UNIVERSIDAD NACIONAL DE COLOMBIA

Proyecto Final
Inferencia Estadística
Profesor: Mario E. Arrieta Prieto

Pablo Gonzalez Baron pgonzalezb@unal.edu.co
Beimar Jose Naranjo Morales bnaranjom@unal.edu.co
Juan Diego Murcia Porras jmurciap@unal.edu.co
Carlos Enrique Nosa Guzmán cnosa@unal.edu.co

Parte A

Distribución de un estimador máximo-verosímil cuando no hay condiciones de regularidad

Consideramos una muestra aleatoria $X_1, ..., X_n$ de una distribución $U[0, \theta = 3]$, sabiendo que $\hat{\theta}_{MLE} = X^{(n)}$.

Definimos una función en R que nos permite generar m simulaciones de una muestra aleatoria de tamaño n que sigue una distribución $U[0, \theta = 3]$.
En esta función, para cada muestra almacenamos la estimación maximo-verosimil que ya sabemos que es $X^{(n)}$, es decir el valor máximo de la muestra de tamaño n, y posteriormente generamos un histograma de todas las estimaciones MLE sobrelapando la función de densidad teorica de $X^{(n)}$ que la calculamos como:

$$f_{X^{(n)}}(x) = nf_X(x) \cdot [F_X(x)]^{n-1} = \frac{n}{3} \cdot \left(\frac{x}{3}\right)^{n-1} = \frac{nx^{n-1}}{3^n}$$

a). Ahora, para el primer literal ejecutamos la función con los parámetros $m = 1000,\ n = 10$ y obtenemos la siguiente gráfica:

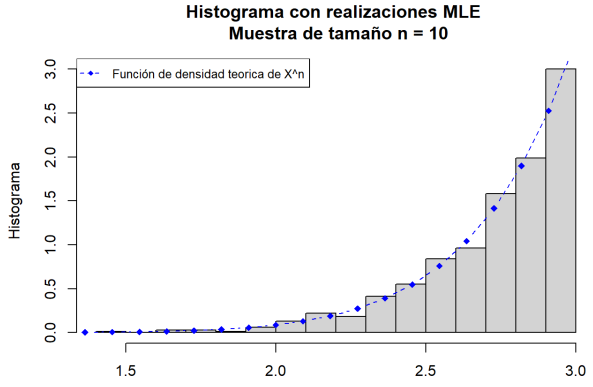
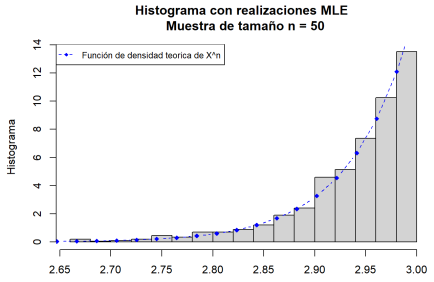
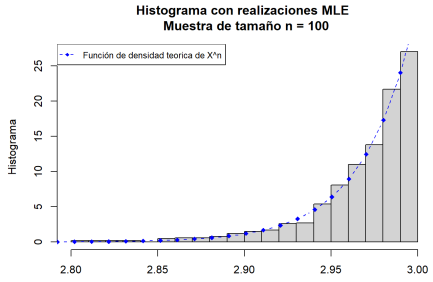


Figura 1: Histograma de las realizaciones del MLE y función de densidad teorica de $X^{(n)}$ con $n = 10$

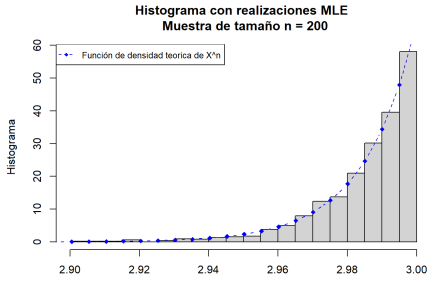
b). Para el segundo literal, ejecutamos la función nuevamente variando el tamaño de muestra con los valores $n = 50, 100, 200, 500, 1000$ y obtenemos la siguiente serie de gráficas:



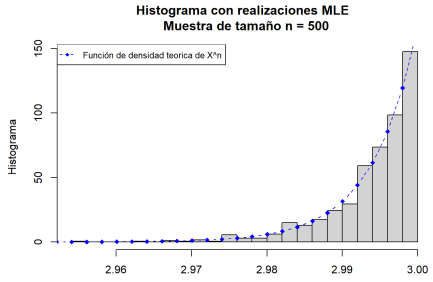
(a) Histograma de las realizaciones del MLE y función de densidad teorica de $X^{(n)}$ con $n = 50$



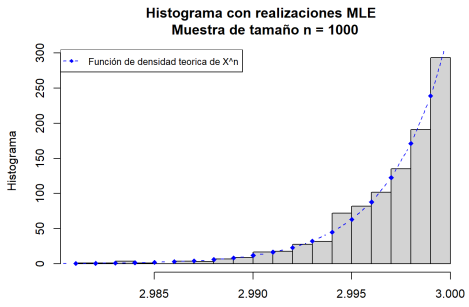
(b) Histograma de las realizaciones MLE y función de densidad teorica de $X^{(n)}$ con $n = 100$



(c) Histograma de las realizaciones del MLE y función de densidad teorica de $X^{(n)}$ con $n = 200$



(d) Histograma de las realizaciones del MLE y función de densidad teorica de $X^{(n)}$ con $n = 500$



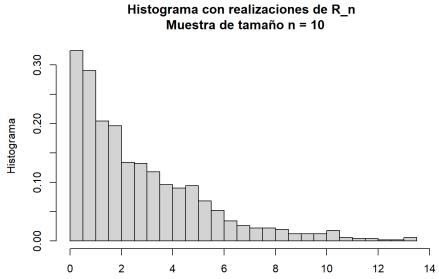
(e) Histograma de las realizaciones del MLE y función de densidad teorica de $X^{(n)}$ con $n = 1000$

c). Con las graficas anteriores es posible ver que a medida que el tamaño de la muestra aumenta, las realizaciones del estimador máximo-verosímil se concentran en *3.0*, y además la curva de la función de densidad teorica de $X^{(n)}$ se acopla al histograma.

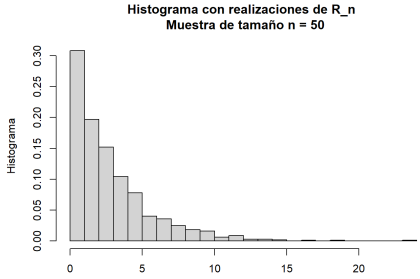
d). Ahora para este literal, definimos una función en R que nos permite generar m muestras aleatorias de tamaño n, y para cada una de estas muestras deseamos almacenar las realizaciones de R_n que viene definida como sigue:

$$R_n = n(3 - X^{(n)})$$

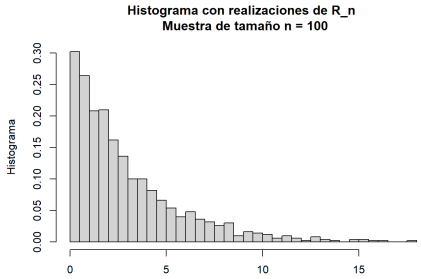
Y posteriormente realizar histogramas de los valores obtenidos para cada n. Por lo tanto, habiendo definido la función, la ejecutamos con los parámetros $m = 1000$ y $n = 10, 50, 100, 200, 500$ y 1000, obteniendo las siguientes gráficas:



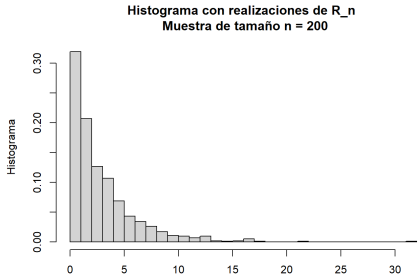
(f) Histograma de las realizaciones de R_n con $n = 10$



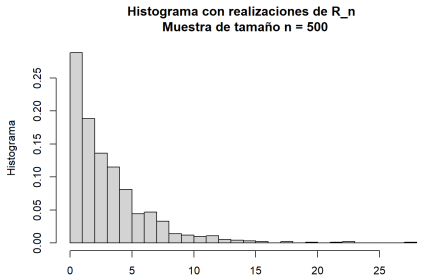
(g) Histograma de las realizaciones de R_n con $n = 50$



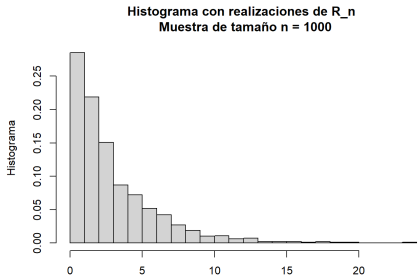
(h) Histograma de las realizaciones de R_n con $n = 100$



(i) Histograma de las realizaciones de R_n con $n = 200$



(j) Histograma de las realizaciones de R_n con $n = 500$



(k) Histograma de las realizaciones de R_n con $n = 1000$

Podemos ver que a medida que n aumenta, el histograma se asemeja a una **distribución exponencial** con parámetro $\lambda = 1/3$.

e). Habiendo conjeturado lo anterior, procedemos a demostrar lo siguiente:

$$R_n = n(3 - X^{(n)}) \xrightarrow{d} Exp(1/3)$$

Demostración. Habiendo calculado al principio la función de densidad teórica de $X^{(n)}$, tenemos que la función de distribución para $X^{(n)}$ es:

$$F_{X^{(n)}}(x) = \int_0^x f_{X^{(n)}}(u)du = \int_0^x \frac{nu^{n-1}}{3^n}du = \left(\frac{x}{3}\right)^n$$

Entonces, calculando la función de distribución de R_n:

$$F_{R_n}(x) = p\Big(R_n \leq x\Big) = p\Big(n(3 - X^{(n)}) \leq x\Big) = p\Big(X^{(n)} \geq 3 - \frac{x}{n}\Big) = 1 - F_{X^{(n)}}\Big(3 - \frac{x}{n}\Big)$$

Calculando la función de distribución de $X^{(n)}$ en $3 - x/n$:

$$F_{X^{(n)}}\Big(3 - \frac{x}{n}\Big) = \left(\frac{3 - \frac{x}{n}}{3}\right)^n$$

Por lo tanto tenemos que la función de distribución de R_n viene dada por:

$$F_{R_n}(x) = 1 - \left(\frac{3 - \frac{x}{n}}{3}\right)^n$$

Entonces hallando el límite de la expresión anterior cuando n tiende a infinito:

$$\lim_{n\to\infty} 1 - \left(\frac{3 - \frac{x}{n}}{3}\right)^n = 1 - \lim_{n\to\infty}\left(\frac{3 - \frac{x}{n}}{3}\right)^n = 1 - \lim_{n\to\infty}\left(\frac{1}{3}\cdot\left(3 - \frac{x}{n}\right)\right)^n = 1 - e^{-x/3}$$

Por lo tanto tenemos que:

$$\lim_{n\to\infty} F_{R_n}(x) = 1 - e^{-x/3}$$

Ahora, como la función de distribución de un modelo exponencial con parámetro λ es:

$$F_X(x) = 1 - e^{-\lambda x}, \qquad x \geq 0$$

Entonces podemos concluir que la variable aleatoria $R_n = n(3 - X^{(n)})$ converge en distribución a un modelo exponencial con parámetro $\lambda = 1/3$. □

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

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con distribución *Cauchy*($\theta = \mathbf{3}$,1)

a). Los siguientes gráficos corresponden a los histogramas del valor aproximado de las estimaciones por MLE y de las medianas de las simulaciones.

