DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

# Simulación en R aplicada a la Ley de los Grandes Números y Teorema Central del Límite

Probabilidad y estadística


| Integrante | LU | Correo electrónico |
|---|---|---|
| Florencia Zanollo | 934/11 | florenciazanollo@gmail.com |
| Christian Murga | 982/12 | christianmmurga@gmail.com |

# Índice

# Introducción

En este trabajo trataremos de validar empíricamente los resultados de la Ley de los Grandes Números y el Teorema Central del límite. Para ello vamos a realizar una serie de simulaciones utilizando el lenguaje R. Además graficaremos nuestros resultados utilizando la librería ggplot2.

En la primer sección simularemos una distribución Uniforme y tomaremos el promedio de variables aleatorias independientes e idénticamente distribuidas; observando y comparando cada uno de los resultados utilizando diferentes técnicas como histogramas y boxplots. Analizaremos su comportamiento al incrementar la cantidad de variables promediada.

En la segunda sección intentaremos hacer lo mismo con la distribución Cauchy.

Los experimentos fueron ejecutados a partir de la semilla: 73284.

# 1. Distribución Uniforme

Hicimos múltiples simulaciones aumentando la cantidad de variables aleatorias de distribución U(0,1) promediadas; [1, 2, 5, 30, 500, 1200] las cuáles llamaremos simulación A a F respectivamente. A continuación mostraremos los estadísticos de las observaciones obtenidas en cada una.

A

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.0009489 | 0.2613544 | 0.5200669 | 0.5088680 | 0.7688815 | 0.9996037 |

B

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.01307 | 0.38364 | 0.51752 | 0.51381 | 0.65900 | 0.96787 |

C

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.1190 | 0.4000 | 0.4926 | 0.4933 | 0.5912 | 0.8425 |

D

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.3376 | 0.4647 | 0.5007 | 0.5002 | 0.5351 | 0.6719 |

E

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.4591 | 0.4908 | 0.5003 | 0.4997 | 0.5086 | 0.5365 |

F

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.4715 | 0.4944 | 0.4999 | 0.4998 | 0.5051 | 0.5285 |

En todas la media y mediana se encuentra cerca de la teórica ($\mu = 0,5$). La distancia intercuartíl y la diferencia entre sus máximos y mínimos disminuye a medida que se aumenta la cantidad de promediadas; esto es acorde a lo dictado por la Ley de los Grandes Números.

En cuanto a la varianza muestral:

A: 0.08381345

B: 0.0407693

C: 0.01728931

D: 0.002557235

E: 0.0001706746

F: 6.666837e-05

Podemos ver que se al comienzo vale casi exactamente la varianza teórica de una U(0,1), es decir, $\frac{(1-0)^2}{12} = 0,8$. Pero a medida que crece el n se acerca a 0.

Lo que sigue en este capítulo son una serie de gráficos que resaltan diferentes características de las observaciones. Con esto esperamos mostrar de una manera más visual lo que se observa en estas simples tablas.

## 1.1. Histogramas

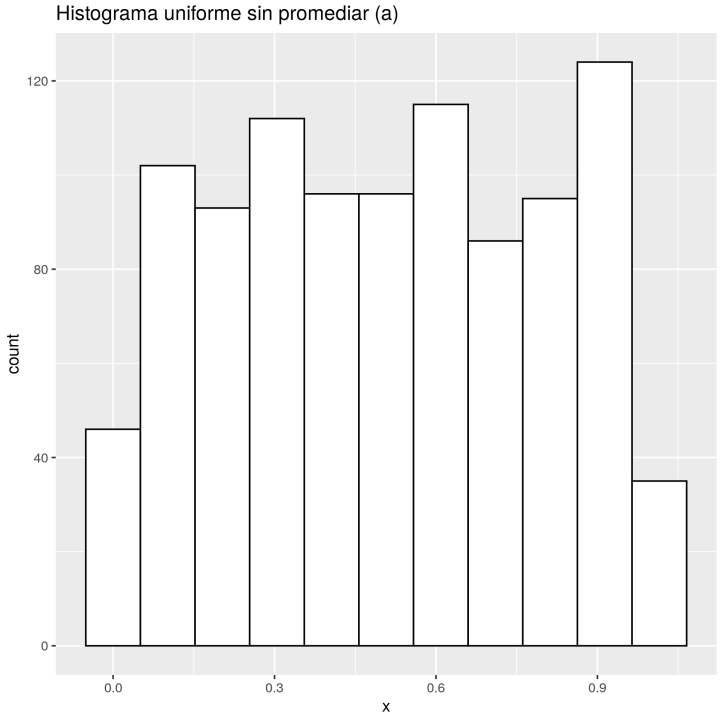


Figura 1: Histograma para simulación A, uniforme

El histograma (Fig.1) se parece al de una uniforme, podemos ver que casi todas las barras tienen una altura similar. Esto es esperable ya que si tomamos una cantidad alta de muestras deberíamos ver que los resultados se acerquen cada vez mas a la distribución original. *La probabilidad experimental tiende a la probabilidad teórica.*

Ahora bien, a medida que aumentamos la cantidad de variables aleatorias consideradas los promedios se aproximan cada vez más a la media (0,5) como podemos ver en Fig.5. Esto concuerda con lo propuesto por la Ley de los Grandes Números que dicta: *si* $X1, X2, X3...$ *es una sucesión infinita de variables aleatorias independientes que tienen el mismo valor esperado* $\mu$ *y varianza* $\sigma^2$*, entonces el promedio de las variables aleatorias converge en probabilidad a* $\mu$*.*

Por otro lado también podemos observar que se va pareciendo más a una distribución normal que una uniforme. Esto muestra empíricamente el Teorema Central del Límite que indica: *si* $X_n$ *es la suma de n variables aleatorias independientes, idénticamente distribuidas; con valor esperado* $\mu$ *y de varianza* $0 < \sigma^2 < \infty$ *entonces la función de distribución de* $X_n$ *se "aproxima bien" a una distribución normal, el teorema asegura que esto ocurre cuando el n es lo suficientemente grande.* Es decir, a medida que aumentamos el n más se va a parecer a una normal.

