como

$$ICBoot_{100(1-\alpha)\,\%}(p) = [\hat{p}_{n,orig} - \hat{\delta}_{1-\frac{\alpha}{2}}, \hat{p}_{n,orig} - \hat{\delta}_{\frac{\alpha}{2}}].$$

Nota: Aunque existe el paquete boot que tiene la función boot.ci, en este ejercicio deben hacer sus propias implementaciones para generar los intervalos correspondientes.

Para el informe:

1. Muestren el procedimiento que siguieron para deducir la expresión del intervalo en la posibilidad 1.
2. Implementen el siguiente algoritmo:

a. Para cada valor de p entre $0.05, 0.1, 0.15, ..., 0.85, 0.9, 0.95$:

b. Para cada tamaño de muestra n entre 5, 10, 50, 100, 200, 500, 1000:

c. Repetir este procedimiento $m = 1000$ veces:

c1. Generen una muestra aleatoria de tamaño n de una Bernoulli con parámetro p.

c2. Calculen para dicha muestra los intervalos del 95% de confianza obtenidos por cada una de las 5 posibilidades[a].

c3. Almacenen para cada método (posibilidad) y para cada una de las m simulaciones, un 1 si dicho intervalo contiene al verdadero parámetro (p), 0 en otro caso; y en caso de que contenga al verdadero valor del parámetro, guarden la longitud del intervalo (LS-LI).

b1. Una vez hecho esto para todas las m repeticiones, resuman para cada posible método de intervalos, la cobertura promedio y la longitud media del intervalo.

$$Cob.prom = \frac{\text{Cantidad de intervalos que contuvieron al parámetro}}{1000}$$

$$Long.prom = \frac{\text{Suma de longitud de los int. que contuvieron al parámetro}}{\text{Cantidad de intervalos que contuvieron al parámetro}}$$

[a]Para las posibilidades 4 y 5, es necesario que utilicen $B = 1000$ muestras para cada una de las posibilidades. Lo idea, sería trabajar con funciones para cada posibilidad, que se llamen en el código principal.

Al final, ustedes deben contar con las medidas de cobertura promedio y longitud promedio para cada una de las posibilidades, y para cada combinación de valor del parámetro y del tamaño de muestra.

3. Para cada posibilidad (1-5) hagan dos gráficos. Uno que muestre en el eje horizontal los valores del parámetro p y en el eje vertical las coberturas promedio para cada tamaño de muestra (deben aparecer varias curvas, una para cada tamaño de muestra). El segundo gráfico debe mostrar en el eje vertical, la longitud promedio de los intervalos.

4. Concéntrense ahora en los tamaños de muestra $n = 5, 50, 200, 1000$. Para cada tamaño de muestra, comparen en un mismo gráfico la cobertura promedio de las 5 posibilidades en función del parámetro p en el eje horizontal. Hagan otros cuatro gráficos comparando la longitud promedio de los intervalos de las 5 posibilidades.

5. ¿Qué conclusiones respecto a la efectividad de las diferentes posibilidades en función del tamaño de muestra y del valor del parámetro sacarían? Pueden hacer otros gráficos

que ayuden a soportar sus conclusiones.

6. Para cada una de las siguientes dos situaciones, calculen un intervalo de confianza de la proporción de interés usando el método que consideran más adecuado, de acuerdo con la información suministrada[1].

   a) En un pequeño estudio hecho, se verificó que, de 10 componentes de aire acondicionado testeados, 8 cumplieron con los estándares de producción. ¿Qué podría decirse de la proporción de componentes que cumplen con los estándares en la población con una confianza del 90 %?

   b) Se realizó una encuesta virtual a 100 estudiantes de la UNAL-sede Bogotá, seleccionados al azar, con el fin de conocer cómo emplean su tiempo libre y cuáles son sus hobbies favoritos. Se les preguntó cuántas horas al día dedican a actividades ocio y qué tipo de actividades realizan. Con base en esta muestra se obtuvieron los siguientes resultados:

Tipo de actividad	Frecuencia absoluta	Frecuencia relativa
Jugar videojuegos	28	28 %
Ver televisión	34	34 %
Salir con amigos	15	15 %
Leer un libro	10	10 %
Dormir	7	7 %
Otro	6	6 %

   ¿Qué podría decirse de la proporción de estudiantes en la población que prefieren leer un libro? Evalúe este resultado con una confianza del 99 %.