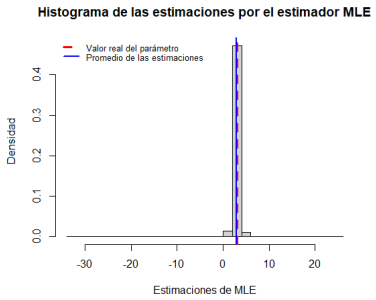


Figura 2: MLE con un tamaño de muestra $n = 10$

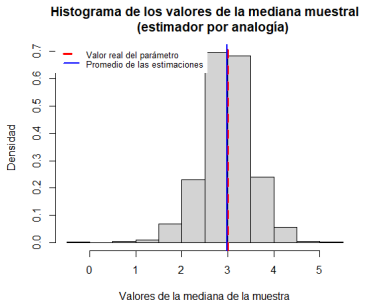


Figura 3: M_e con un tamaño de muestra $n = 10$

El promedio y la varianza respectivos a cada uno de los estimadores se muestra a continuación

	Promedio	Varianza
$\widehat{\theta}_{MLE}$	3,008810	0,646882
M_e	2,978649	0,339135

Según lo observado en las anteriores gráficas, las lineas correspondientes al valor real del parámetro y al promedio de las simulaciones de los dos estimadores parecen ser próximas, esto también lo confirma el valor promedio de los estimadores mostrados en la tabla anterior, por ende en este caso es válido afirmar que los dos estimadores parecen ser insesgados.

Según las simulaciones, encontremos cuál de estos dos estimadores resulta ser mejor en términos de error cuadrático medio (MSE); en primer lugar se tiene que

	Sesgo al cuadrado	Varianza	MSE
$\widehat{\theta}_{MLE}$	0,00007762482	0,646882	0,646904
M_e	0,0004558359	0,339135	0,331369

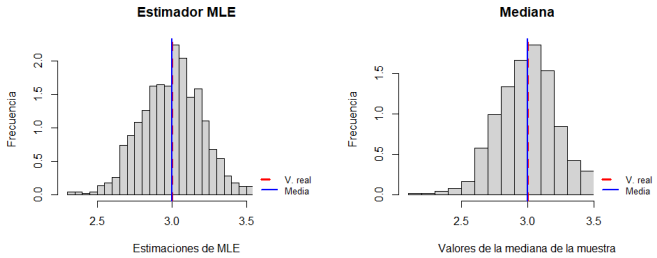
por lo tanto, la eficiencia relativa es

$$RE(\widehat{\theta}_{MLE}, M_e) = \frac{MSE(\widehat{\theta}_{MLE}, \theta)}{MSE(M_e, \theta)} = \frac{0{,}646904}{0{,}331369} = 1{,}952384 > 1$$

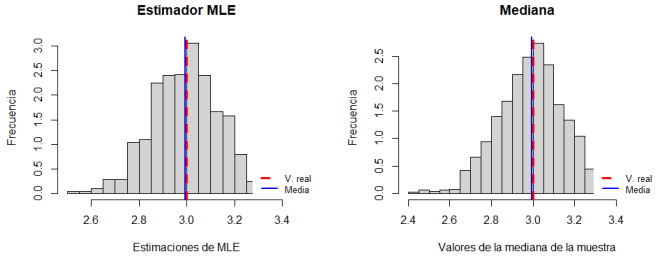
de lo anterior se deduce que la mediana es más eficiente que el estimador MLE para esta simulación.[1]

b). Repitiendo el proceso del literal **a** con tamaños de muestra $n = 50, 100, 200, 500, 1000$ obtenemos los siguientes histogramas

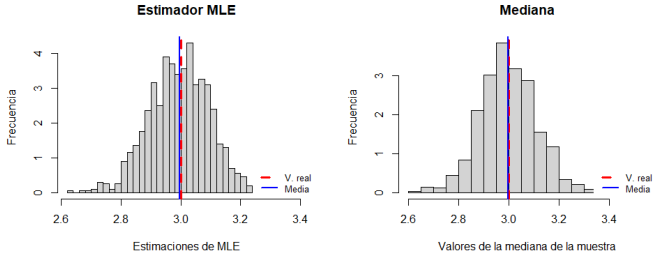
[1] Obsérvese que otra manera de poder asumir que los estimadores son insesgados es el sesgo cercano a 0.



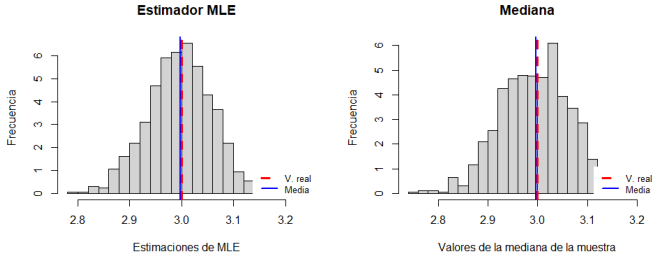
(a) Tamaño de muestra $n = 50$



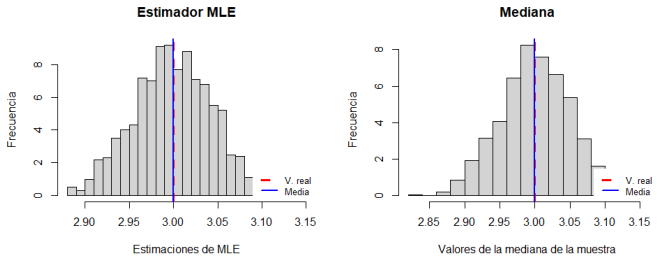
(b) Tamaño de muestra $n = 100$



(c) Tamaño de muestra $n = 200$



(d) Tamaño de muestra $n = 500$



(e) Tamaño de muestra $n = 1000$

Los promedios y varianzas registradas para cada uno de los tamaños de muestra se resume en la siguiente tabla

Estimador	n	Promedio	Varianza
$\widehat{\theta}_{MLE}$	50	3.012441	0.04328238
	100	3.001057	0.0204716
	200	2.995537	0.01072755
	500	3.003593	0.004352049
	1000	3.001595	0.001980125
M_e	50	3.00985	0.05114264
	100	3.001976	0.02570468
	200	2.996428	0.01336497
	500	3.004504	0.00525236
	1000	3.002046	0.002397749

el error cuadrático medio en cada uno de los casos es

Estimador	n	Sesgo al cuadrado	Varianza	MSE
$\widehat{\theta}_{MLE}$	50	0.0001547893	0.04328238	0.04343717
	100	0.00000111733	0.0204716	0.02047272
	200	0.00001991479	0.01072755	0.01074746
	500	0.00001291246	0.004352049	0.004364962
	1000	0.000002542533	0.001980125	0.001982667
M_e	50	0.00009703215	0.05114264	0.05123967
	100	0.000003906462	0.02570468	0.02570859
	200	0.00001275931	0.01336497	0.01337773
	500	0.00002028806	0.00525236	0.005272648
	1000	0.000004186437	0.002397749	0.002401935

de esta manera, la eficiencia relativa en cada caso corresponde a

n	$RE(\widehat{\theta}_{MLE}, M_e)$
50	0.8477253
100	0.7963378
200	0.8033845
500	0.82785
1000	0.8254458

Para las simulaciones realizadas variando el tamaño de muestra se puede notar que el error cuadrático medio es menor a uno en todos los casos, esto significa que el estimador máximo verosímil es mejor en términos de error cuadrático medio, contrario a lo que se mostró en el anterior ítem cuando el tamaño de muestra solo era de diez variables aleatorias.

c). Considere las variables aleatorias definidas como

$$V_n = \sqrt{n}(\widehat{\theta}_{MLE} - \theta) \qquad W_n = \sqrt{n}(M_e - \theta)$$

para todo $n \in \mathbb{N}_{\geq 1}$.

- Dado que una muestra aleatoria de una población con distribución *Cauchy* cumple las condiciones de regularidad[2] puesto que el soporte de ésta distribución corresponde al conjunto de los números reales, es válido aplicar el teorema que indica el comportamiento asintótico del estimador máximo verosímil, así

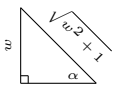
$$V_n = \sqrt{n}(\widehat{\theta}_{MLE} - \theta) \xrightarrow[n\to\infty]{d} N(0, (I(\theta))^{-1})$$

donde $I(\theta)$ es la información de Fisher del parámetro θ.

[2]Asumimos que se tiene un caso regular de estimación cuando el soporte de una variable aleatoria no depende del parámetro desconocido.

Obsérvese que

$$
\begin{aligned}
I(\theta) &:= E\left[\left(\frac{\partial}{\partial\theta}\ln f_X(X,\theta)\right)^2\right] \\
&= E\left[\left(\frac{\partial}{\partial\theta}\ln\left(\frac{1}{\pi(1+(X-\theta)^2)}\right)\right)^2\right] \\
&= E\left[\frac{\partial}{\partial\theta}(-\ln\pi-\ln(1+(X-\theta)^2))\right] \\
&= E\left[\left(\frac{2(X-\theta)}{1+(X-\theta)^2}\right)^2\right] \\
&= E\left[\frac{4(X-\theta)^2}{(1+(X-\theta)^2)^2}\right] \\
&= \int_{-\infty}^{\infty}\frac{4(x-\theta)^2}{(1+(x-\theta)^2)^2}\frac{1}{\pi(1+(x-\theta)^2)}\mathrm{d}x \\
&= \frac{4}{\pi}\int_{-\infty}^{\infty}\frac{(x-\theta)^2}{(1+(x-\theta)^2)^3}\mathrm{d}x \\
&= \frac{4}{\pi}\int_{-\infty}^{\infty}\frac{w^2}{(1+w^2)^3}\mathrm{d}w \qquad \left\{\begin{array}{rcl} w & = & x-\theta \\ \mathrm{d}w & = & \mathrm{d}\theta\end{array}\right. \\
&= \frac{4}{\pi}\int_{-\infty}^{\infty}\frac{1+w^2-1}{(1+w^2)^3}\mathrm{d}w \\
&= \frac{4}{\pi}\int_{-\infty}^{\infty}\left(\frac{1}{(1+w^2)^2}-\frac{1}{(1+w^2)^3}\right)\mathrm{d}w
\end{aligned}
$$



$$
\begin{aligned}
&= \frac{4}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\left(\frac{1}{\sec^4(\alpha)}-\frac{1}{\sec^6(\alpha)}\right)\sec^2(\alpha)\mathrm{d}\alpha \qquad \left\{\begin{array}{rcl} \tan(\alpha) & = & w \\ \sec^2(\alpha)\mathrm{d}\alpha & = & \mathrm{d}w\end{array}\right. \\
&= \frac{4}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\left(\frac{1}{\sec^2(\alpha)}-\frac{1}{\sec^4(\alpha)}\right)\mathrm{d}\alpha \\
&= \frac{4}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}(\cos^2(\alpha)-\cos^4(\alpha))\mathrm{d}\alpha \\
&= \frac{4}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos^2(\alpha)(1-\cos^2(\alpha))\mathrm{d}\alpha \\
&= \frac{4}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}(\sin(\alpha)\cos(\alpha))^2\mathrm{d}\alpha \\
&= \frac{4}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\left(\frac{\sin(2\alpha)}{2}\right)^2\mathrm{d}\alpha \\
&= \frac{1}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\sin^2(2\alpha)\mathrm{d}\alpha \\
&= \frac{1}{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}(1-\cos(4\alpha))\mathrm{d}\alpha \\
&= \frac{1}{2\pi}\left(\alpha-\frac{\sin(4\alpha)}{4}\right)_{\alpha=-\frac{\pi}{2}}^{\alpha=\frac{\pi}{2}} \\
&= \frac{1}{2\pi}(\pi) \\
&= \frac{1}{2}
\end{aligned}
$$

Finalmente se tiene que

$$V_n = \sqrt{n}(\widehat{\theta}_{MLE} - \theta) \xrightarrow[n\to\infty]{d} N(0,2)$$

- Para calcular el comportamiento asintótico de la secuencia $\{W_n\}_{n\in\mathbb{N}_{\geq 1}}$, en primer lugar nótese que