Para poder comparar correctamente los histogramas de los distintos conjuntos de datos sería necesario tenerlos dibujados en la misma escala tanto para el eje horizontal como para el vertical. Por eso, en general es más cómodo hacer boxplots para comparar distintos conjuntos de datos.

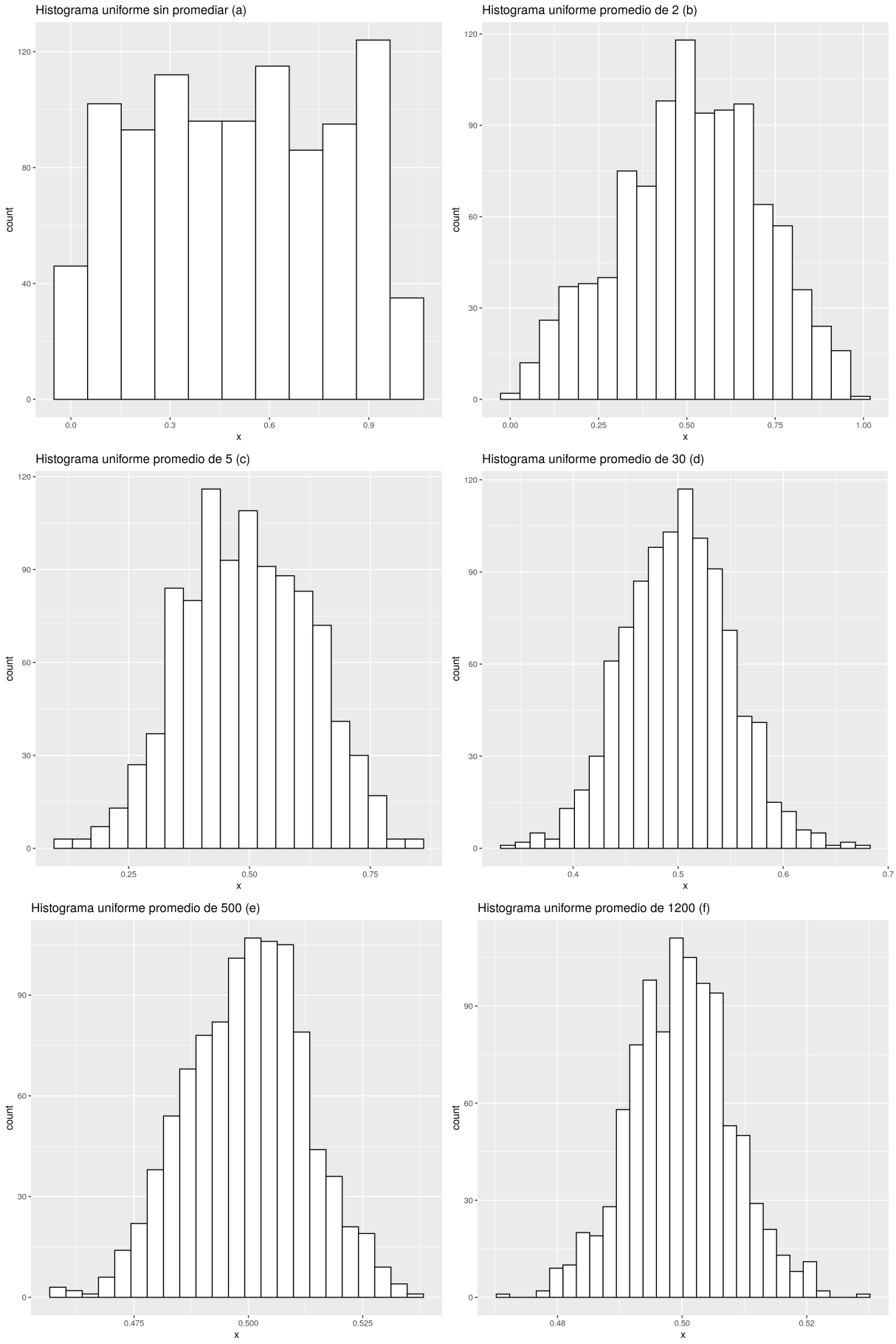


Figura 2: Histogramas para las diferentes simulaciones uniforme

## 1.2. Boxplots

En los boxplots (Fig.3) podemos ver que en todos la mediana se mantiene alrededor del mismo valor (0,5) también se aprecia que el rango intercuartílico disminuye notablemente a medida que aumentamos la cantidad de variables promediadas, *o sea disminuye la dispersión acumulándose más puntos cerca de la mediana.*

Esto último se puede apreciar en los "bigotes"que se muestran a los costados de cada box. Además en los casos de d, e y f podemos observar *outliers* (datos atípicos) que quedaron como tal debido a la poca distancia intercuartíl que tienen.