[1]Si el método seleccionado es un método de remuestreo, pueden reconstruir los datos originales creando un vector que tenga la cantidad de ceros y unos necesarios.

Solución

1.

Partamos de que

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}} \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

Notemos que esto es una desigualdad para un valor absoluto

$$\mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}}\right| \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha,$$

$$\mathbb{P}\left(\frac{|\hat{p}_n - p|}{\sqrt{p(1-p)}} \leq \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right) \approx 1-\alpha,$$

Por comodidad, llamemos

$$c = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \geq 0,$$

Si despejamos p obtenemos un polinomio cuadrático sobre p

$$\mathbb{P}(|\hat{p}_n - p| \leq c\sqrt{p(1-p)}) = 1-\alpha \Rightarrow)\mathbb{P}((\hat{p}_n - p)^2 \leq c^2 p(1-p)) = 1-\alpha$$

$$\Rightarrow)\mathbb{P}(\hat{p}_n^2 - 2\,\hat{p}_n p + p^2 \leq c^2 p - c^2 p^2) = 1-\alpha$$

$$\Rightarrow)\mathbb{P}((1+c^2)p^2 - (2\,\hat{p}_n + c^2)p + \hat{p}_n^2 \leq 0) = 1-\alpha$$

Como $1 + c^2 > 0$ lo último se puede escribir como

$$\mathbb{P}\left(p^2 - \frac{2\,\hat{p}_n + c^2}{1+c^2}p + \frac{\hat{p}_n^2}{1+c^2} \leq 0\right) = 1-\alpha,$$

Usando la fórmula cuadrática para hallar las raíces del polinomio, tenemos que

$$p = \frac{2\,\hat{p}_n + c^2 \pm \sqrt{(2\,\hat{p}_n + c^2)^2 - 4(1+c^2)\hat{p}_n^2}}{2(1+c^2)}.$$

Notemos que

$$\begin{aligned}(2\,\hat{p}_n + c^2)^2 - 4(1+c^2)\hat{p}_n^2 &= 4\,\hat{p}_n^2 + 4c^2\,\hat{p}_n + c^4 - 4\,\hat{p}_n^2 - 4c^2\hat{p}_n^2 \\ &= c^4\end{aligned}$$

Por lo que las raíces de $p_{1,2}$ están dadas por

$$p = \frac{2\,\hat{p}_n + c^2 \pm c^2}{2(1+c^2)},$$

De esta forma, las raíces son

$$p_1 = \frac{\hat{p}_n}{1+c^2} \qquad\qquad p_2 = \frac{\hat{p}_n + c^2}{1+c^2}$$

Ahora, podemos factorizar el polinomio cuadrático como $(p-p_1)(p-p_2)$, así que para simplificar el cálculo de

$$\mathbb{P}\left(p^2 - \frac{2\hat{p}_n + c^2}{1+c^2}p + \frac{\hat{p}_n^2}{1+c^2} \leq 0\right) = 1-\alpha.$$

Basta ver cómo cambian los signos de la función $f(p) = (p-p_1)(p-p_2)$, con $0 \leq p_1 \leq p_2$ c.s.; Observando la gráfica de $f(p)$ podemos determinar cual es el intervalo donde $f(p) < 0$.

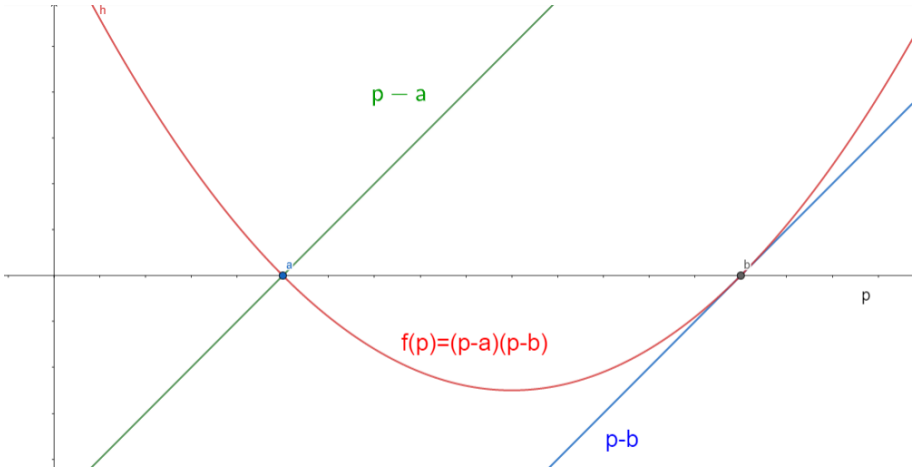


Figura 27: Gráfica de la función $f(p)$, donde $a = p_1$ y $b = p_2$.