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \frac{1}{\pi}\frac{1}{1+(t-\theta)^2}\mathrm{d}t \\ &= \frac{1}{\pi}(\arctan(t-\theta))_{t\to-\infty}^{t=x} \\ &= \frac{1}{\pi}\left(\arctan(x-\theta) - \frac{-\pi}{2}\right) \\ &= \frac{1}{\pi}\arctan(x-\theta) + \frac{1}{2} \end{aligned}$$

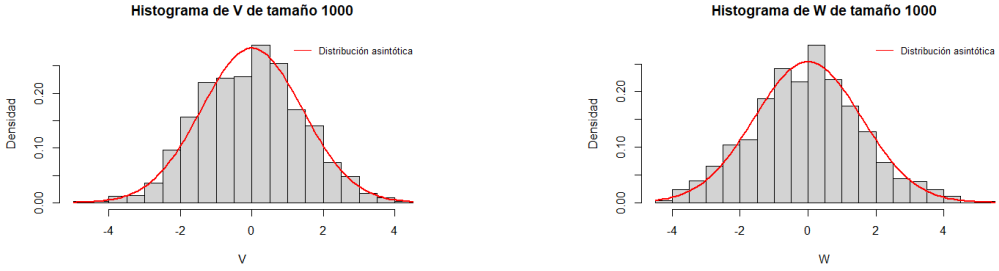
de esta forma,

$$\begin{aligned} F_X(x) = 0{,}5 &\Longrightarrow \frac{1}{\pi}\arctan(x-\theta) + \frac{1}{2} = 0{,}5 \\ &\Longrightarrow \frac{1}{\pi}\arctan(x-\theta) = 0 \\ &\Longrightarrow \arctan(x-\theta) = 0 \\ &\Longrightarrow x - \theta = 0 \\ &\Longrightarrow x = \theta \end{aligned}$$

es decir, θ corresponde a la mediana poblacional. Dado que f_X es positiva y continua en un vecindario alrededor de θ es factible aplicar el teorema de la distribución asintótica de una estadística de orden, el cual establece que

$$W_n = \sqrt{n}(M_e - \theta) \xrightarrow[n\to\infty]{d} N\left(0, \frac{0{,}5(1-0{,}5)}{f_X^2(\theta)}\right) = N\left(0, \frac{\pi^2}{4}\right)$$

d). A continuación se presentan los dos histogramas correspondientes a las $m = 1000$ realizaciones de las variables aleatorias V_{1000} y W_{1000}



Las anteriores gráficas sugieren que hay un ajuste de los datos a la distribución asíntotica de ambas secuencias de variables aleatorias.

Por otra parte, teniendo en cuenta el anterior ítem,

$$ARE(V,W) = \frac{2}{\frac{\pi^2}{4}} = \frac{8}{\pi^2} \approx 0{,}81 < 1$$

por lo tanto el estimador máximo verosímil estandarizado es más eficiente asintóticamente que la mediana estandarizada.

e). **Bonus**: Para comparar los estimadores máximo verosímil, mediana y moda[3] del parámetro θ observemos la siguiente tabla

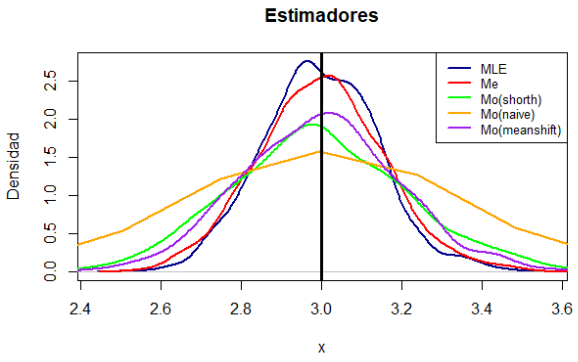
[3]La moda fue calculada con tres métodos diferentes proporcionados por R.

Estimador	n	Promedio	Varianza	MSE
$\widehat{\theta}_{MLE}$	50	2.996014	0.04223191	0.04224780
	100	2.998172	0.02026824	0.02027158
	200	2.996917	0.009830832	0.00984034
	500	3.002189	4.038119e-03	4.042913e-03
	1000	3.000063	0.001907256	0.001907260
M_e	50	2.998450	0.05076466	0.05076707
	100	2.999015	0.02575419	0.02575516
	200	2.993535	0.012036991	0.01207879
	500	3.003359	4.843179e-03	4.854459e-03
	1000	2.999616	0.002447371	0.002447519
M_o (shorth)	50	2.993396	0.09203550	0.09207912
	100	2.994204	0.04335340	0.04338699
	200	3.000168	0.027597079	0.02759711
	500	3.003123	1.290002e-02	1.290977e-02
	1000	2.997869	0.007227213	0.007231755
M_o (naive)	50	2.981625	8.21203391	8.21237157
	100	3.010676	4.35397400	4.35408798
	200	3.011642	4.665814474	4.66595000
	500	3.262144	1.338212e+02	1.338900e+02
	1000	3.197926	73.552473827	73.591648509
M_o (meanshift)	50	2.990943	0.06935517	0.06943720
	100	2.994649	0.03679206	0.03682069
	200	2.997392	0.022262845	0.02226965
	500	3.003083	1.078577e-02	1.079528e-02
	1000	2.997294	0.007386058	0.007393380

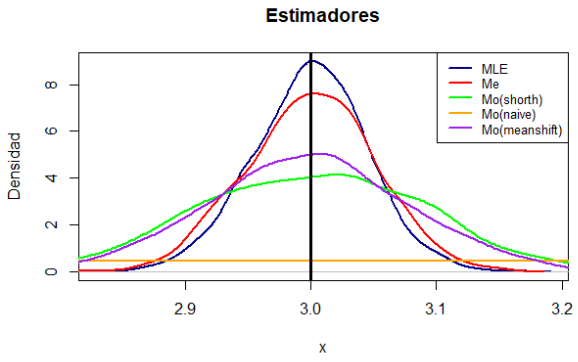
Estos datos sugieren que

- Los tres estimadores son insesgados: El promedio está 'cerca' del valor real del parámetro y cuando la muestra aumenta la varianza tiende a ser igual a cero.[4]
- En términos de error cuadrático medio el estimador máximo verosímil resulta ser el más eficiente comparado con la moda y la mediana de la muestra aleatoria.

La siguientes gráficas muestran la densidad aproximada de cada uno de los estimadores y apoyan los enunciados dichos anteriormente



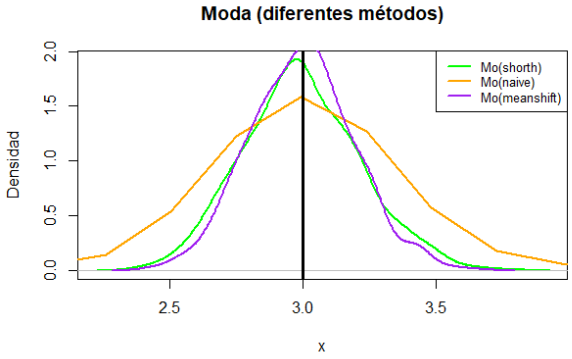
(a) Tamaño de muestra $n = 100$



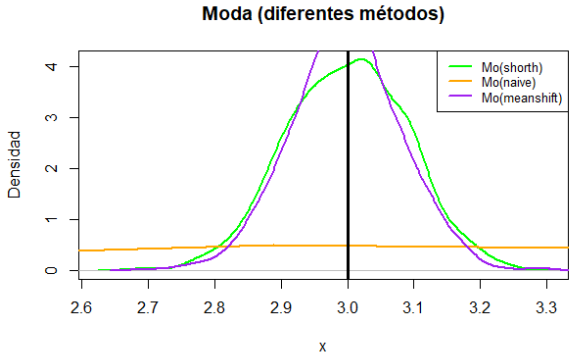
(b) Tamaño de muestra $n = 1000$

Por otra parte, respecto a los métodos para calcular la moda de un conjunto de datos (realizaciones) con ayuda del paquete `modeest` en *R* se escogió tres métodos dentro de los doce posibles en la función `mlv`, estos son *shorth, naive* y *meanshift*.

[4]Excluyase de esta afirmación el método *naive* para calcular la moda.



(a) Tamaño de muestra $n = 100$



(b) Tamaño de muestra $n = 1000$

A continuación, se presenta un breve descripción[5] de cada uno de los métodos para el cálculo de esta estadística

- Método *shorth*: Este método hace uso del estimador LMS (Least Median of Squares) para la moda el cuál pertenece a los procesos de 'mode-seeking'[6]. Según lo observado en la tabla y las gráficas anteriores este método arroja valores óptimos para la estimación de la moda.
- Método *naive*: Este método hace uso del estimador Chernoff de la moda el cuál se define como el centro del intervalo que contiene la mayor cantidad de datos posibles dada una longitud fija. Según lo observado en las gráficas anteriores, este estimador sufre de problemas de estabilidad pues, su varianza no tiende a cero ni se reduce cuando el tamaño de la muestra aumenta.
- Método *meanshift*: La estimación de la moda por medio de éste método hace uso de un algoritmo iterativo que cambia cada dato por el promedio de un vecindario alrededor de ese dato dado un valor aproximado inicial del valor que se desea hallar.[7] Comparado con los anteriores métodos, este parece tener una menor varianza respecto a las estimaciones de la moda.

[5]La información mencionada aquí puede ser consultada principalmente en *Poncet, P. (2019). Modeest: Mode estimation (2.4.0) [Computer software]. https://CRAN.R-project.org/package=modeest* y en *Modeest documentation. (n.d.). Recuperado el ¿cuándo? del 2022, desde https://rdrr.io/cran/modeest/man/*.

[6]Para más información acerca de cómo funciona este método consulte *Meer, P. (1991). Robust Regression Methods for Computer Vision: A Review. International Journal of Computer Vision. Recuperado el ¿cuándo? de 2022 desde https://sites.rutgers.edu/peter-meer/wp-content/uploads/sites/69/2018/12/meerrob91.pdf*

[7]Información tomada de *Yizong Cheng. (1995). Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(8), 790–799. https://doi.org/10.1109/34.400568*.