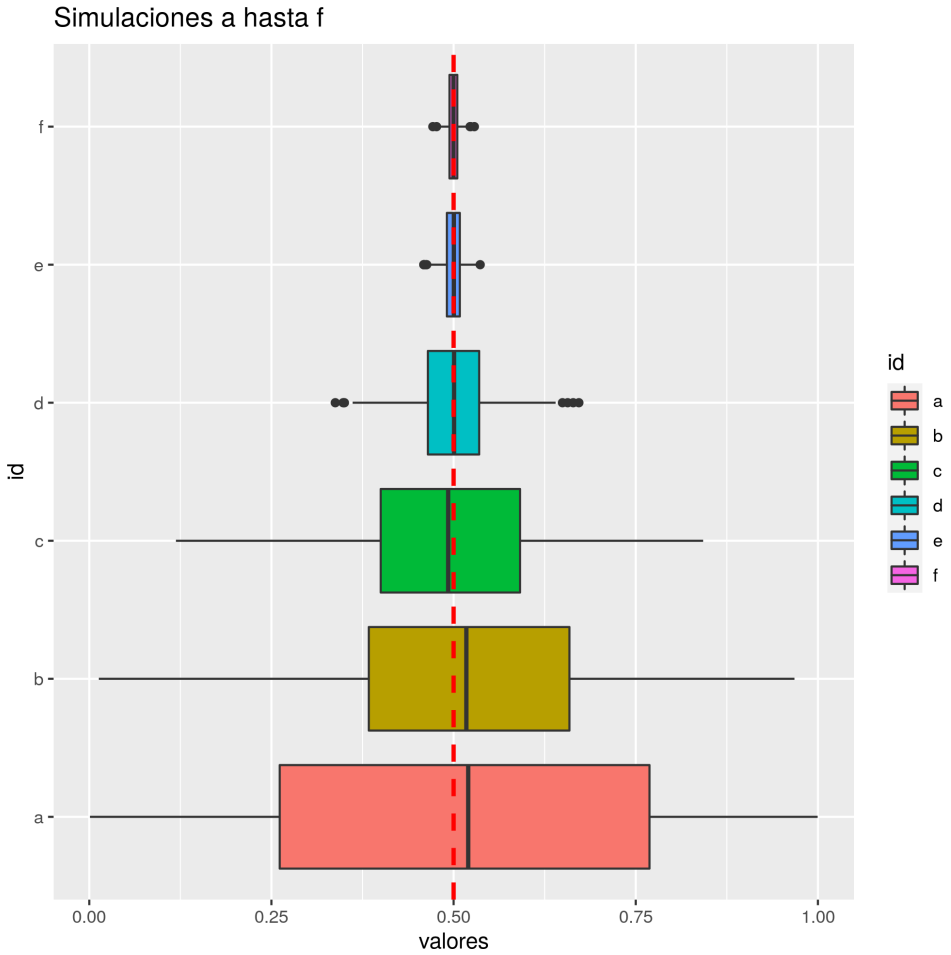


Figura 3: Boxplots paralelos para todas las simulaciones uniforme

## 1.3. QQplots

El QQplot nos ayuda a comparar 2 distribuciones, en este caso comparamos nuestras distribuciones con la normal estándar (N(0,1)), en la Fig.4 podemos ver que en los primeros gráficos (con pocas variables aleatorias) solo algunos puntos caen sobre la linea de la normal pero a medida que aumentamos la cantidad de variables se acercan más a tal punto que en el último casi todos están sobre la recta, esto nos indica que nuestra distribución esta aproximadamente normalmente distribuida.