También podríamos usar la regla del cementerio, así, obtenemos la siguiente equivalencia

$$\mathbb{P}\left(p^2 - \frac{2\hat{p}_n + c^2}{1+c^2}p + \frac{\hat{p}_n^2}{1+c^2} \leq 0\right) = 1-\alpha \iff \mathbb{P}(p_1 \leq p \leq p_2) = 1-\alpha.$$

En conclusión, para la posibilidad 1 una vez tenemos los datos, obtenemos el siguiente intervalo de confianza

$$ICA_{100(1-\alpha)\,\%}(p) = \left[\frac{\hat{p}_n}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}, \frac{\hat{p}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{n}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}\right]$$

Haciendo multiplicaciones entre fracciones, obtenemos una forma más sencilla de observar el

intervalo de confianza para la posibilidad 1

$$ICA_{100(1-\alpha)\,\%}(p) = \left[\frac{n\hat{p}_n}{n + z^2_{1-\frac{\alpha}{2}}}, \frac{n\hat{p}_n + z^2_{1-\frac{\alpha}{2}}}{n + z^2_{1-\frac{\alpha}{2}}} \right].$$

■

2.

Se adjunta el código hecho en R de este algoritmo, el código puede tardar hasta 6 horas en dar resultados para las 5 posibilidades.

3.

Observemos los resultados obtenidos con el código anterior, cada linea representa la cobertura promedio obtenida para un tamaño de muestra n. Para dar una mayor representación de lo que está ocurriendo, comparemos las coberturas y longitudes obtenidas en diferentes tamaños de muestra, para cada posibilidad.

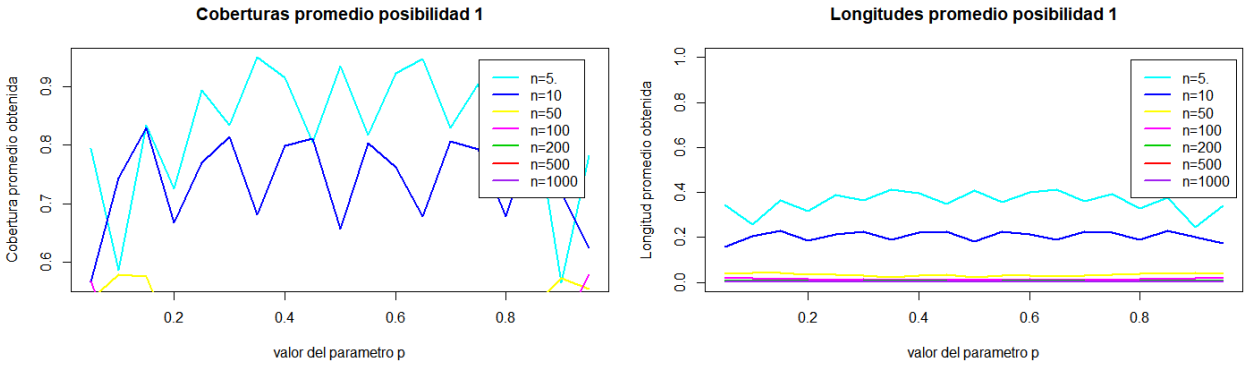


Figura 28: Cobertura y longitud promedio obtenida a partir de las simulaciones en la posibilidad 1