Comparación del estimador ML y de momentos en un modelo Doble Exponencial

a). A continuación[8] se presentan los histogramas de las m estimaciones máximo verosímil y por momentos:

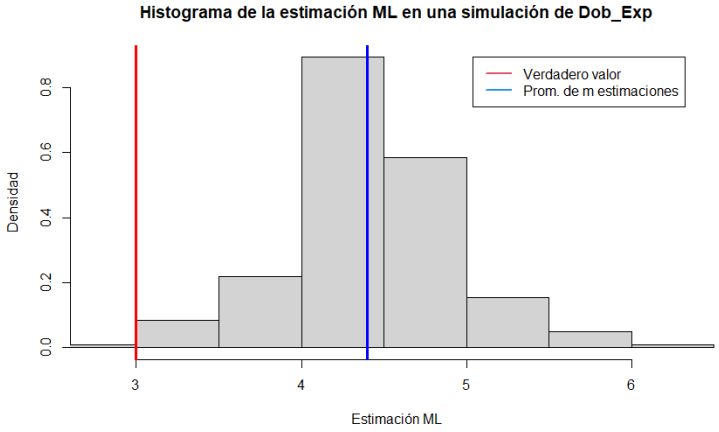


Figura 4: ML con n=10

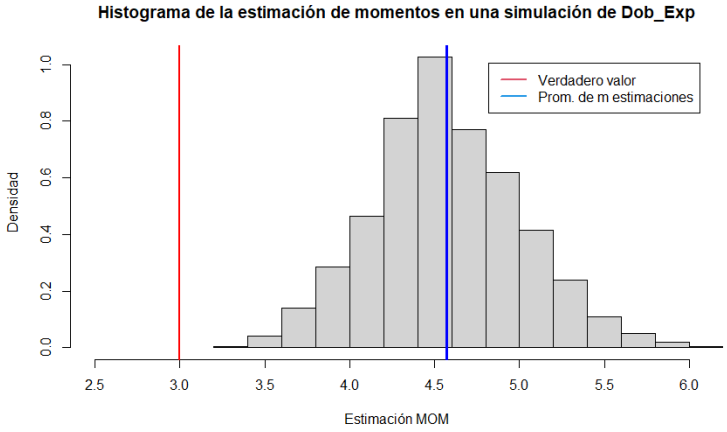


Figura 5: MOM con n=10

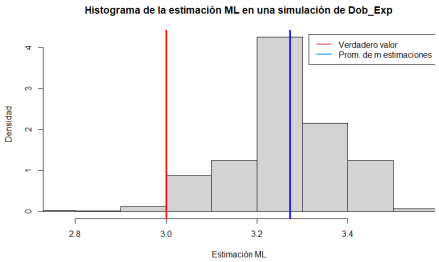
observando las anteriores gráficas no se puede afirmar que a simple vista los estimadores sean insesgados pues como se puede ver el verdadero valor del parámetro (linea roja) está alejado del promedio muestral (linea azul), En la siguiente tabla se registran los valores del promedio y la varianza obtenida de cada uno de los estimadores:

	Varianza	Promedio
MLE	0.3034418	4.397963
MOM	0.1969085	4.571380

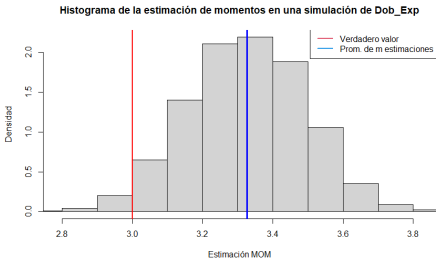
Al calcular la estimación del sesgo del estimador ML (simulado) obtenemos aproximadamente el valor $1{,}39796$ y para el estimador por momentos se obtiene $1{,}57138$ con lo que sería más conveniente afirmar en este caso que no hay insesgamiento. El error cuadrático medio estimado del estimador ML (simulado) es aproximadamente $(4{,}397963-3)^2+0{,}3034418 \approx 2{,}2577$ y el del estimador por momentos $(4{,}571380-3)^2+0{,}1969085 \approx 2{,}666144$ y por tanto a pesar de que las estimaciones por momentos tienen menor varianza, las estimaciones ML resultan mejores en términos del error cuadrático medio.

b). A continuación se presentan los histogramas de las m estimaciones ML y por momentos

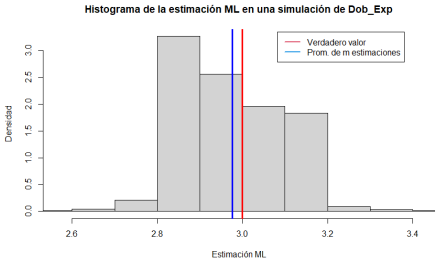
[8] Todos los resultados numéricos aquí presentados están previamente calculados en el script ???.



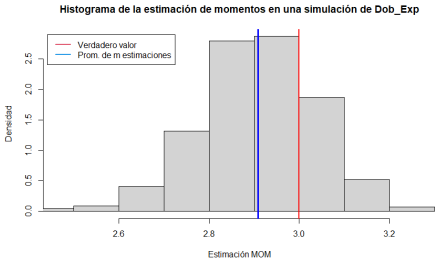
(a) MLE con n=50



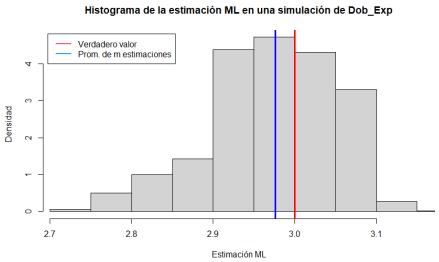
(b) MOM con n=50



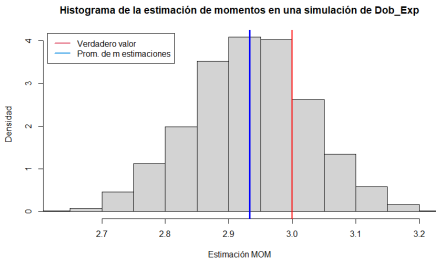
(c) MLE con n=100



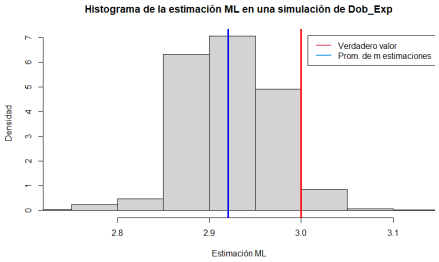
(d) MOM con n=100



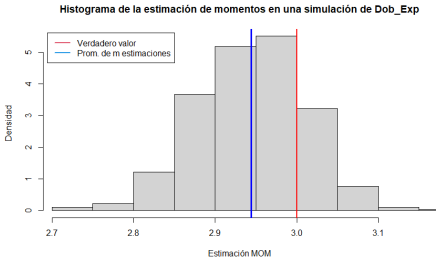
(e) MLE con n=200



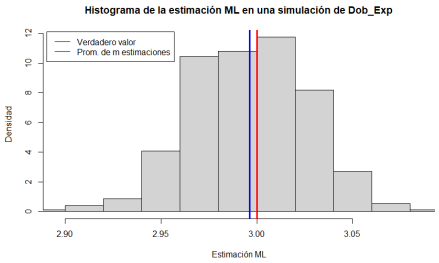
(f) MOM con n=200



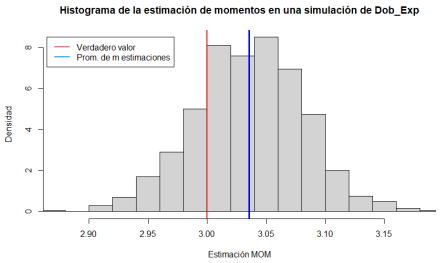
(g) MLE con n=500



(h) MOM con n=500



(i) MLE con n=1000



(j) MOM con n=1000

y se registran los valores de promedio y varianza de las m estimaciones con diferentes tamaños de muestra:

n=50	Varianza	Promedio
MLE	0.01326942	3.272249
MOM	0.02719302	3.327106

n=100	Varianza	Promedio
MLE	0.01387023	2.976884
MOM	0.01647585	2.908454

n=200	Varianza	Promedio
MLE	0.005740612	2.976002
MOM	0.008787678	2.932594

n=500	Varianza	Promedio
MLE	0.002315869	2.920505
MOM	0.004371080	2.944375

n=1000	Varianza	Promedio
MLE	0.0009305188	2.996256
MOM	0.0020924081	3.035356

calculando los errores cuadráticos medios (estimados) obtenemos los valores

n	MLE	MOM
50	0.08739	0.13419
100	0.0144	0.02485
200	0.0063	0.01333
500	0.0086	0.00746
1000	0.0009	0.00334

y como se puede apreciar, el estimador ML (simulado) presenta menor error cuadrático medio estimado en todos los casos a excepción cuando $n = 500$, manteniendo casi siempre la conclusión del ítem anterior.

c). Por el teorema central del límite, sabemos que

$$\frac{\sqrt{n}(\bar{X}_n - \theta)}{\sqrt{2}} \xrightarrow{d} \mathcal{N}(0,1)$$

cuando $n \to \infty$. Ahora, sea $g(x) = x\sqrt{2}$, derivando, $g'(x) = \sqrt{2}$, así aplicando el método delta, obtenemos:

$$\frac{\sqrt{n}(g(\bar{X}_n) - g(\theta))}{\sqrt{2}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, [g'(\theta)]^2 * 1)$$

esto es

$$Z_n = \sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,2)$$

Finalmente, veamos a qué converge M_n. Se tiene que $Me \xrightarrow{c.s.} 3$ cuando $n \to \infty$ puesto que[9] si buscamos un me tal que

$$\int_{-\infty}^{me} \frac{1}{2} e^{-|x-3|}\, dx = \frac{1}{2}$$

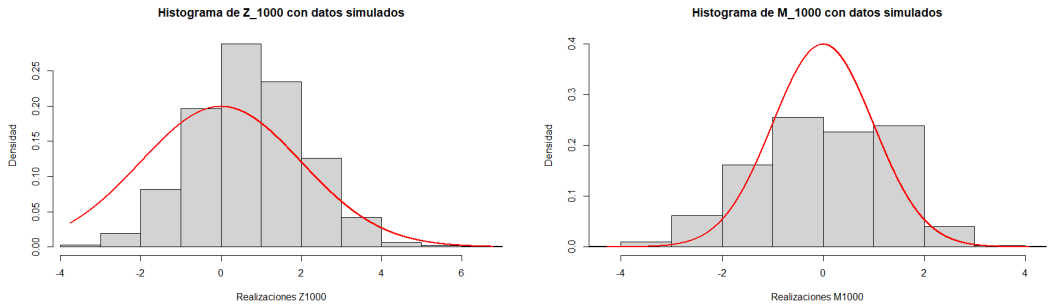
dado que el integrando es una función simétrica, con eje de simetría en $x = 3$, obtenemos que $me = 3$ es la mediana poblacional a la cuál, sabemos, converge casi siempre la estadística Me, además como $(f_X(3))^2 = \left(\frac{1}{2}\exp(-|3-3|)\right)^2 = \frac{1}{4}$ entonces por lo visto en clase sobre la distribución asintótica de una estadística de orden,

$$\sqrt{n}(Me - 3) \xrightarrow[n\to\infty]{d} \mathcal{N}\Big(0, \frac{1/4}{1/4}\Big)$$

$$M_n = \sqrt{n}(Me - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

[9]Observe que este resultado también se puede obtener con ayuda de R (ver script).

d). Se presentan los histogramas de las realizaciones de Z_{1000} y M_{1000} respectivamente:



aparentemente los modelos parecen ajustar, sin embargo sería bueno ejecutar varias veces las estimaciones de Z_{1000} y M_{1000} para dar mayor convencimiento. Por la teoría vista en clase sabemos que la eficiencia asintótica es:

$$ARE(Z_n, M_n) = 2/1 = 2 > 1$$

y por tanto, M_n es un estimador más eficiente asintóticamente que Z_n.

e). **Bonus**: Se comparan[10] los estimadores moda, máximo verosímil y de momentos en términos del error cuadrático medio estimado:

n	MLE	MOM	Moda
10	2.25774	2.66614	2.23568
50	0.08739	0.13419	0.11554
100	0.0144	0.02485	0.09526
200	0.0063	0.01333	0.02536
500	0.0086	0.00746	0.05358
1000	0.0009	0.00334	0.00195

Según lo observado, la moda es un estimador más eficiente (al menos en error cuadrático medio estimado) que el estimador máximo verosímil y por momentos para valores pequeños de n, sin embargo, a medida que n se hace grande se vuelve el estimador con la menor eficiencia de los tres.

[10]Ver script.