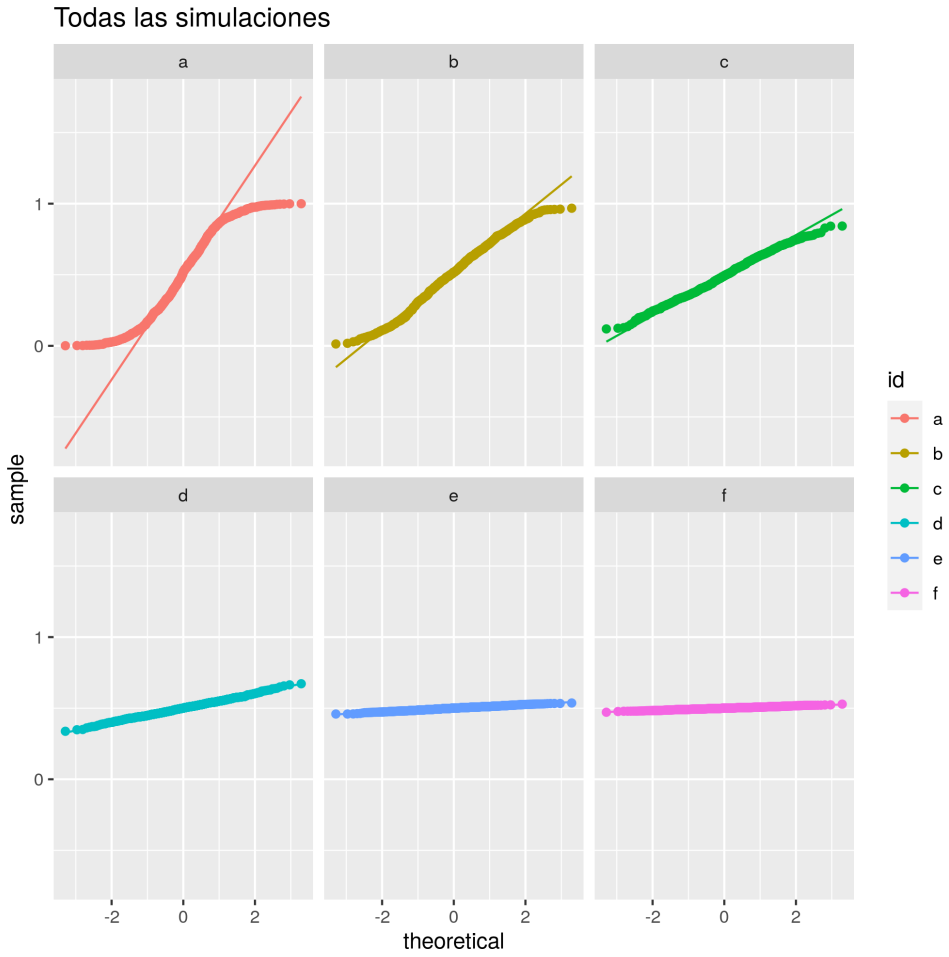


Figura 4: QQplots para todas las simulaciones uniforme

## 2. Uniforme Normalizada

El teorema Central del Límite dice, además, que si realizamos la siguiente transformación

$$\frac{\bar{X}_n - E(X_1)}{\sqrt{\frac{Var(X_1)}{n}}} \tag{1}$$

sobre los promedios la distribución se aproximará no sólo a una normal sino a la estándar, siempre contando con un n suficientemente grande. Por consiguiente aplicando dicha transformación sobre los datos y volviendo a observarlos esperamos que se parezcan a una distribución normal estándar.

A

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -54.6683 | -26.1423 | 2.1982 | 0.9714 | 29.4545 | 54.7289 |

B

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -53.341 | -12.746 | 1.919 | 1.513 | 17.418 | 51.252 |

C

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -41.7415 | -10.9534 | -0.8143 | -0.7345 | 9.9902 | 37.5202 |

D

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -17.78587 | -3.86367 | 0.07149 | 0.02444 | 3.84839 | 18.83334 |

E

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -4.48158 | -1.00845 | 0.03763 | -0.03676 | 0.93761 | 3.99784 |

F

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -3.11745 | -0.61496 | -0.01307 | -0.01912 | 0.55826 | 3.11979 |

Al ver los estadísticos de los datos normalizados ya notamos cómo para n=1 no se parece en nada a una normal estándar pero, a medida que aumenta el n, la media se acerca a la teórica ($\mu = 0$). Además si miramos la varianza muestral:

A: 1005.761

B: 489.2316

C: 207.4717

D: 30.68682

E: 2.048095

F: 0.8000204

También para n=1 (A) está muy lejos (¡y es muy alta!) pero va bajando de a poco a medida que aumenta el n y en n=1200 (F) notamos cómo se acerca a la teórica ($\sigma^2 = 1$).

Entonces para poder comparar visualmente con los resultados de la primer sección realizamos la misma serie de gráficos, los cuáles mostramos a continuación.

## 2.1. Histogramas

Claramente podemos ver que van acercándose a una simetría con respecto al 0. En el último caso cayendo en su mayoría dentro de [-2, 2]. El gráfico se parece cada vez más al de la distribución normal estándar.