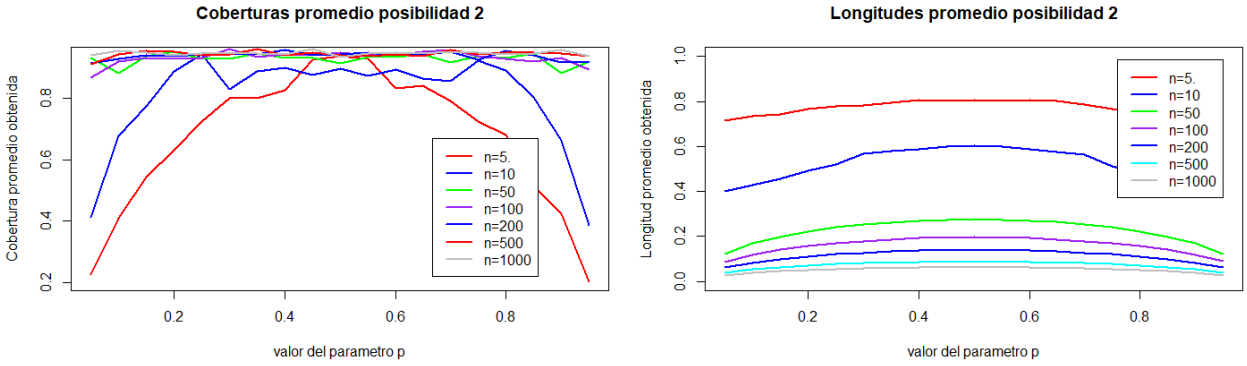


Figura 29: Cobertura y longitud promedio obtenida a partir de las simulaciones en la posibilidad 2

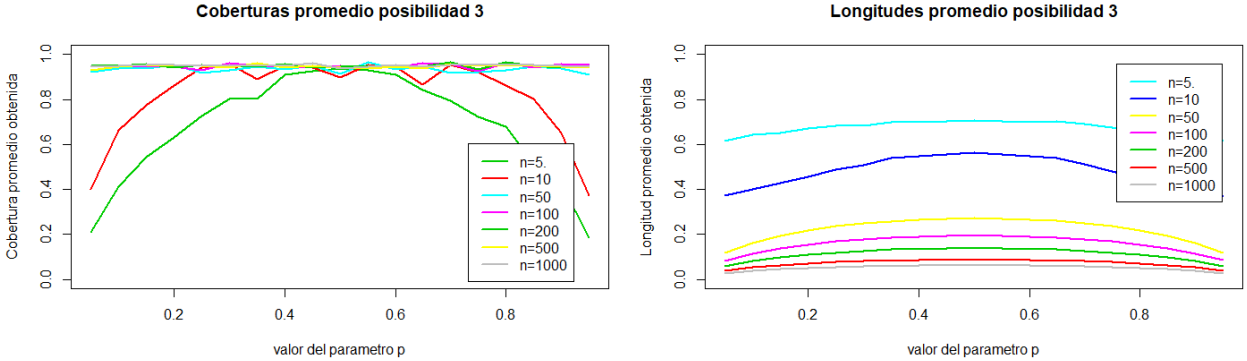


Figura 30: Cobertura y longitud promedio obtenida a partir de las simulaciones en la posibilidad 3

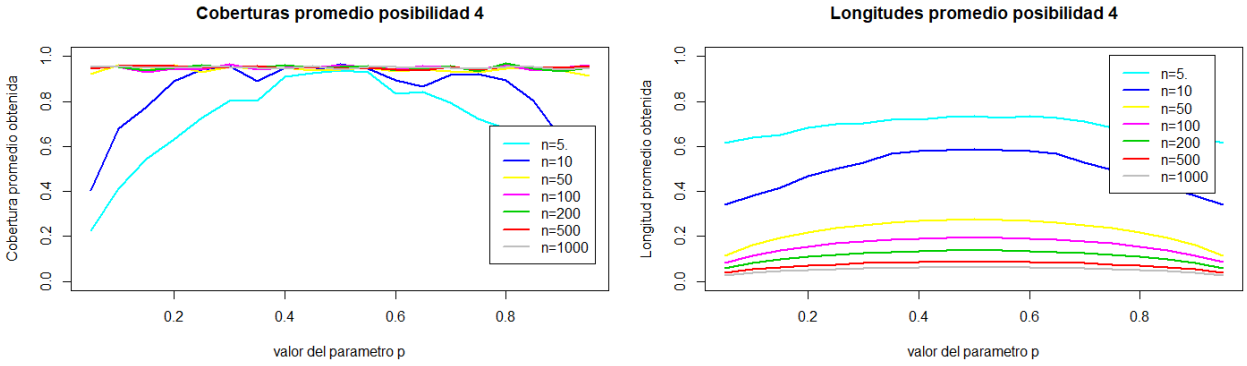


Figura 31: Cobertura y longitud promedio obtenida a partir de las simulaciones en la posibilidad 4