Técnica de remuestreo (Bootstrap)

a). Generamos una única muestra simulada de tamaño n=10 de una distribución *N(3,9)*, y guardamos su media y varianza en las variables *meandata* y *vardata*. En nuestra simulación, la muestra simulada es:

6.253867 -2.733987 10.745161 2.633719 -4.340087 2.370666 1.459805 3.864737 6.922613 6.317181

Tenemos además que: $\overline{x_{10}} \approx 3{,}3494$ y $s_{10}^2 = 20{,}6834$ El promedio resulta relativamente cercano al parámetro μ (con una diferencia de 0.3494). No obstante, la varianza de los datos termina siendo bastante alejada del parámetro σ^2. Ello se debe principalmente a que el tamaño de muestra no es lo suficientemente grande (n=10) (Aunque cabe anotar que, en términos de MSE, la varianza muestral es menos eficiente que el estimador $\frac{n-1}{n}S_n^2$).

b). Calculamos la función empírica de los datos, obteniendo la gráfica:

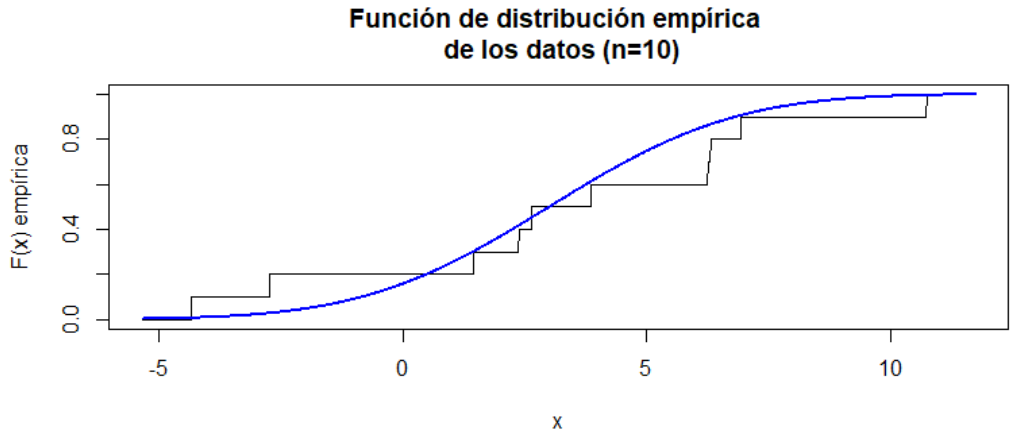


Figura 6: Función de distribución empírica. En azul, se muestra la función de distribución real de la población

Se puede ver cómo la función de distribución empírica $\hat{F}_n(x)$ tiene un comportamiento similar a la de la distribución poblacional. Sin embargo, dedido de nuevo al pequeño tamaño del muestra, tal similitud no es tan marcada como se podría querer.

c). Olvidemos ahora que sabemos de dónde provienen los datos, y realicemos la técnica de remuestreo sobre nuestra muestra (a la que ahora llamaremos en el código como *data*). En el código, se guardan $B = 1000$ muestras *bootstrap* de nuestra muestra inicial en la matriz *mat*.

d). Ahora, calculamos para cada una de estas muestras *bootstrap* la media $\{\overline{x_i}\}_{i=1}^B$, la expresión de la varianza $\{\frac{(n-1)s_i}{\hat{\sigma}^2}\}_{i=1}^B$ (donde $\hat{\sigma}^2$ es la estimación de la varianza en a.), y el coeficiente de variación $\{\frac{s_i}{\overline{x_i}}\}_{i=1}^B$.
Por lo visto en clase, sabemos que, en una población normal, $\overline{X_i} \sim N(\mu, \sigma^2/n)$. En nuestro caso específico, $\overline{X_{10}} \sim N(3, 9/10)$. Adicionalmente, en clase se vio que, para una población normal, $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$. En nuestro caso particular, $\frac{(10-1)S_n^2}{\sigma^2} \sim \chi^2(10-1)$. Para los histogramas de la media y la expresión de la varianza, se encuentran superpuestas las distribuciones teóricas correspondientes a las variables aleatorias relacionadas con cada estadística:

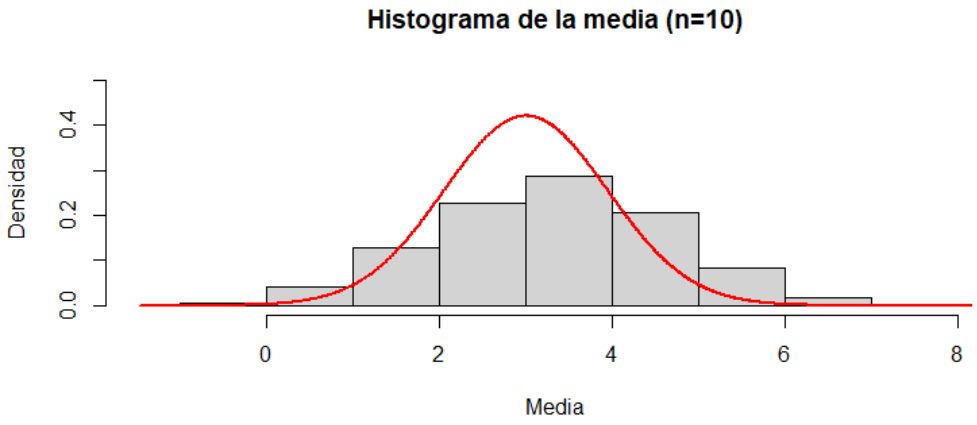


Figura 7: Histograma de la media de las muestras *bootstrap* de la muestra simulada inicial. En rojo, la función de densidad de una distribución $N(3, 9/10)$.

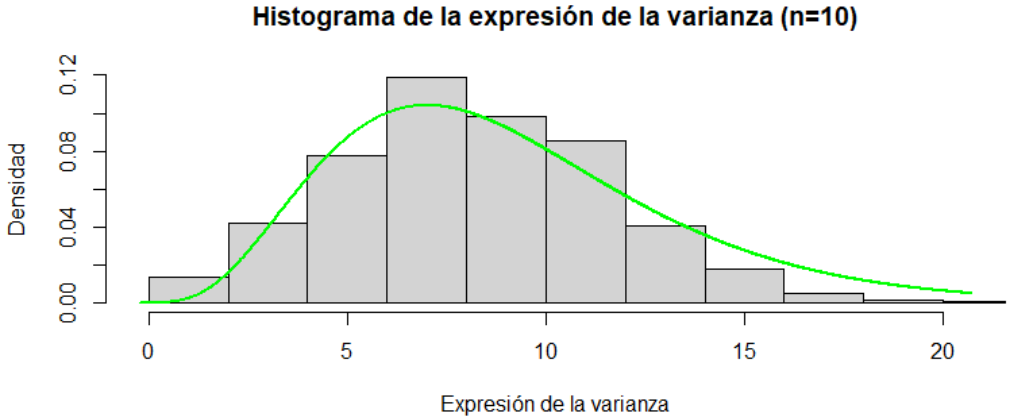


Figura 8: Histograma de la expresión $\frac{(10-1)s_i}{\tilde{\sigma}^2}$ de las muestras *bootstrap* de la muestra simulada inicial. En verde, la función de densidad de una distribución $\chi^2(10-1)$.

A continuación, presentamos el histograma arrojado para el coeficiente de variación:

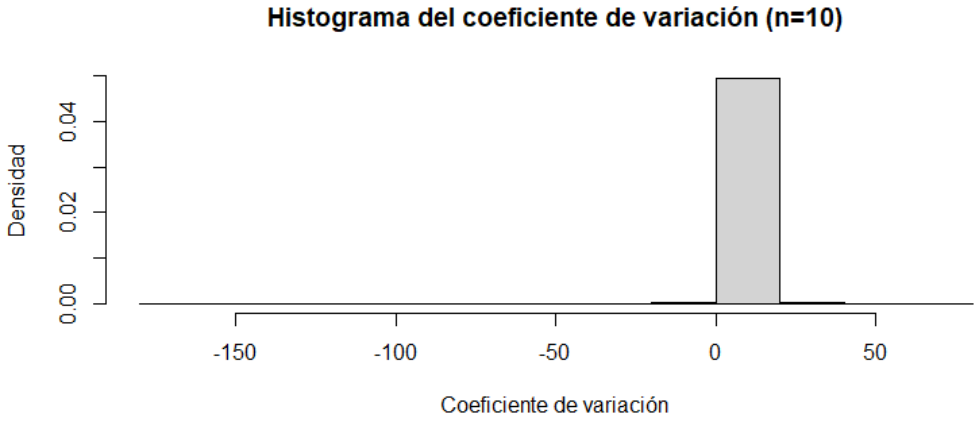


Figura 9: Histograma del coeficiente de variación de las muestras *bootstrap* de la muestra simulada inicial.

Como podrá notar, el histograma generado no es muy diciente respecto a la distribución de la variable aleatoria correspondiente al coeficiente de variación. Esto se debe, principalmente, a que el coeficiente de variación es una estadística que tiene un "buen comportamiento" cuando el promedio de nuestra muestra está lo suficientemente alejado de 0. En nuestro caso, puede que se hayan obtenido muestras con promedios que no eran lo suficientemente alejados de este número. Por consiguiente, tales muestras produjeron una serie de datos atípicos, que se ven reflejados en nuestro histograma. Para corregir este problema, restringamos el histograma sobre el intervalo [0,20], aumentando el número de clases, para visualizar un poco mejor la distribución de la variable aleatoria:

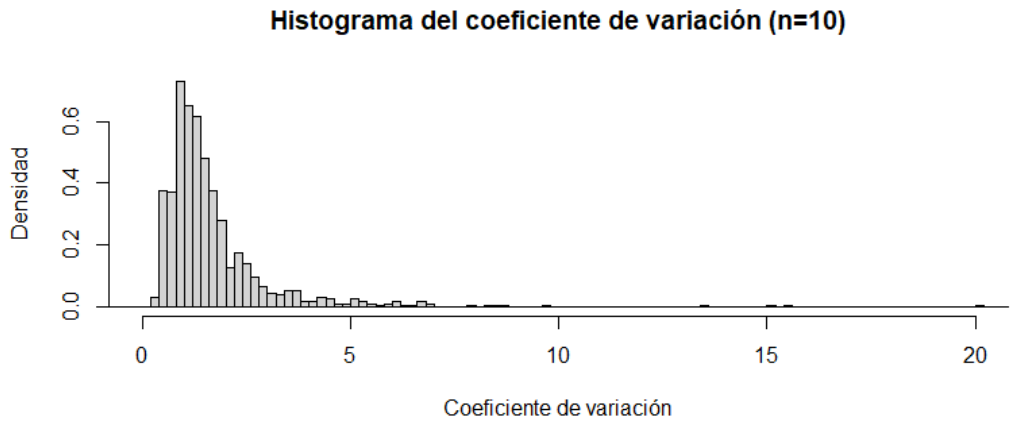


Figura 10: Histograma del coeficiente de variación de las muestras *bootstrap* de la muestra simulada inicial sobre el intervalo [0,20].

e). De los resultados vistos en (d), podemos concluir que la técnica de remuestreo nos da buenas aproximaciones de la distribución de cada variable aleatoria propuesta, teniendo en cuenta el tamaño de muestra dado.
Existe una clara relación directa entre lo cercana que es la función de distribución a la distribución real de la función, y la confiabilidad de los datos arrojados. Como en este caso, la muestra tuvo un promedio un poco mayor que el de la población, podemos ver un ligero sesgo a la derecha de nuestro histograma con respecto a la curva teórica.

f). En el caso del coeficiente de variación, se presentaron pequeños inconvenientes, debido a su comportamiento caótico con un promedio cercano a cero. Sin embargo, supimos controlar de alguna forma tal problema, y concluir que bajo la técnica de remuestreo (ignorando datos atípicos), la variable aleatoria correspondiente al coeficiente de variación, parece tener una distribución de forma acampanada con asimetría positiva.
El valor real del coeficiente de variación poblacional es de $\frac{\sigma}{\mu} = \frac{3}{3} = 1$,. Adicionalmente, la media muestral de todos los datos fue de $1{,}617344$. En el siguiente histograma se pueden visualizar mejor tales diferencias:

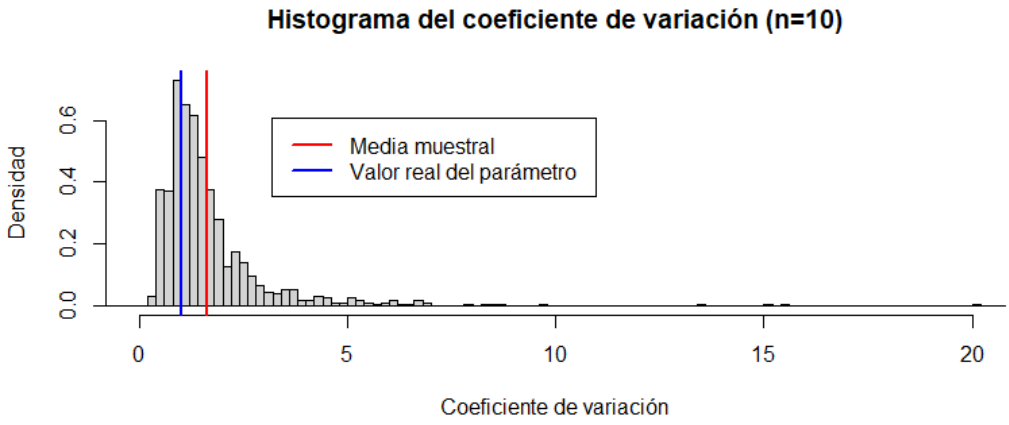


Figura 11: Histograma del coeficiente de variación de las muestras *bootstrap* de la muestra simulada sobre el intervalo [0,20], con su respectiva media y valor real del parámetro.

