Trabajo Práctico: Ley de los Grandes Números y Teorema Central del Límite

Julián Barrios L.U: 718/18
Alejandra Delgado L.U.: 777/14

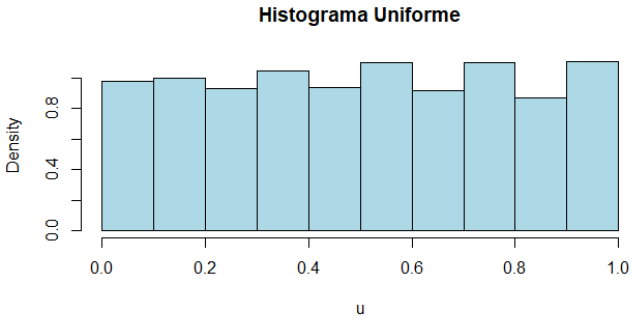
En el siguiente trabajo, tanto la generación y análisis de datos, como los gráficos fueron hechos en R. Se utilizó como semilla el siguiente número: 71818.

Primera parte

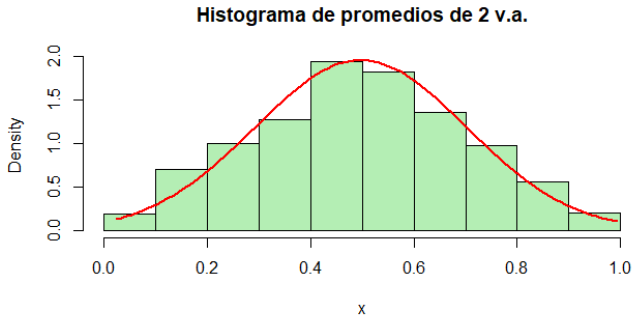
1. El objetivo de este primer punto es estudiar la distribución del promedio de variables aleatorias i.i.d., a medida que va aumentando n, siendo n el número de variables aleatorias.

   (a) En primer lugar procedimos a generar nuestro primer conjunto de datos, formado por 1000 variables aleatorias i.i.d. con distribución $U(0,1)$.

   Con dichos datos construímos un histograma (Fig.1 (a)), donde podemos observar que la muestra se distribuye de manera uniforme:



(a) Histograma v.a. uniforme estándar



(b) Histograma de los promedios de dos v.a.

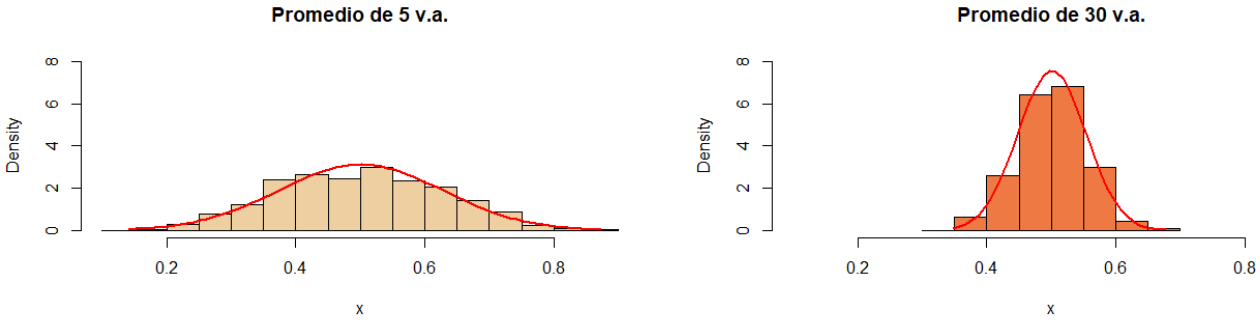
Figure 1: Histogramas de promedios de variables aleatorias para distintos valores de n.

   (b) Luego consideramos el promedio de dos variables aleatorias i.i.d., X_1 y X_2, ambas con distribución uniforme estándar $U(0,1)$. Luego de replicar 1000 veces el promedio entre ambas, obtuvimos el histograma de la Fig.1 (b).