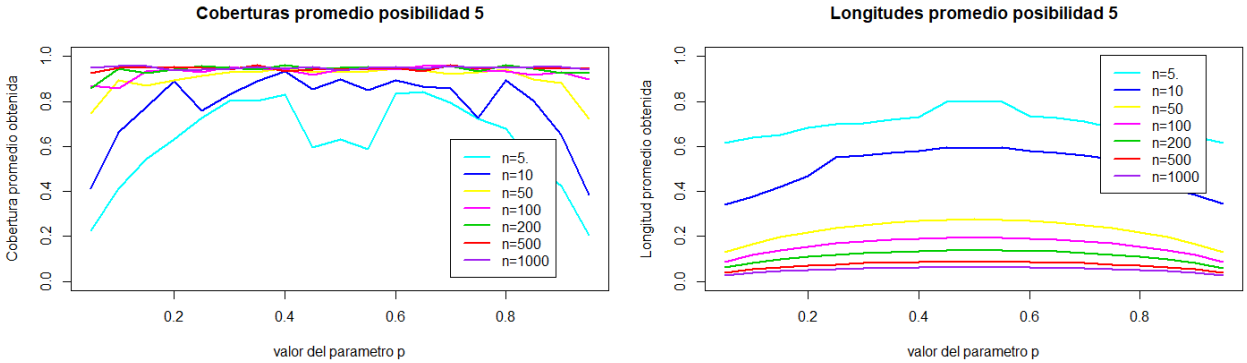


Figura 32: Cobertura y longitud promedio obtenida a partir de las simulaciones en la posibilidad 5

4.

Comparemos de nuevo las coberturas y longitudes obtenidas, esta vez las compararemos con las coberturas y longitudes obtenidas en todas las posibilidades, para un tamaño de muestra n fijo.

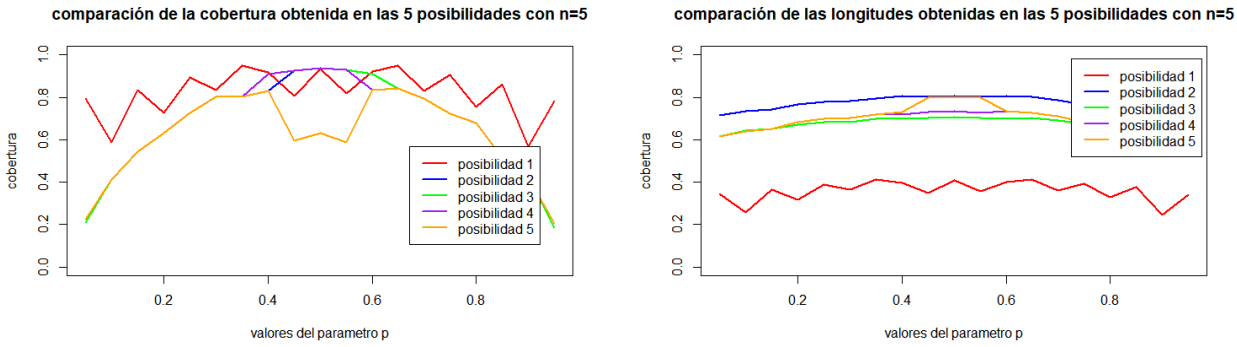


Figura 33: Comparación de las coberturas y longitudes obtenidas en las 5 posibilidades con un tamaño de muestra $n = 5$