Bajo esta información, puede parecer que el estimador $S_n/\overline{X_n}$ fuera insesgado. Sin embargo, debido al pequeño tamaño de muestra, es recomendable ser muy cautelosos con lo que se pueda decir al respecto. El tamaño de tal sesgo, si existe, es un dato aún más difícil de deducir con un tamaño de muestra tan reducido.

g). Ahora, tomemos un tamaño de muestra más grande: $n = 1000$. Veamos cómo ello afecta en la confiabilidad de los resultados obtenidos. Respondamos a cada uno de los items anteriores con este nuevo tamaño de muestra:

- Generamos una única muestra simulada de tamaño $n = 1000$ de una distribución $N(3, 9)$, y guardamos su media y varianza en las variables *meandata* y *vardata*. Tenemos que: $\overline{x_{1000}} \approx 2{,}976291$ y $s^2_{1000} \approx 8{,}738599$, los cuales son valores bastante cercanos a los verdaderos parámetros. Resultados que eran de esperarse, debido a que ambas estadísticas convergen casi seguramente hacia la media y varianza poblacional, respectivamente.
- A continuación, se presenta la gráfica de la función de distribución empírica de los datos:

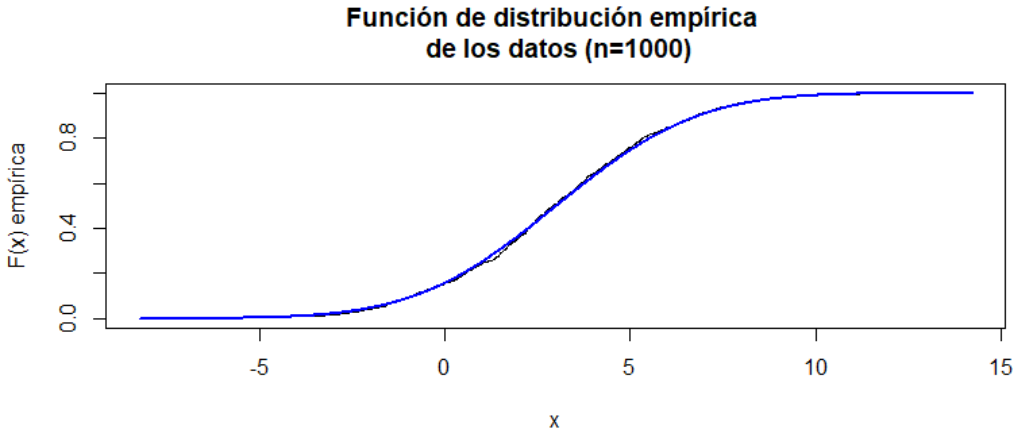


Figura 12: Función de distribución empírica. En azul, se muestra la función de distribución real de la población

Podemos observar cómo la función de distribución empírica $\hat{F}_n(x)$ tiene un comportamiento muy similar al de la función de distribución poblacional (mucho más que con la muestra de tamaño 10).

- Guardemos $B = 1000$ muestras *bootstrap* de la muestra inicial (que ahora llamamos *data*) en la matriz *mat*.

- Ahora, calculamos para cada una de estas muestras *bootstrap* la media $\{\overline{x_i}\}_{i=1}^{B}$, la expresión de la varianza $\{\frac{(n-1)s_i}{\hat{\sigma}^2}\}_{i=1}^{B}$ (donde $\hat{\sigma}^2$ es la estimación de la varianza en a.), y el coeficiente de variación $\{\frac{s_i}{\overline{x_i}}\}_{i=1}^{B}$. Similar a como se vió para la muestra simulada de tamaño 10, podemos deducir que $\overline{X_{1000}} \sim N(3, 9/1000)$ y, además, $\frac{(1000-1)S_n^2}{\sigma^2} \sim \chi^2(1000-1)$. Las curvas de densidad de tales distribuciones se encuentran superpuestas al histograma de sus estadísticas correspondientes.

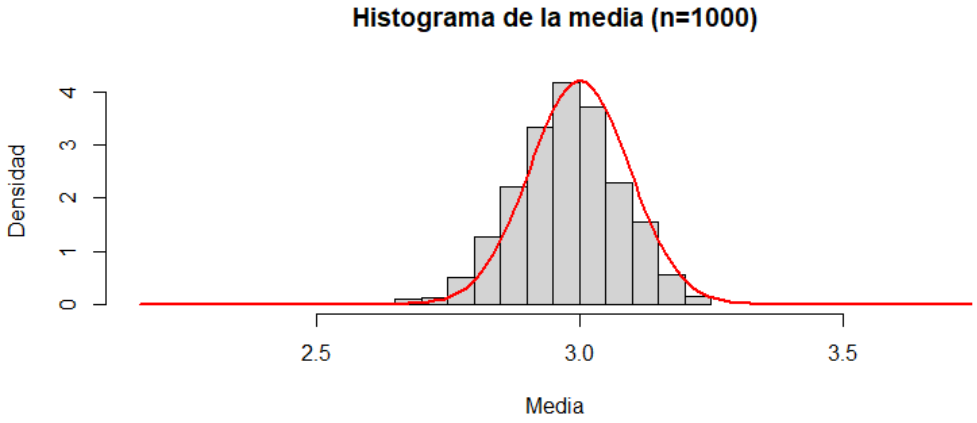


Figura 13: Histograma de la media de las muestras *bootstrap* de la muestra simulada inicial. En rojo, la función de densidad de una distribución $N(3, 9/1000)$.

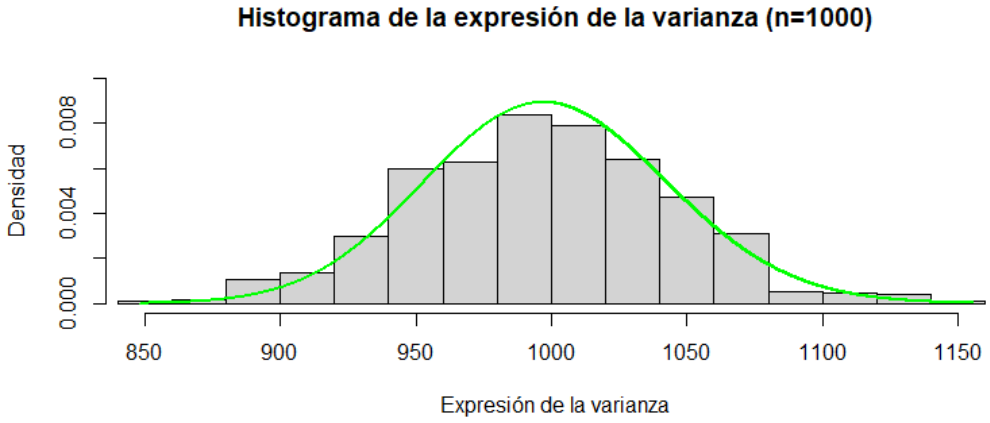


Figura 14: Histograma de la expresión $\frac{(1000-1)s_i}{\hat{\sigma}^2}$ de las muestras *bootstrap* de la muestra simulada inicial. En verde, la función de densidad de una distribución $\chi^2(1000-1)$.

Para el caso del coeficiente de variación, debido a que hubo mayor precisión en el promedio (un mayor tamaño de muestra), no hubo datos atípicos, y se puede ver claramente el comportamiento de tal estadística:

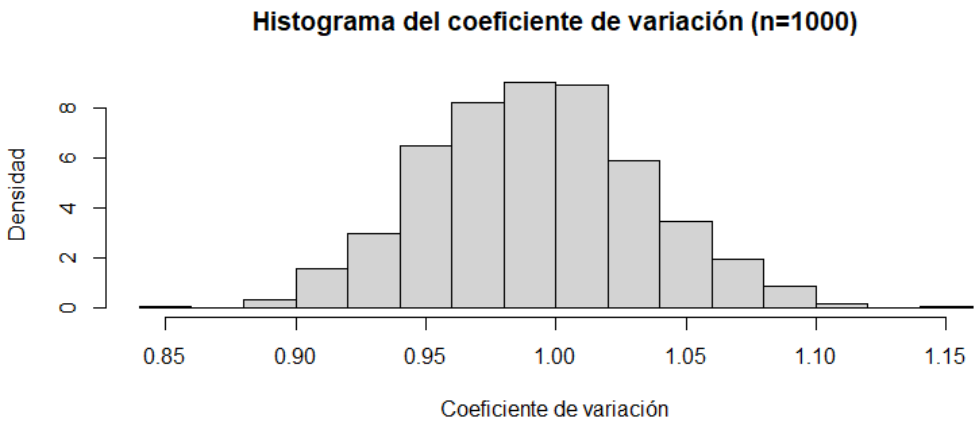


Figura 15: Histograma del coeficiente de variación de las muestras *bootstrap* de la muestra simulada inicial.

- Hay una mayor similitud entre los histogramas de las estadísticas y las curvas teóricas en los primeros dos casos, gracias a que trabajamos con un mayor tamaño de muestra, y su función de distribución empírica es mucho más similar a la función de distribución real.
- En cuanto al coeficiente de variación, el histograma 15 nos ayuda a ver que el estimador $S_{1000}/\overline{X_{1000}}$ tiene una distribución acampanada con asimetría positiva, con una varianza bastante menor a la correspondiente al tamano de muestra de $n = 10$.
  Para estudiar el sesgo de nuestro estimador, calculamos el promedio de los coeficientes de variación correspondientes. El promedio muestral es de aproximadamente $0{,}99$. El siguiente histograma muestra las diferencias entre el valor real del parámetro, y el promedio muestral:

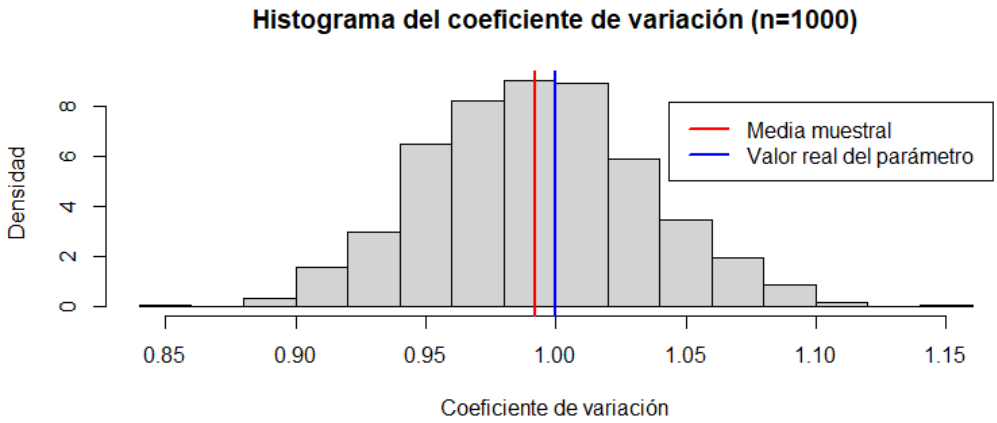


Figura 16: Histograma del coeficiente de variación de las muestras *bootstrap* de la muestra inicial, con su respectiva media y el valor real del parámetro.