   Podemos observar en este histograma que la densidad presenta una forma acampanada, donde el 50% de los valores se encuentra entre 0.35 y 0.64. La distribución de los datos se asemeja a una normal -graficada en rojo- con μ igual al promedio del conjuntos de datos (es decir, centrándose aproximadamente en el valor 0.5, que coincide con la esperanza de una v.a. uniforme estándar) y cuya varianza corresponde a la varianza de los promedios (ver Tabla 1).

   (c) Considerando ahora cinco variables aleatorias i.i.d. con distribución $U(0,1)$, computamos el promedio 1000 veces, obteniendo el gráfico de la Fig.2 (a). Observamos que el 50% de los datos

se encuentra entre 0.4 y 0.6, y el 73% entre 0.35 y 0.65, lo que implica una distribución de datos más concentrada alrededor del valor 0.5, en comparación con la muestra anterior.

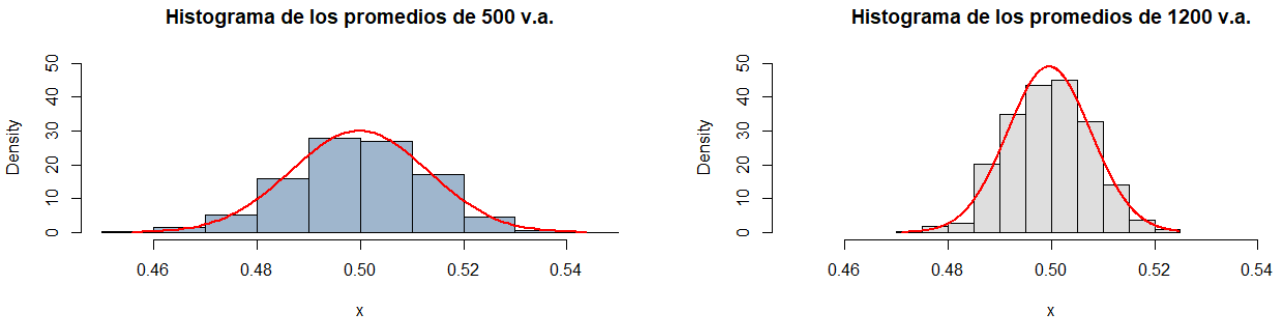


(a) n=5 (b) n=30

Figure 2: Histograma de los promedios de variables aleatorias uniformes estándar.

(d) Aumentamos una vez más el número de variables aleatorias i.i.d. $U(0,1)$ a 30 y calculamos los promedios. Obtuvimos un histograma (Fig. 2 (b)) A diferencia del caso anterior, el 94.2% de datos se encuentra entre 0.4 y 0.6, con lo cual se evidencia una alta acumulación de datos alrededor del valor 0.5; es decir, el conjunto de datos tiene cada vez menos dispersión.

(e) Replicando la anterior con 500 variables aleatorias i.i.d. $X_i \sim U(0,1)$ obtuvimos el gráfico que se observa en la Fig. 3 (a):



(a) n=500 (b) n=1200

Figure 3: Histograma de los promedios de variables aleatorias uniformes estándar.

Podemos ver que a medida que vamos aumentando el tamaño de la muestra, la dispersión va disminuyendo (ya que la varianza disminuye como 1/n) y los datos van acumulándose más alrededor de μ.

Además, podemos observar que el 50% de los datos se concentran entre 0.49 y 0.51, y en comparación con la muestra anterior, el 100% de los datos se encuentra entre 0.46 y 0.54.

(f) Finalmente definimos 1200 variables aleatorias i.i.d. $X_i \sim U(0,1)$ y realizamos nuevamente un histograma de los promedios. En este caso observamos que todos los datos se encuentran entre

0.47 y 0.53. En comparación con el caso anterior la concentración de datos alrededor de 0.5 es mayor.