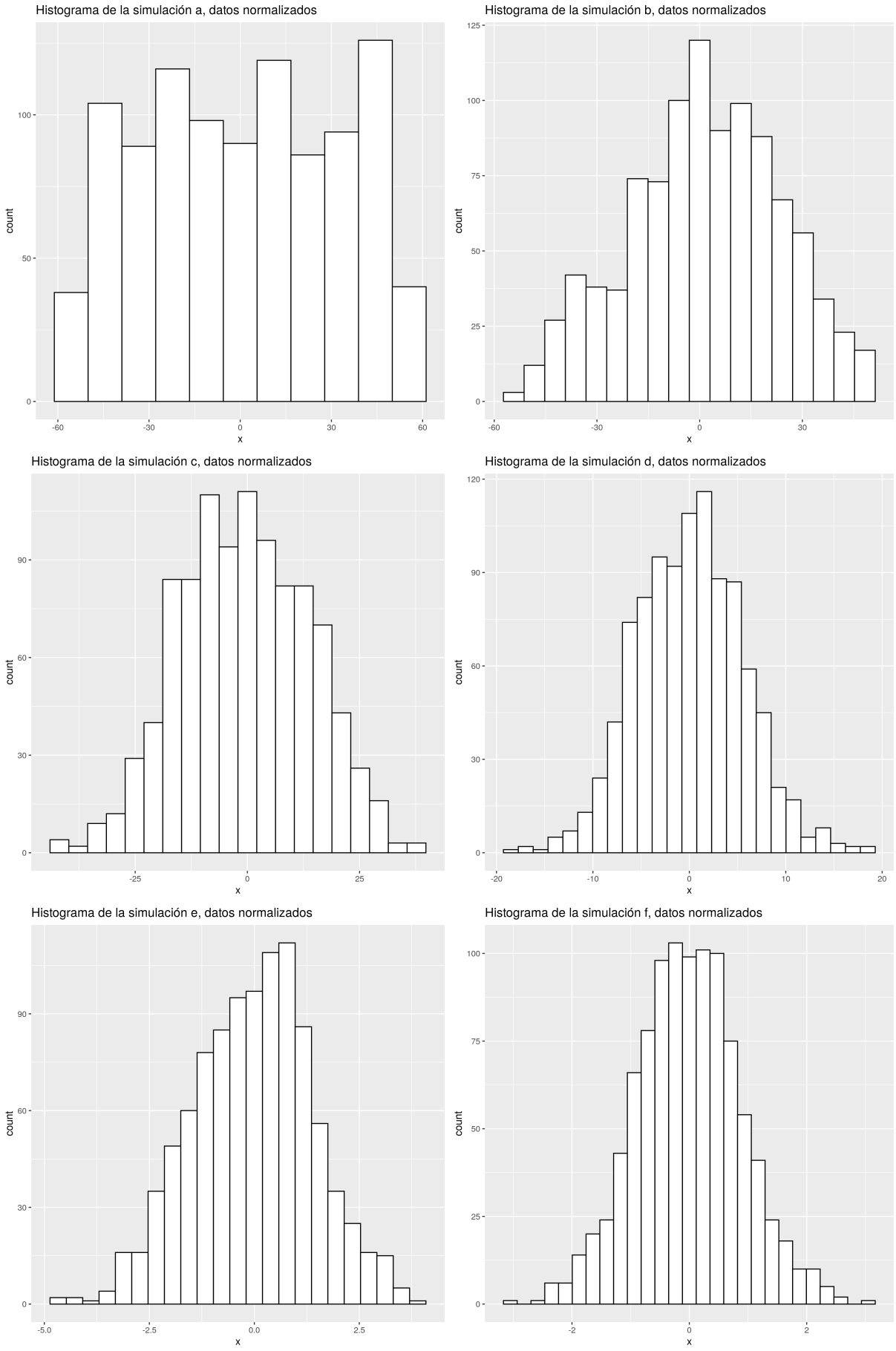


Figura 5: Histogramas para las diferentes simulaciones uniforme normalizada

## 2.2. Boxplots

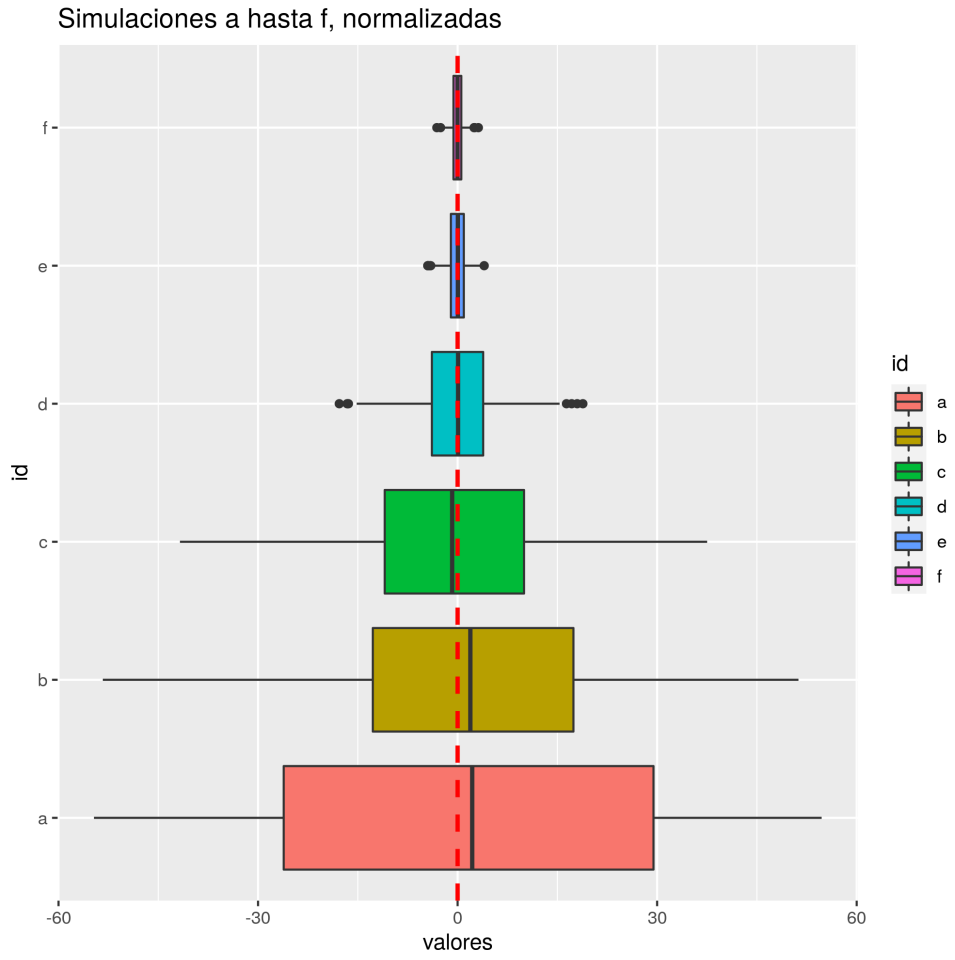


Figura 6: Boxplots paralelos para todas las simulaciones uniforme normalizada

Podemos ver claramente cómo lo dicho en la primer sección se repite para las diferentes simulaciones pero esta vez es respecto de una normal estándar, es decir, la simetría se sigue observando pero ahora la media tiende a 0.

# 3. Distribución Cauchy

Hicimos múltiples simulaciones aumentando la cantidad de variables aleatorias de distribución U(0,1) promediadas; [1, 2, 5, 30, 500, 1200] las cuáles llamaremos simulación A a F respectivamente. A continuación mostraremos los estadísticos de las observaciones obtenidas en cada una.