Podemos deducir entonces que S_n/X_n es un estimador para σ/μ insesgado, o con un sesgo que se va reduciendo a medida que aumenta n (el valor absoluto del sesgo tuvo una disminución significativa cuando pasamos de $n = 10$ a $n = 1000$). Lo que sí es bastante seguro es que la presición de tal estimador aumenta a medida que el tamaño de muestra crece.

Parte B

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, tenemos en principio la siguiente posibilidad:

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

donde $\hat{p}_n = \overline{X}_n$, es decir la media muestral.
Aunque la variable aleatoria no es monótona en p, a continuación mostramos el procedimiento para despejar al parámetro de alli y poder obtener un intervalo de confianza.

Procedimiento.
En principio tenemos que:

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

Elevando al cuadrado a ambos lados obtenemos:

$$p\left(\frac{n(\hat{p}_n - p)^2}{p(1-p)} \leq z^2_{1-\frac{\alpha}{2}} \right) = p\left(\frac{\hat{p}_n^2 - 2\hat{p}_n p + p^2}{p(1-p)} \leq \frac{z^2_{1-\frac{\alpha}{2}}}{n} \right)$$

Multiplicando por $p(1-p)$ a ambos lados de la desigualdad obtenemos:

$$p\left(\hat{p}_n^2 - 2\hat{p}_n p + p^2 \leq p(1-p) \cdot \frac{z^2_{1-\frac{\alpha}{2}}}{n} \right) = p\left(\hat{p}_n^2 - 2\hat{p}_n p + p^2 \leq p \cdot \frac{z^2_{1-\frac{\alpha}{2}}}{n} - p^2 \cdot \frac{z^2_{1-\frac{\alpha}{2}}}{n} \right)$$

Si pasamos la expresion de la derecha a la parte izquierda de la igualdad, agrupando terminos por p^2 y p obtenemos:

$$p\left(\left(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}\right) \cdot p^2 - \left(2\hat{p}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{n}\right) \cdot p + \hat{p}_n^2 \leq 0 \right)$$

Como $\hat{p}_n = \overline{X}_n$ y $z^2_{1-\frac{\alpha}{2}}$ son constantes conocidas, la expresión tiene la forma de una ecuación cuadrática de la forma $ax^2 + bx + c$ donde:

$$a = \left(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}\right),\ b = -\left(2\hat{p}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{n}\right),\ c = \hat{p}_n^2,\ x = p$$

Por lo tanto las soluciones de la ecuación cuadrática son:

$$p_{1,2} = \frac{\left(2\hat{p}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{n}\right) \pm \sqrt{\left(2\hat{p}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{n}\right)^2 - 4\hat{p}_n^2 \cdot \left(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}\right)}}{2\left(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}\right)}$$

Resolviendo el termino que está adentro de la raiz cuadrada y simplificando la ecuación obtenemos:

$$p_{1,2} = \frac{\hat{p}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{2n}}{1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{z^2_{1-\frac{\alpha}{2}}}{4n^2}}}{1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}}$$

Entonces obtenemos el siguiente resultado:

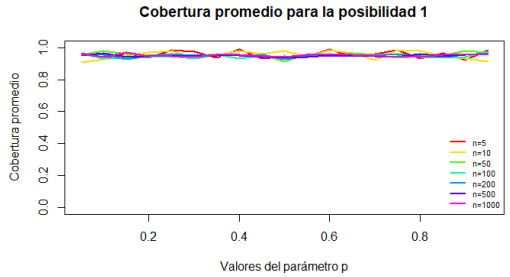
$$p\left(\frac{\hat{p}_n+\frac{z^2_{1-\frac{\alpha}{2}}}{2n}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}-z_{1-\frac{\alpha}{2}}\cdot\frac{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}+\frac{z^2_{1-\frac{\alpha}{2}}}{4n^2}}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}\leq p\leq\frac{\hat{p}_n+\frac{z^2_{1-\frac{\alpha}{2}}}{2n}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}+z_{1-\frac{\alpha}{2}}\cdot\frac{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}+\frac{z^2_{1-\frac{\alpha}{2}}}{4n^2}}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}\right)\approx 1-\alpha$$

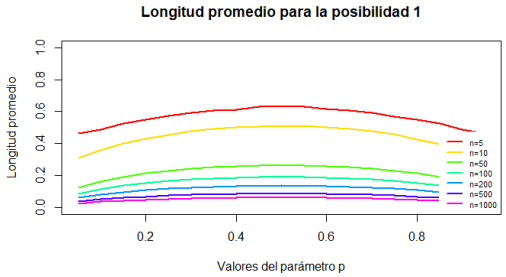
Por lo tanto el intervalo de confianza asintótico para la proporción p viene dado por:

$$ICA_{100(1-\alpha)\,\%}(p)=\frac{\hat{p}_n+\frac{z^2_{1-\frac{\alpha}{2}}}{2n}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}\pm z_{1-\frac{\alpha}{2}}\cdot\frac{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}+\frac{z^2_{1-\frac{\alpha}{2}}}{4n^2}}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}$$

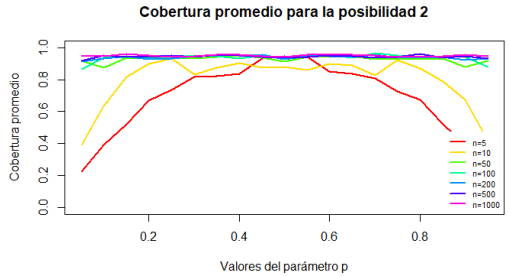
donde $\hat{p}_n=\overline{X}_n$.
Para hacer una primera comparación gráfica entre las cinco posibilidades, a continuación se presentan las gráficas de los datos obtenidos al ejecutar el algoritmo haciendo la comparación entre el parámetro como variable independiente, y la cobertura promedio y la longitud promedio de los intervalos como variables dependientes.

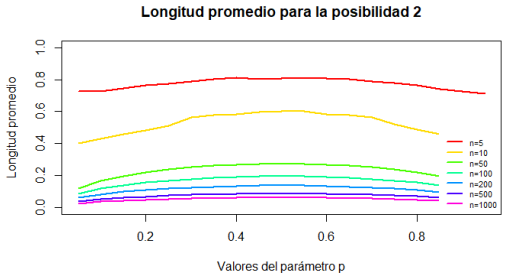
- **Posibilidad 1**



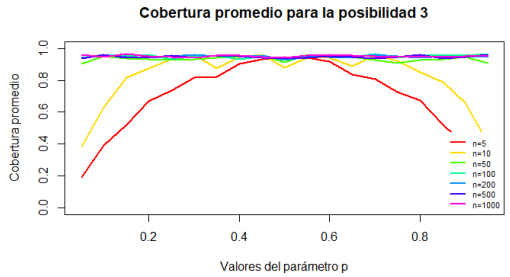


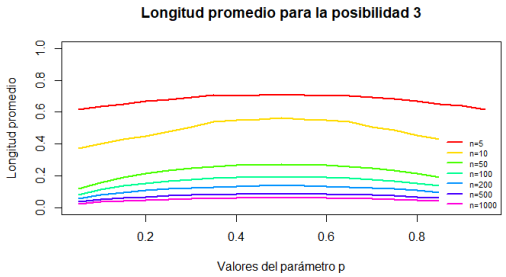
- **Posibilidad 2**



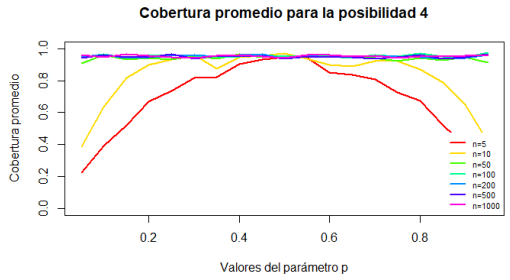


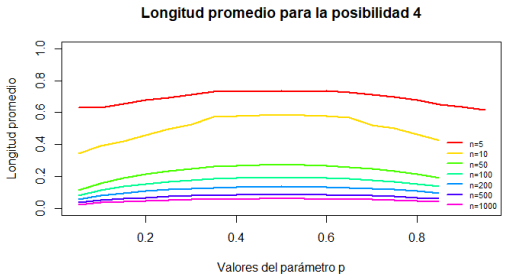
- **Posibilidad 3**



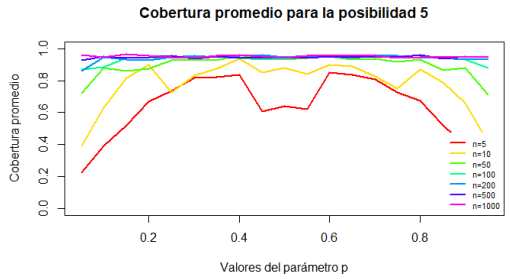


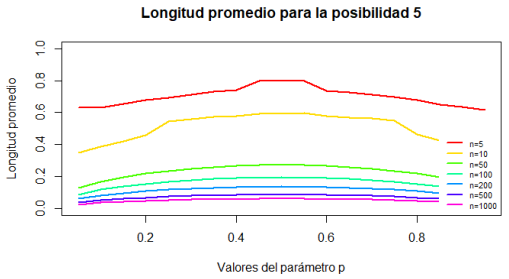
- **Posibilidad 4**





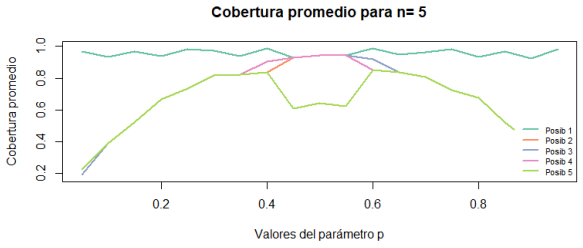
- **Posibilidad 5**

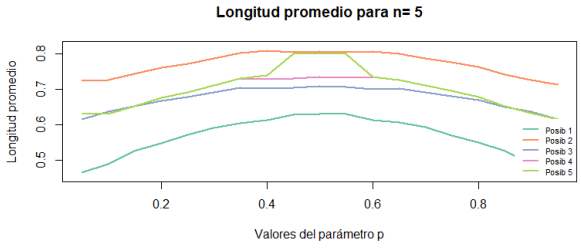




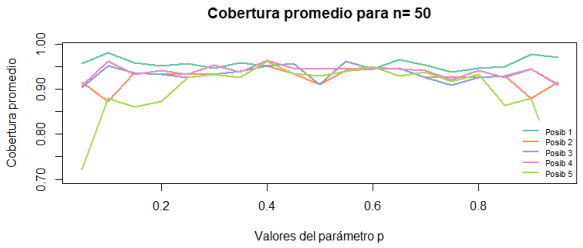
Ahora, nos concentraremos sobre los tamaños de muestra $n = 5, 50, 200, 1000$. Para cada uno de ellos, realizamos una comparación gráfica entre la cobertura promedio de cada una de las posibilidades. Ejecutamos el mismo procedimiento para comparar las longitud promedio de los intervalos.