Para poder comparar mejor los datos, realizamos un boxplot con los 6 conjuntos de promedios, como se muestra en la siguiente figura (Fig. 4):

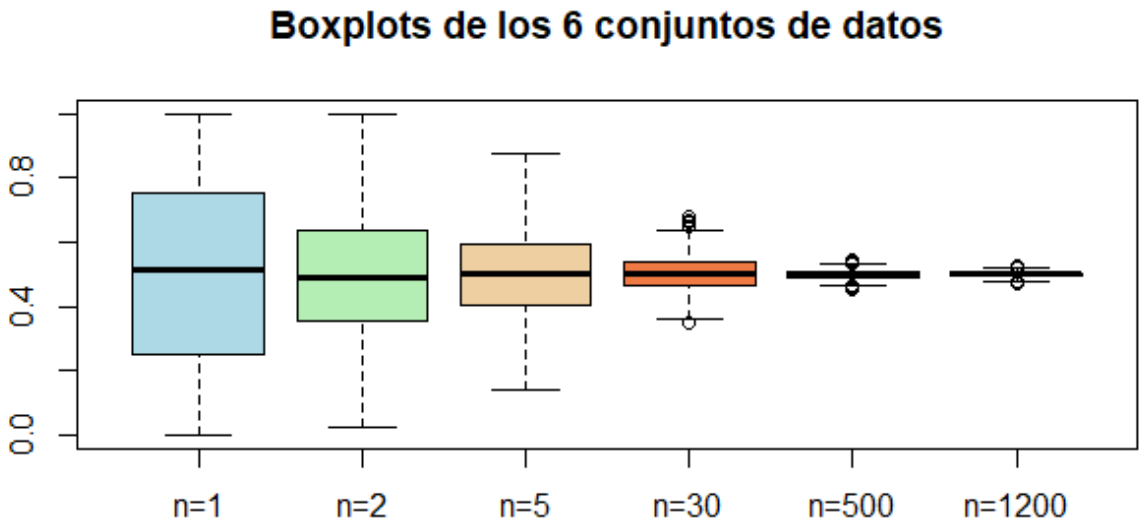


Figure 4: Boxplots de los 6 conjuntos de datos.

En este gráfico podemos observar a medida que aumentamos el número de variables promediadas, los valores tienden a concentrarse alrededor de la esperanza $E(X_i)$ de una distribución uniforme estándar $X_i \sim U(0,1)$ cuyo valor es 0.5. Esto tiene sentido ya que la Ley de los Grandes Números afirma que cualquier variable aleatoria con varianza y esperanza finitas, al promediarse n repeticiones, convergerá (en probabilidad) a su esperanza.

A continuación calculamos la media y la varianza muestral para cada conjunto de datos. Los resultados se presentan en la siguiente tabla (Tabla 1):

	n=1	n=2	n=5	n=30	n=500	n=1200
Media	0,5039398	0,49741	0,5006182	0,5010485	0,4997806	0,4996323
Varianza	8,382845e-02	4,169934e-02	1,641283e-02	2,796396e-03	1,756260e-04	6,632763e-05

Table 1: Medias y varianzas muestrales de los 6 conjuntos de valores.

Los valores teóricos los podemos calcular considerando que:

$$E(\bar{X}_n) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} E(X_i)}{n} \overset{i.i.d.}{=} \frac{\sum_{i=1}^{n} E(X_1)}{n} = \frac{nE(X_1)}{n} = E(X_1) = \mu \qquad (1)$$

$$V(\bar{X}_n) = V\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{V(\sum_{i=1}^{n} X_i)}{n^2} \overset{i.i.d.}{=} \frac{\sum_{i=1}^{n} V(X_i)}{n^2} \overset{i.i.d.}{=} \frac{nV(X_1)}{n^2} = \frac{\sigma^2}{n} \tag{2}$$

Teniendo en cuenta las ecuaciones 1 y 2 encontramos los siguientes valores teóricos de las medias y las varianzas:

	n=1	n=2	n=5	n=30	n=500	n=1200
Media	0,5	0,5	0,5	0,5	0,5	0,5
Varianza	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{60}$	$\frac{1}{360}$	$\frac{1}{6000}$	$\frac{1}{14400}$

Table 2: Medias y varianzas de los 6 conjuntos de valores.

Observamos que los valores de varianza obtenidos concuerdan con lo comentado anteriormente. Los mismos tienden a disminuir a medida que el número de variables aleatorias aumenta debido a que se concentran alrededor de la media muestral. A su vez, los valores de media y varianza muestrales coinciden, de forma aproximada, con sus respectivos valores teóricos.