A

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -923.5973 | -1.1159 | -0.0092 | -1.4078 | 1.0249 | 107.9371 |

B

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -718.1324 | -1.0250 | -0.0365 | -0.5774 | 0.8101 | 688.4354 |

C

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -514.8324 | -1.0809 | -0.0855 | -0.2199 | 0.8246 | 365.6574 |

D

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -308.5902 | -1.1112 | -0.0390 | -0.0200 | 0.9521 | 205.8355 |

E

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -163.2997 | -0.8477 | 0.0570 | -0.2614 | 1.1490 | 327.8878 |

F

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -1553.5773 | -0.9580 | 0.0405 | -2.8507 | 1.0167 | 221.2646 |

## 3.1. Histogramas

Los primeros histogramas que realizamos no aportaban demasiada información (Fig.7) ya que tomábamos un ancho para los rectángulos de manera dinámica utilizando la distancia intercuartíl y por la naturaleza de los datos, que tienen cierta concentración alrededor del 0 pero con outliers muy groseros, su distancia intercuartíl es muy grande por ende el ancho resultaba poco y en la figura parece haber una distribución acampanada cuando no la hay.

Entonces decidimos tratar de graficar los histogramas de otra forma, agrandando el tamaño de los intervalos (ver Fig.8). Podemos ver que, similar a lo visto en la sección 1, los datos se acumulan alrededor de un valor esperado, en este caso 0; pero a diferencia de las otras secciones los outliers hacen que nunca obtengamos esa figura acampanada y simétrica.