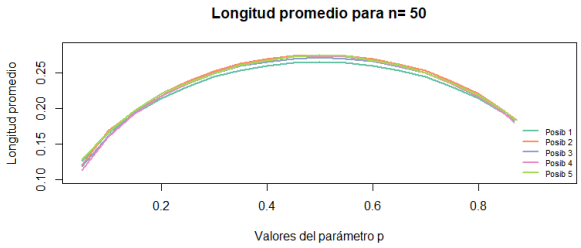
- **n=5**



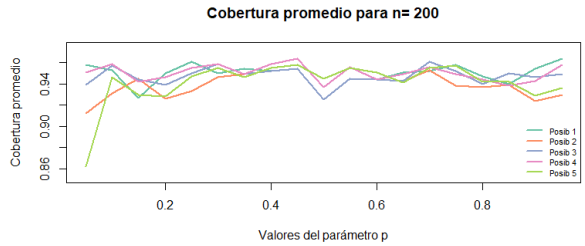


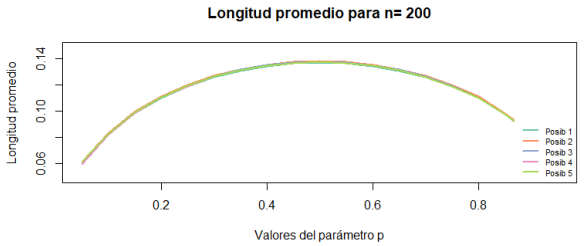
- **n=50**



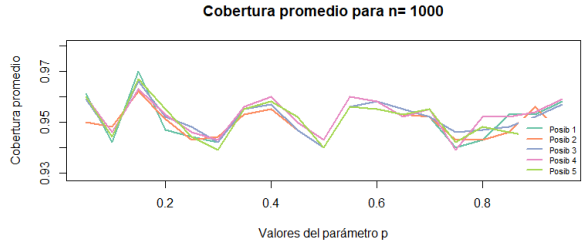


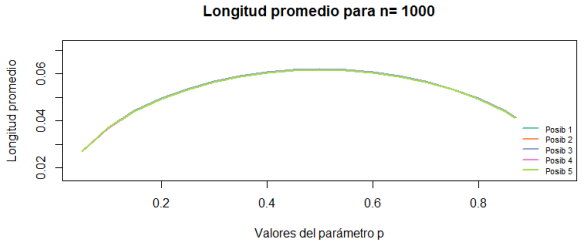
- **n=200**





- **n=1000**

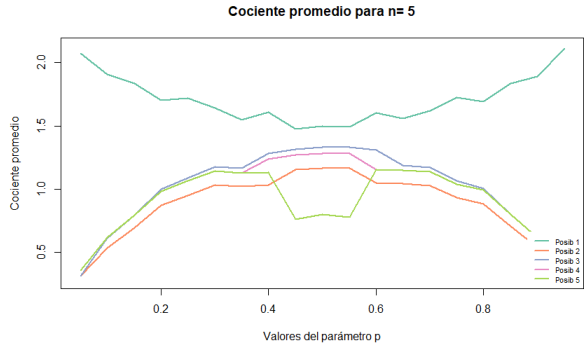


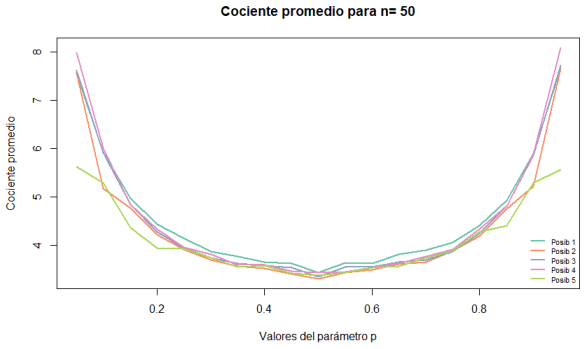


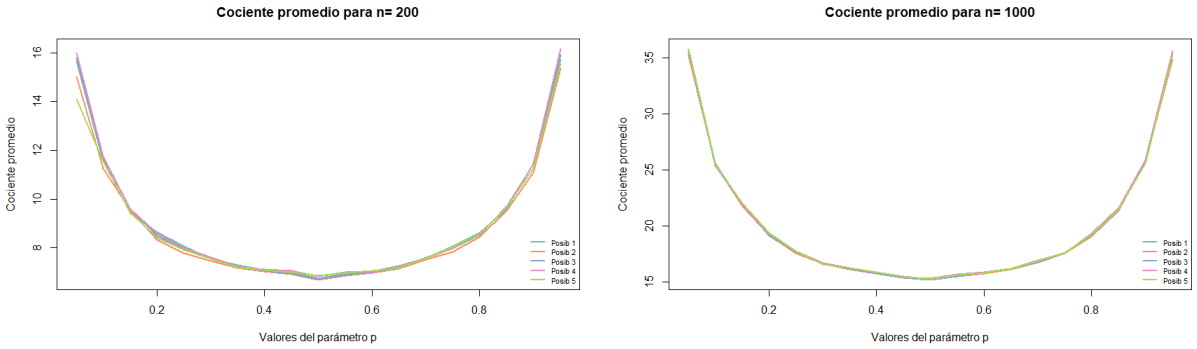
Concluimos así, según lo anterior, que

- Para tamaños de muestra pequeños la posibilidad 1 es bastante más eficiente que los demás, pues en cuanto a cobertura, es estable en valores cercanos a 1 y para todos los posibles valores de p, y en cuanto a longitud del intervalo, es quien tiene las menores.
- A medida que n crece, la posibilidad 1 sigue siendo muy eficiente respecto al resto, sin embargo, las demás posibilidades se van tornando eficientes, aproximándose y para algunos valores de p incluso mejorando dicha eficiencia.
- Cuando $|p| \approx 0{,}5$ la longitud de todas las posibilidades se hace máxima, para casi cualquier valor de n.
- Por lo visto, es preferible siempre tener el tamaño de muestra más grande posible en el estudio, tanto para tener una longitud más acotada para el parámetro, como para que los métodos sean eficientes en cuánto a cobertura del mismo.
- Si se tiene un valor de n suficientemente grande, cualquier método funciona de manera similar, por tanto, cualquiera de las posibilidades las podríamos implementar en nuestro estudio.

Finalmente, se muestra, gráficamente, el cociente cobertura-longitud de cada una de las posibilidades





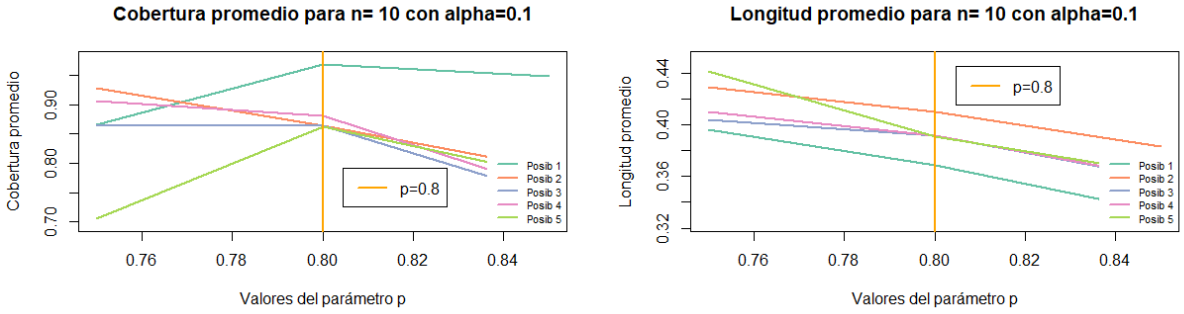


avalando nuestras conclusiones respecto a la eficiencia (cobertura-longitud), de las posibilidades. Además, observamos que para tamaños de muestras "grandes", las curvas tienen su mínimo en valores cercanos a $p = 0{,}5$ lo que una vez más, nos muestra que la longitud se hace máxima ahí.
Por último, se muestran dos ejemplos de las aplicaciones de este resultado al contexto de los intervalos de confianza:

Caso 1: 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 % ?

Caso 2: 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 estos resultados, se obtuvo que un 10 % de los encuestados dedican su tiempo de ocio a leer un libro. ¿Qué podría decirse de la proporción de estudiantes en la población que prefieren leer un libro? Evalúe este resultado con una confianza del 99 %.

Solución Caso 1: Debido a que estamos trabajando con un tamaño de muestra pequeño ($n = 10$), podemos deducir que el más eficiente de los métodos vistos es el primero . Esto se puede ver reflejado en las siguientes gráficas, donde se realizó un procedimiento muy similar a lo realizado en la simulación de las posibilidades con $\alpha = 0{,}05$, cambiándolo esta vez por $\alpha = 0{,}1$. Se muestra el desempeño de cada método alrededor de nuestra estimación $\hat{p} = 0{,}8$:



Luego, por lo visto en la primera parte de esta sección sabemos que

$$ICA_{100(1-\alpha)\,\%}(p) = \frac{\hat{p}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{2n}}{1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{z^2_{1-\frac{\alpha}{2}}}{4n^2}}}{1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}}$$

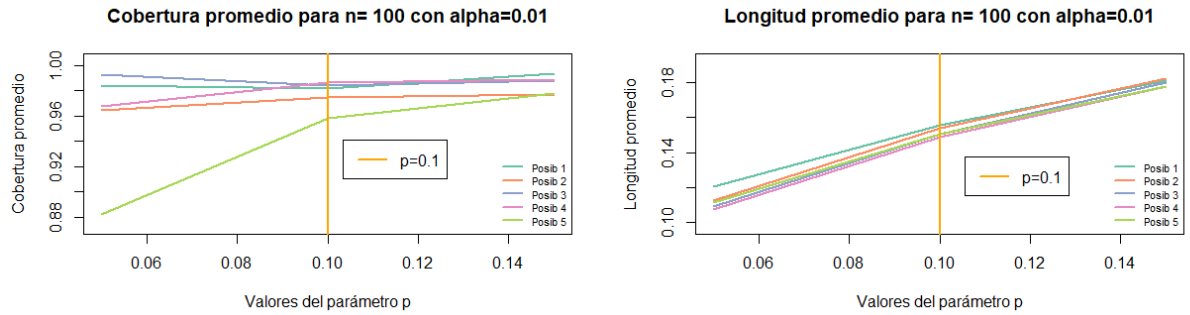
donde en este caso $\hat{p}_{10} = 0{,}8$, $\alpha = 0{,}1$ y $n = 10$. Haciendo las cuentas en R[11] obtenemos que

$$ICA_{90\,\%}(p) \approx (0{,}677, 0{,}884)$$

[11] Ver script.

Podríamos entonces concluir que la mayoría de los componentes cumplen con los estándares, más específicamente, la proporción de componentes que cumplen con los estándares de producción está entre el 67 % y 88.4 % de los componentes en la población.

Solución Caso 2: A continuación, se pueden ver los gráficos correspondientes al tamaño de muestra $n = 100$. Se resaltan los resultados obtenidos alrededor de nuestra estimación $\hat{p} = 0{,}1$ y $\alpha = 0{,}01$, debido a que esta es la mejor aproximación del parámetro que se tiene para esta situación:



Como se puede notar, el método más eficiente para este tamaño de muestra, con una confianza del 99 %, para esta aproximación de la proporción es el número 4 (*bootstrap* basado en percentiles). En el código adjunto, puede ver el procedimiento realizado para esta posibilidad. Obtenemos que, mediante la posibilidad 4[12]:

$$ICA_{99\,\%}(p) = [0, 0{,}15]$$

Podemos entonces concluir que, con un nivel aproximado de confianza del 99 %, y utilizando el método más efectivo bajo las condiciones dadas (*bootstrap* basado en percentiles), la proporción de estudiantes que prefieren leer un libro está entre 0 y $0{,}4201$.

[12]Ver script

Referencias

Blanco, L. (2010). Probabilidad (2nd ed., Vol. 1). Universidad Nacional de Colombia.
Mayorga, H. (2004). Inferencia Estadística (Primera Edición).Universidad Nacional de Colombia.