Como siguiente paso realizamos qqnorms para los conjuntos de datos trabajados, para comparar los cuantiles "empíricos" con los teóricos.:

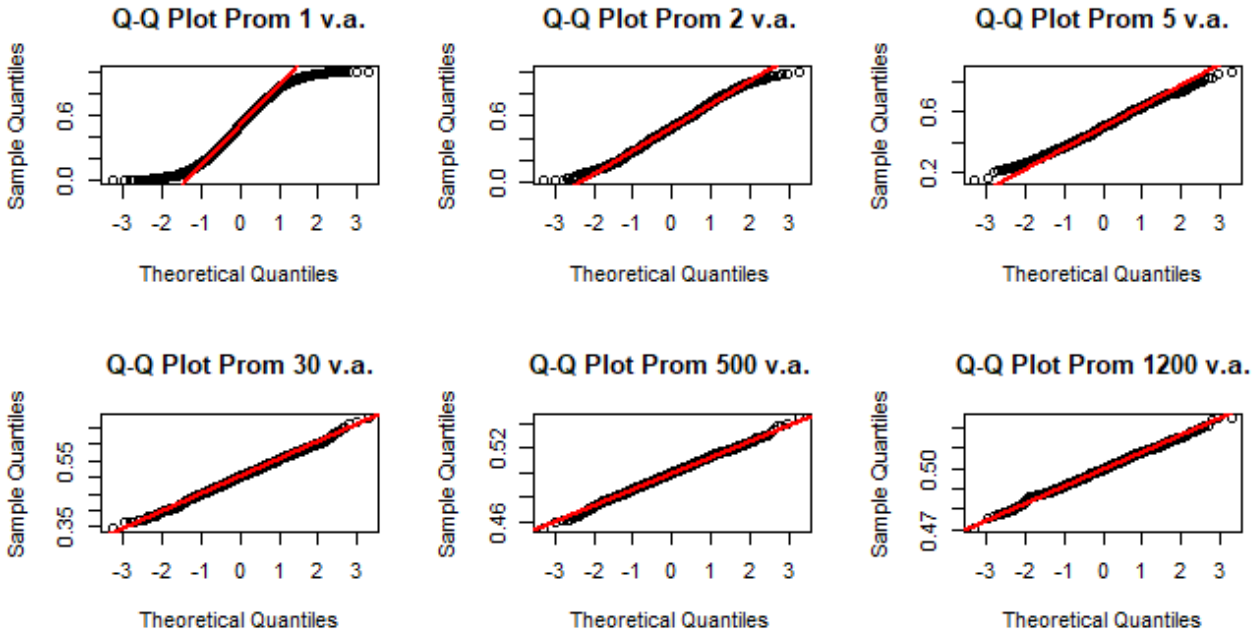


Figure 5: Q-Q Plots de los 6 conjuntos de datos.

Observamos en la Figura 5 que cuando el número de variables aleatorias promediadas aumenta, los puntos del gráfico tienden a posicionarse sobre una recta (con pendiente y ordenada aproximadamente igual σ y a μ respectivamente), indicando que la distribución de los valores se asemeja a una normal.

(g) Para poder comprobar el Teorema Central del Límite, procedimos a estandarizar los promedios. Luego realizamos histogramas y boxplots (Fig. 6 y 7) de los promedios estandarizados.