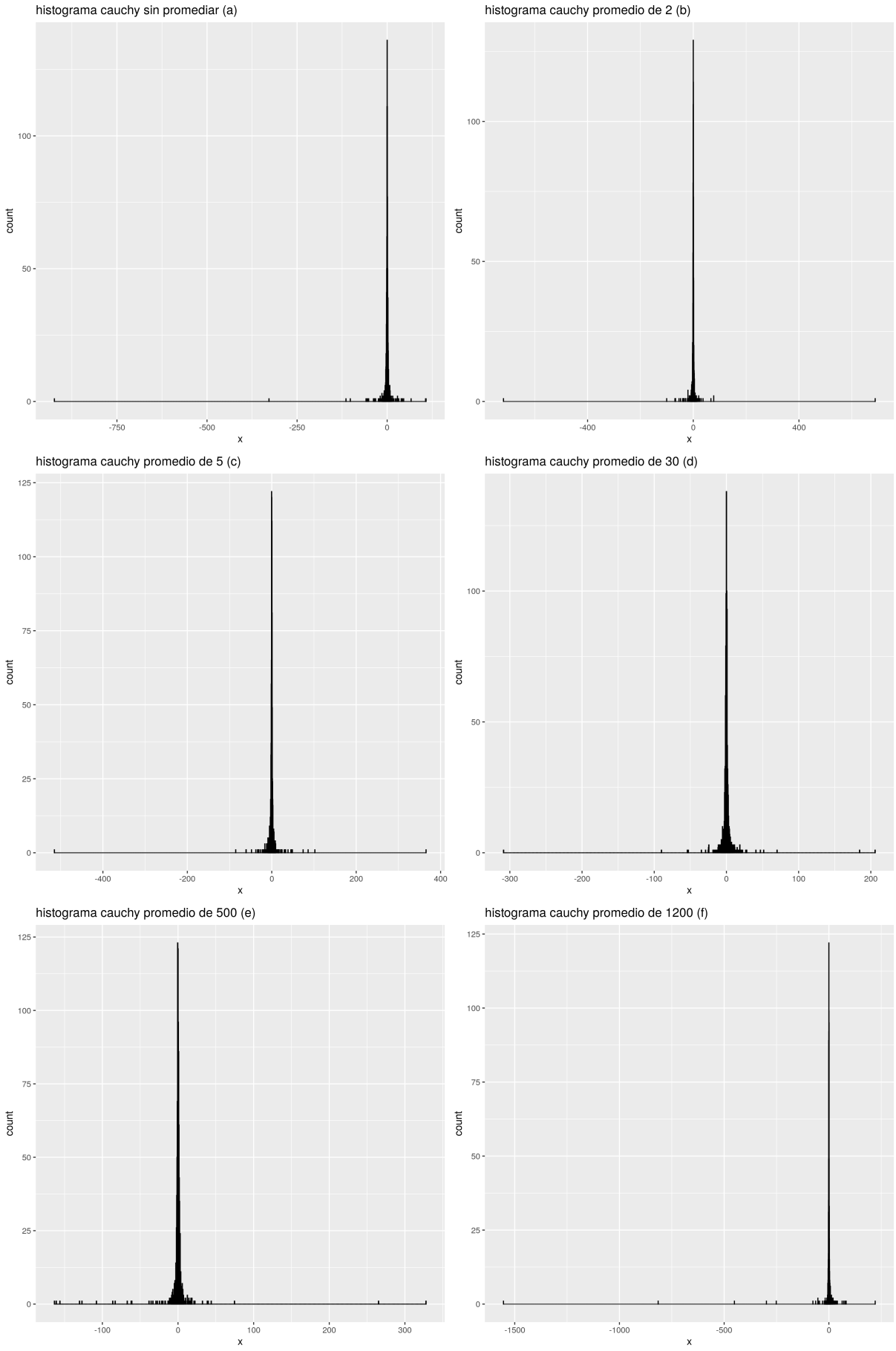


Figura 7: Histogramas para las diferentes simulaciones Cauchy V2

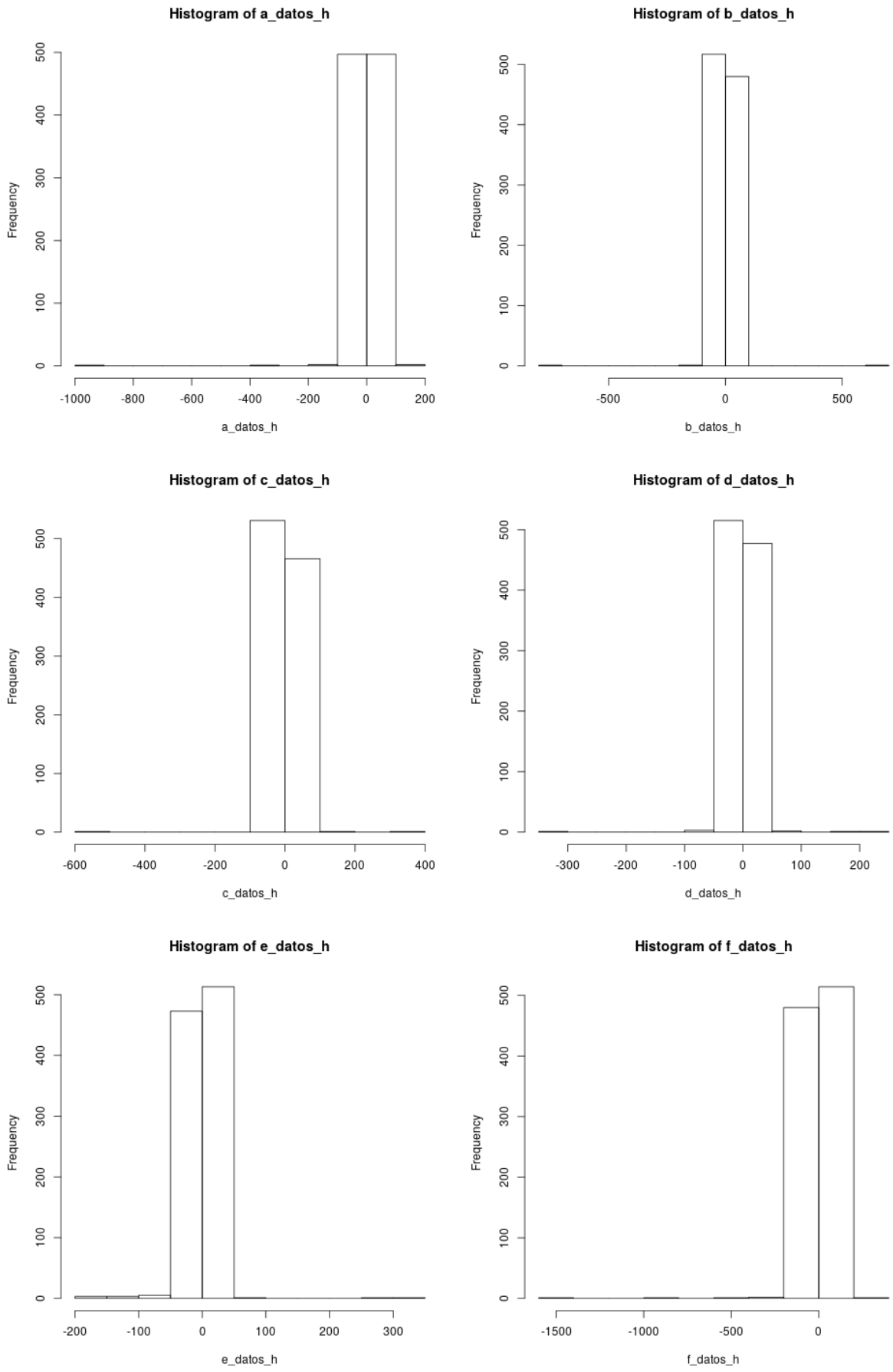


Figura 8: Histogramas para las diferentes simulaciones Cauchy

## 3.2. Boxplots

Si graficamos los boxplots simplemente observamos que no nos dan demasiada información debido a que hay muchos outliers y si hacemos un zoom en el intervalo donde se acumulan una mayor densidad (intervalo: [-50,50]) se aprecia que estas distribuciones se mantienen muy dispersas aún cuando aumentamos la cantidad de variables aleatorias. Podemos ver que igual que las anteriores la mediana se mantiene mantiene alrededor del mismo valor sin embargo en estas el rango intercuartiílico se mantiene constante igual que los bigotes. Esto son indicadores de que este tipo de distribución no cumple con el Teorema Central del Límite.