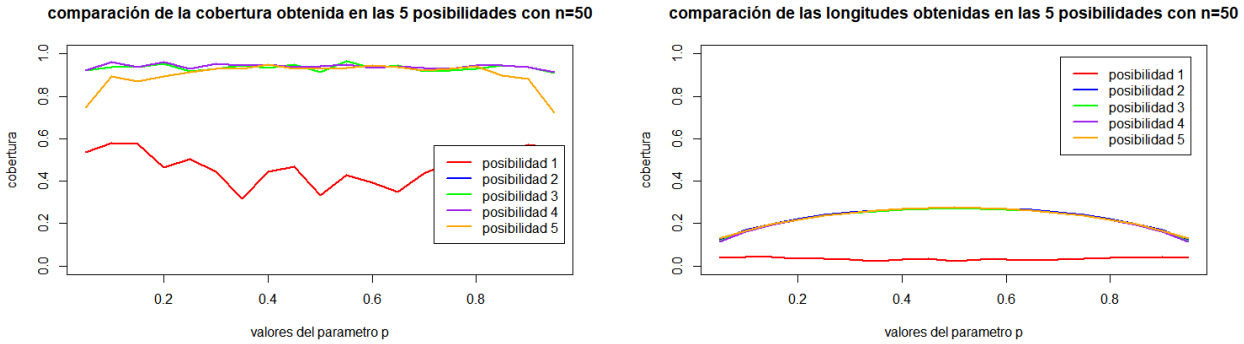


Figura 34: Comparación de las coberturas y longitudes obtenidas en las 5 posibilidades con un tamaño de muestra $n = 50$

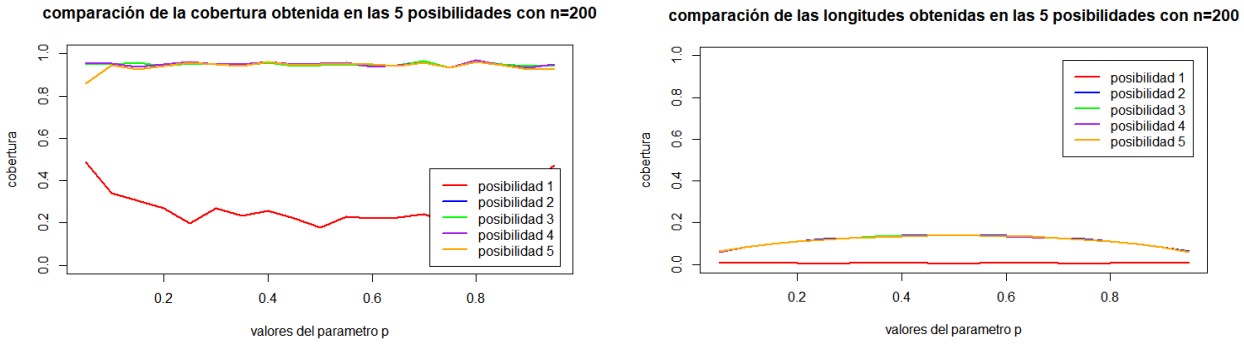


Figura 35: Comparación de las coberturas y longitudes obtenidas en las 5 posibilidades con un tamaño de muestra $n = 200$

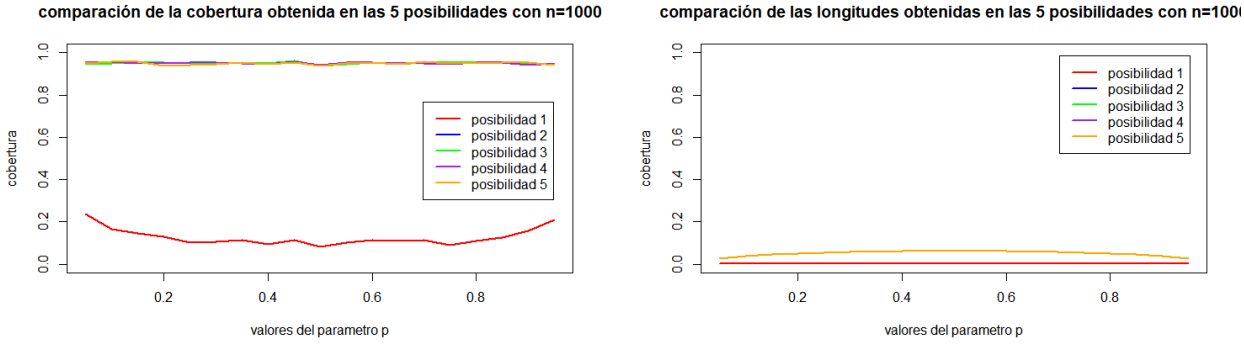


Figura 36: Comparación de las coberturas y longitudes obtenidas en las 5 posibilidades con un tamaño de muestra $n = 1000$

5.