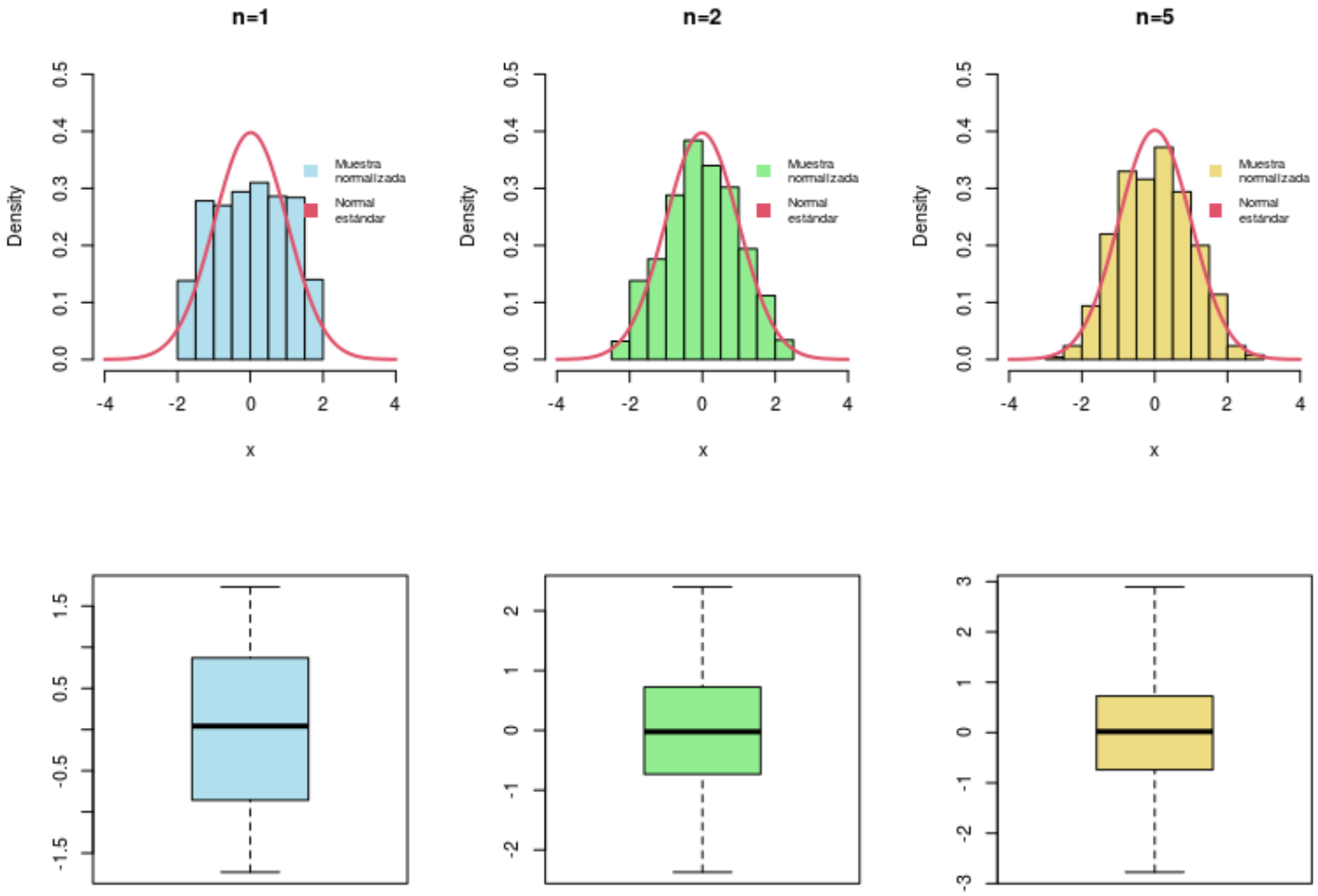


Figure 6: Histogramas y boxplots de los conjuntos de datos estandarizados para n = 1, 2 y 5.