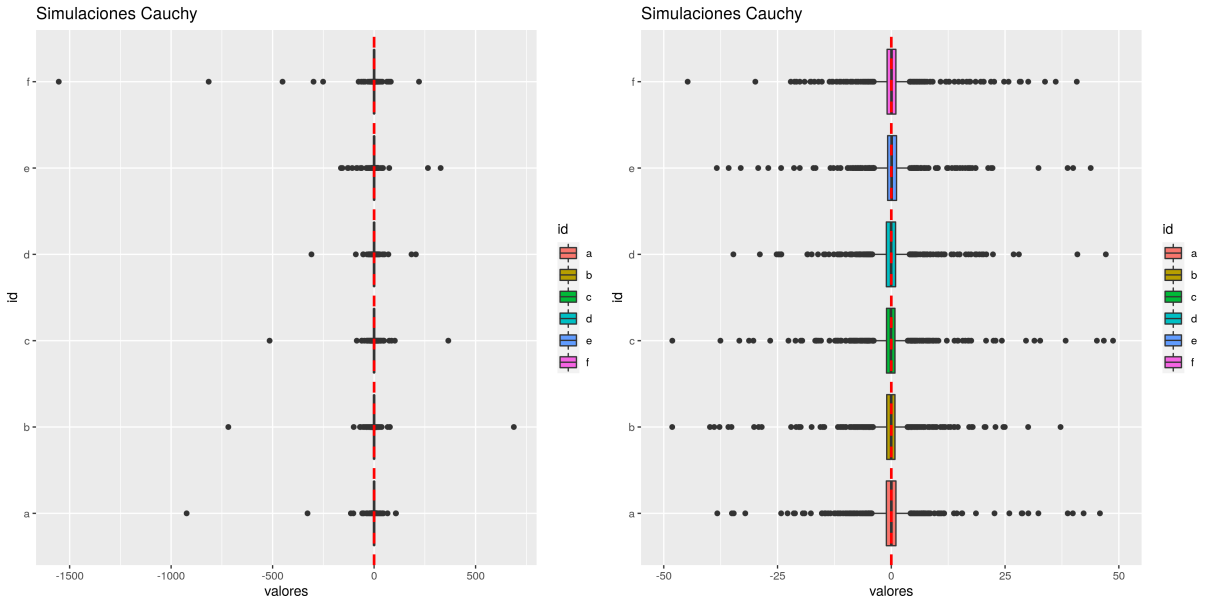


Figura 9: Boxplots paralelos para todas las simulaciones Cauchy

## 3.3. QQplots

Como en las secciones anteriores comparamos nuestras distribuciones con la normal estándar (N(0,1)), en la Fig.10 podemos ver que en este caso todas las distribuciones se comportan similar respecto de la normal estándar. Lo que nos indica que aún aumentando la cantidad de variables aleatorias promediadas nuestra distribución parece seguir comportandose como una cauchy.

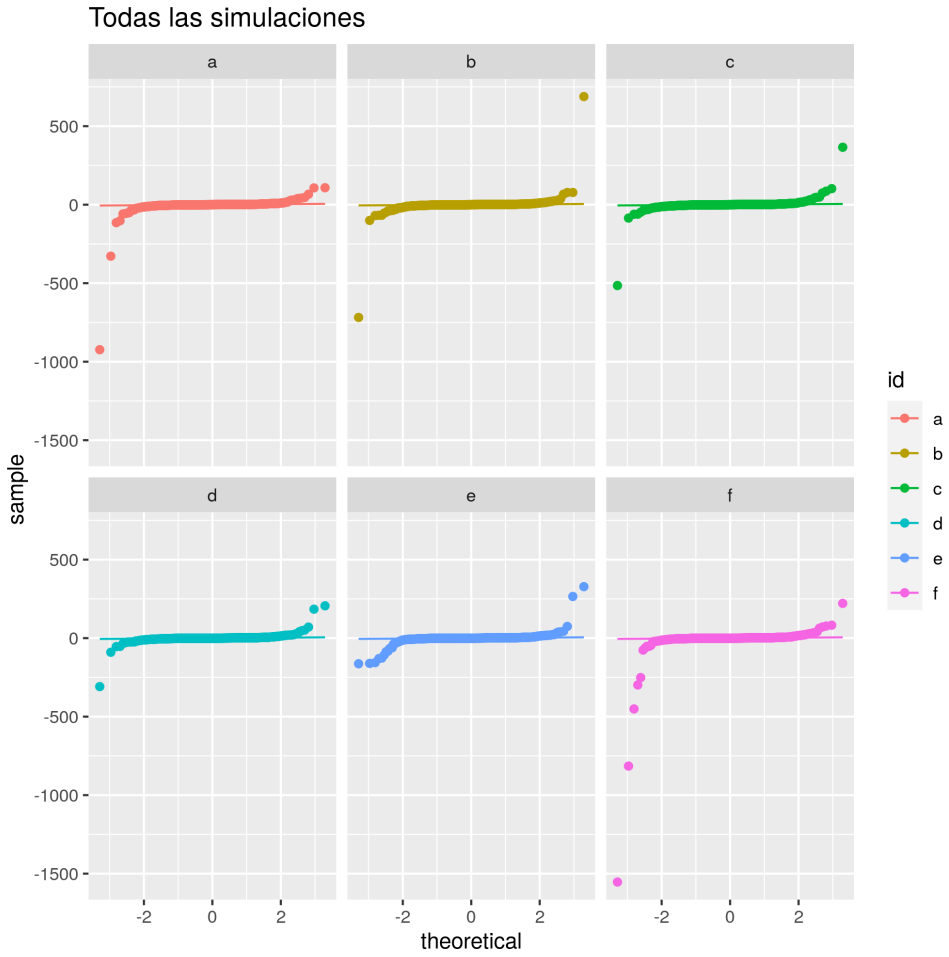


Figura 10: QQplots para todas las simulaciones cauchy

# 4. Conclusiones

En las primeras secciones comprobamos empíricamente la Ley de los grandes números y el Teorema central del límite. Esto resultó como esperábamos dado que las distribuciones utilizadas tienen esperanza y varianza finitas y no nulas.

En la tercer sección, por el contrario, mostramos un caso donde esto no se cumple entonces no se pueden utilizar. Cauchy es una distribución particular, que no tiene esperanza y varianzas finitas. Si analizamos la convergencia podemos ver que no converge a una normal sino a otra Cauchy. Esto se notó fácilmente en los gráficos y valores observados.

Personalmente aprendimos a interpretar los distintos gráficos y comprender su utilidad, ya que muestran propiedades muy diferentes de los datos.