A partir de los gráficos anteriores podemos observar el comportamiento de cada una de las posibilidades. En la posibilidad 1, si observamos las gráficas dadas en 28, vemos que la cobertura promedio (es decir, la cantidad de intervalos que contuvieron al parámetro p) es mayor para tamaños de muestra pequeños, a medida que aumenta el tamaño de muestra, la cobertura promedio disminuye. Esto nos indica que la posibilidad 1 es más efectiva que las demás posibilidades para tamaños de muestra pequeños.
Sin embargo, cuando n es muy grande, los limites del intervalo de confianza de p son valores muy cercanos a p (pero p no se encuentra en el intervalo), esto lo podemos deducir ya que el intervalo de confianza es obtenido a partir del TCL y que la longitud promedio de los intervalos que contuvieron a p disminuye a medida que el tamaño de muestra n aumenta.

Las coberturas promedio tienen un comportamiento parabólico en las posibilidades $2-5$ respecto al parámetro p , lo que nos muestra que la cobertura disminuye cuando $p \longrightarrow 0$ o $p \longrightarrow 1$ (a medida que p se aleja del foco de la parábola). Como el comportamiento de p es parabólico, podemos pensar que alcanza la mayor cobertura posible cuando $p \approx 0.5$, esto es cierto en las posibilidades $3, 4, 2$, siendo esta última posibilidad la que brinda una mayor cobertura para $p = 0.5$, sin embargo, en la posibilidad 5 la cobertura promedio disminuye para valores de p cercanos a 0.5

Como podemos observar en los gráficos de los ejercicios 3 y 4, cuando el tamaño de muestra n es muy grande, la cobertura promedio es aproximadamente 1 en las posibilidades $2-5$, lo que significa que cuando n aumenta, aumentan la cantidad de intervalos que contuvieron al parámetro p. También, al observar las longitudes, podemos observar que la longitud promedio de los intervalos que contuvieron al parámetro p es cada vez mas pequeña, por lo que los intervalos aumentan su exactitud.

6.

1. Usamos la posibilidad 4, obteniendo un histograma para las estimaciones de la muestra Bootstrap de la proporción como el de la Figura 37.

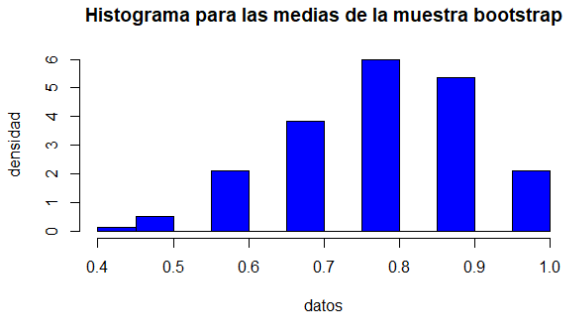


Figura 37: Histograma de las estimaciones Bootstrap de la proporción.

Con ayuda de R, código que se encuentra adjunto a este proyecto, obtenemos un intervalo de confianza para la proporción, del 90 % de confianza, centrado en la estimación puntual obtenida de la muestra original. En este caso

$$ICBoot_{100(1-\alpha)\,\%}(p) = [0.6, 1],$$

es decir que si se repite el muestreo un número suficientemente grande de veces, el 90 % de las veces, la proporción de componentes que cumplen con los estándares de producción estará entre el 60 % y el 100 %.

2. Como el tamaño de la muestra es grande, podemos usar cualquiera de las posibilidades relacionadas al uso del TCL; en particular, al usar la posibilidad 2, según los datos dados tenemos que $\overline{p}_n = 0.1$ y con una confianza del 99 %, $\alpha = 0.01$, así que $1 - \alpha/2 = 0.995$ y $z_{1-\alpha/2} = 2.58$. Por tanto, usando la posibilidad 2,

$$ICA_{100(1-\alpha)\,\%}(p) = 0.1 \mp 2.58\frac{\sqrt{0.1 \cdot 0.9}}{10}$$

$$ICA_{100(1-\alpha)\,\%}(p) = [0.0226, 0.1774],$$

es decir que si se repite el muestreo un número suficientemente grande de veces, el 99 % de las veces, la proporción de estudiantes de la UNAL-sede Bogotá que leen un libro en su tiempo de ocio esta entre el 2 % y el 18 % aproximadamente.