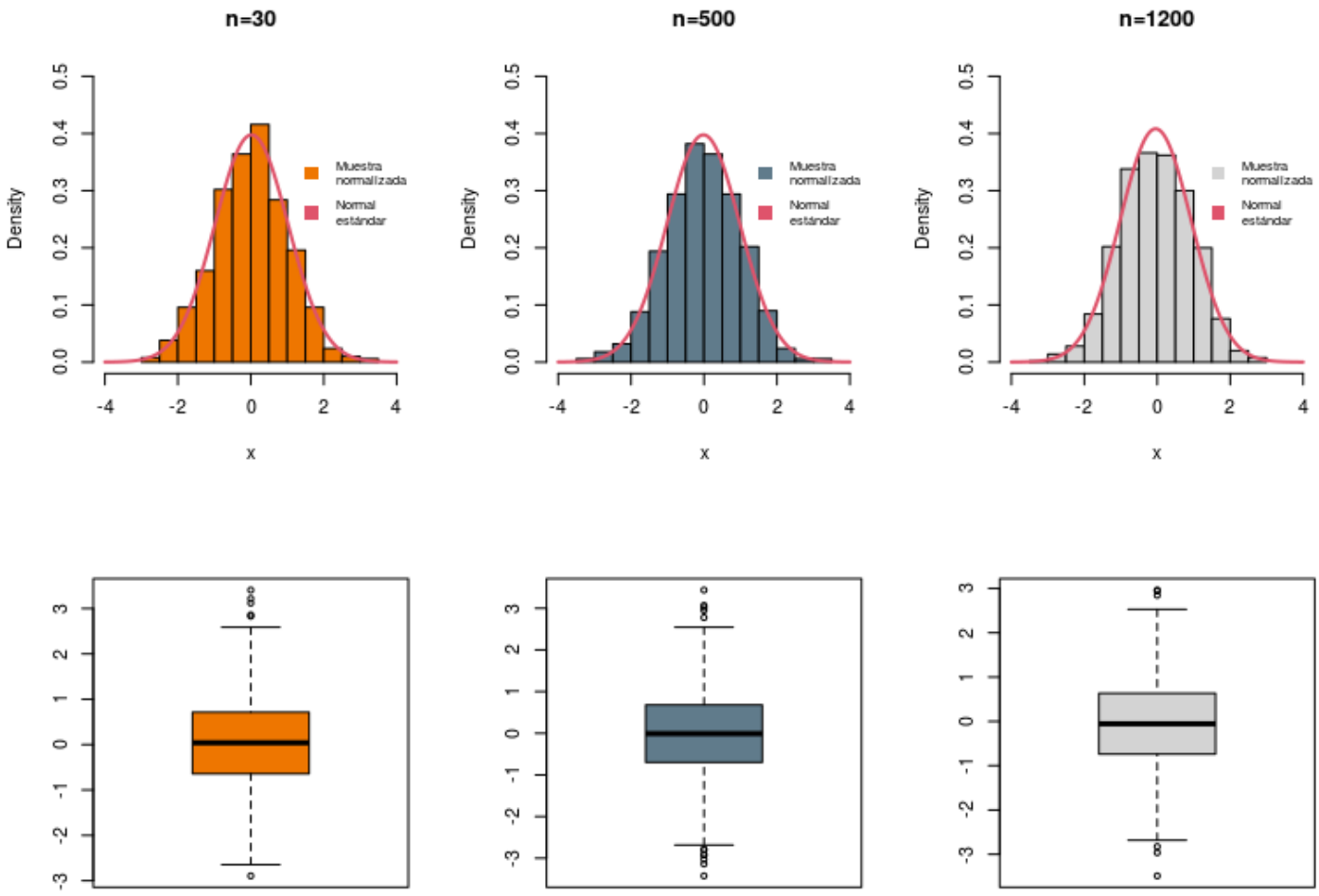


Figure 7: Histogramas y boxplots de los conjuntos de datos estandarizados para n = 30, 500 y 1200.

Podemos observar que los promedios estandarizados tienden a distribuirse como normales estándar, es decir, centradas en el cero y con varianzas muestrales muy cercanas a uno (ver Tabla 3). Esto es razonable ya que el Teorema Central del Límite (TCL) afirma que la distribución de los promedios de n variables aleatorias i.i.d. estandarizados, se parece a una distribución normal estandar a medida que n aumenta. En este caso en particular los resultados del TCL comienzan a notarse con valores de n bajos.

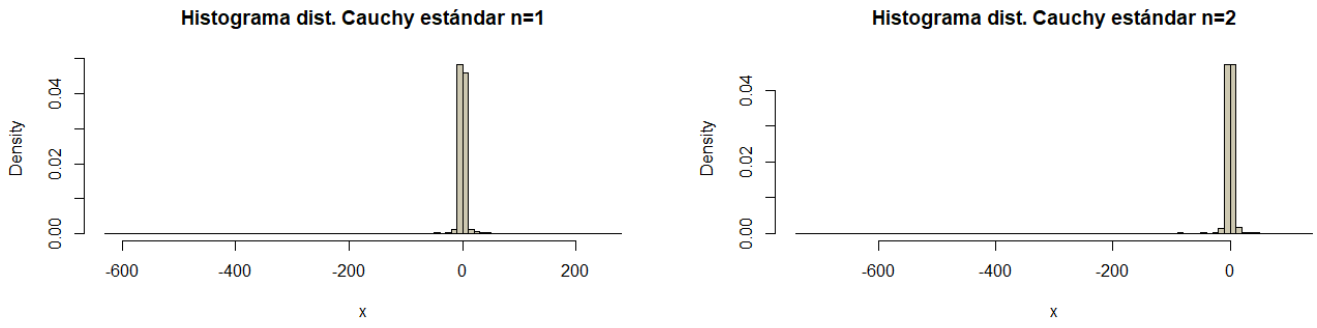
	n=1	n=2	n=5	n=30	n=500	n=1200
Varianza	1.0059414	1.0007842	0.9847698	1.0067025	1.0537559	0.9551179

Table 3: Varianzas de los promedios estandarizados.

Segunda Parte

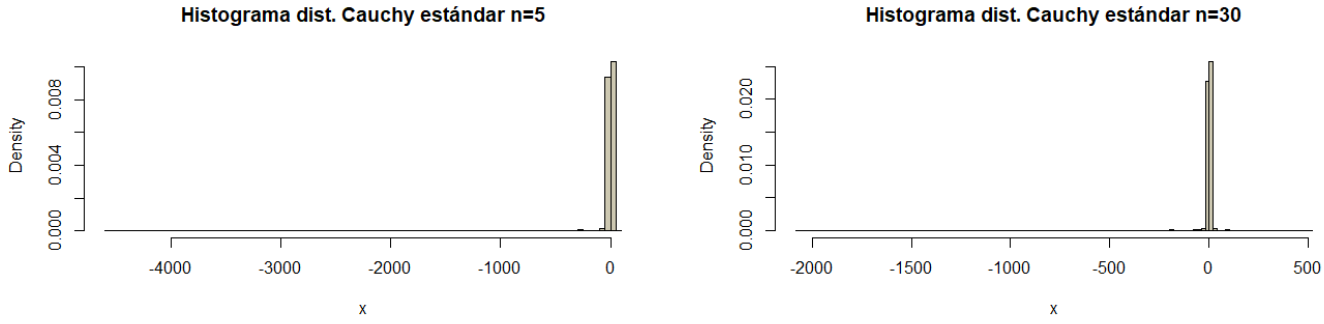
2. Repetimos el mismo procedimiento de la primera parte para variables aleatorias con distribución Cauchy $C(0,1)$ (también denominada Cauchy estándar). Esta densidad es simétrica alrededor del cero, con colas que acumulan más probabilidad que la normal estándar. Además, esta distribución tiene una característica particular, ya que no posee esperanza ni varianza finitas. Tiene moda y mediana bien definidas, iguales a cero en este caso.

   En la Figura 8 podemos ver los histogramas para mil repeticiones de una v.a. $C(0,1)$ (a) y para promedios de dos variables aleatorias $C(0,1)$ (b). En todos casos se evidencia que los promedios se centran en cero. Calculamos para todos los conjuntos de datos el rango intercuartil, medida de dispersión que nos permite ver en qué rango se encuentra el 50% de los datos centrales (ver Tabla 4). El IQR para n=1 es igual a 1.99883, mientras que para n=2 esta medida es igual a 1.92643.



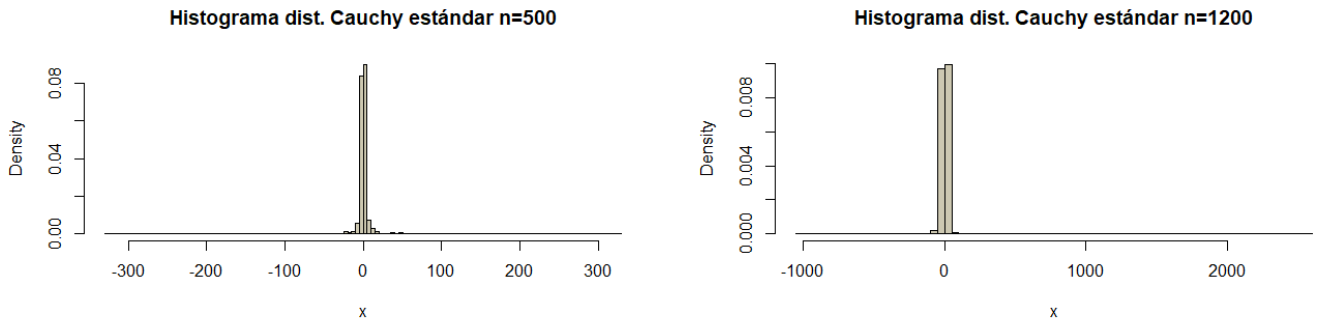
(a) n=1 (b) n=2

Figure 8: Histogramas de una distribución Cauchy estándar para distintos números de variables aleatorias.



(a) n=5 (b) n=30

Figure 9: Histogramas de una distribución Cauchy estándar para distintos números de variables aleatorias.



(a) n=500

(b) n=1200

Figure 10: Histogramas de una distribución Cauchy estándar para distintos números de variables aleatorias.

En las Figuras 9 y 10 se presentan los histogramas para n mayores. Se evidencia que hay datos muy alejados de cero (con muy baja densidad). Estos valores atípicos o outliers se pueden visualizar mejor observando los boxplots de los promedios (Fig. 11).

Si se cumpliera la Ley de los Grandes Números, esperaríamos que a medida que aumentamos n, los promedios tiendan a concentrarse, disminuyendo así el rango intercuartil. Sin embargo, esto no se observa, ya que el IQR no se reduce al aumentar n (ver Tabla 4).

	n=1	n=2	n=5	n=30	n=500	n=1200
IQR	1.998830	1.926430	1.916762	2.001125	1.961860	2.131936

Table 4: IQR de los promedios de v.a. $C(0,1)$

Como mencionamos anteriormente, esta distribución no posee esperanza ni varianza finitas. Calculamos las varianzas muestrales para poder evidenciar esta particularidad y efectivamente no disminuyen ni parecen converger cuando aumenta n (ver Tabla 5).

	n=1	n=2	n=5	n=30	n=500	n=1200
Varianzas muestrales	610.9684	757.9958	21237.8071	5288.4891	439.7285	7707.2920

Table 5: Varianzas muestrales los promedios de v.a. $C(0,1)$

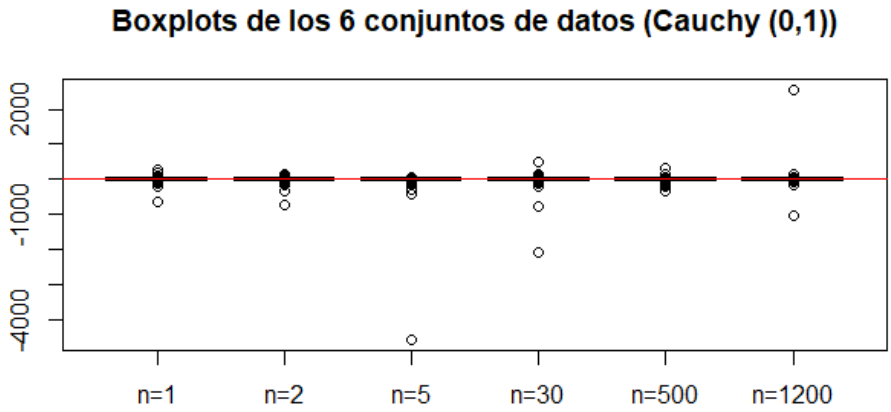


Figure 11: Boxplots de los 6 conjuntos de datos (promedios de v.a. Cauchy(0,1)).

En el boxplot de los seis conjuntos de datos podemos además observar que la mediana de los promedios (linea roja) es cero. Esto tiene sentido ya que el promedio de variables aleatorias $C(0,1)$ también tiene distribución $C(0,1)$.

Al no cumplir esta densidad con las hipótesis necesarias (varianza y esperanza finitas) no es posible realizar la estandarización de los promedios encontrados.

Por consiguiente podemos concluir que los datos evidencian que la distribución de Cauchy no cumple con la Ley de los Grandes Números ni el Teorema Central del Límite.