Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Simulación Estadística
Cali, mayo de 2018

TEOREMA CENTRAL DEL LIMITE PRUEBAS DE NORMALIDAD.

Kevin Steven García[a], Cesar Andres Saavedra[b]

1. TEOREMA CENTRAL DEL LIMITE.

El teorema del límite central es un teorema fundamental de probabilidad y estadística. El teorema describe la distribución de la media de una muestra aleatoria proveniente de una población con varianza finita. Cuando el tamaño de la muestra es lo suficientemente grande, la distribución de las medias sigue aproximadamente una distribución normal. El teorema se aplica independientemente de la forma de la distribución de la población. El teorema de límite central le permite aplicar estos procedimientos útiles a poblaciones que son considerablemente no normales. El tamaño que debe tener la muestra depende de la forma de la distribución original. Si la distribución de la población es simétrica, un tamaño de muestra de 5 podría producir una aproximación adecuada. Si la distribución de la población es considerablemente asimétrica, es necesario un tamaño de muestra más grande.

1.1. Teorema del límite central.

Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n obtenida desde una población con cualquier distribución con media μ_X y varianza σ_{X^2}, entonces cuando n es grande, $\bar{X}$ sigue una distribución normal con $\mu_{\bar{X}} = \mu_X$ y $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$

Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n que se toma de una población $N(\mu_X, \sigma^2_X)$ conocida, entonces:

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_x}{\sqrt{n}}}$$

Entonces Z se aproxima a una variable normal N(0,1), mejorándose la calidad de la aproximación a medida que el n aumenta.

- El teorema del límite central garantiza una distribución aproximadamente normal cuando n es suficientemente grande.
- Existen diferentes versiones del teorema, en función de las condiciones utilizadas para asegurar la convergencia. Una de las más simples establece que es suficiente que las variables que se suman sean independientes, idénticamente distribuidas, con valor esperado y varianza finitas.
- La aproximación entre las dos distribuciones es, en general, mayor en el centro de las mismas que en sus extremos o colas, motivo por el cual se prefiere el nombre "teorema del límite central".

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1628466. E-mail: cesar.saavedra@correounivalle.edu.co

2. VARIABLES ALEATORIAS.

2.1. Introducción.

En esta sección se evidencian los resultados de las simulaciones de las distribuciones de probabilidad para tamaños de muestra y distintos parámetros, así poder observar su comportamiento y dar respuesta a los siguientes interrogantes:

- Para que valor de n se puede afirmar el cumplimiento del teorema.
- La distribución de probabilidad de X afecta la convergencia.
- Los parámetros de la distribución de probabilidad de X afecta la convergencia.

El teorema referenciado en el numeral 1 del presente documento se llevara a cabo haciendo uso de dos distribuciones, donde: la distribución Poisson y la distribución Logística fueron simuladas por medio de las funciones ya incluidas en el programa estadístico r. Teniendo en cuenta estas simulaciones se aplico la prueba Cramér-Von Mises y se comparó los resultados para así saber si las simulaciones para los distintos tamaños de muestra y parámetros de verdad cumplen con el Teorema Central Del Limite y saber si estas convergen o no a la distribución normal. Posteriormente, se concluye acerca de los métodos de simulación usados teniendo en cuenta los resultados de la prueba.

2.2. Descripción general de las distribuciones y las pruebas de bondad de ajuste.

2.2.1. Distribución Poisson.

Esta distribución es una de las más importantes distribuciones de variable discreta. Sus principales aplicaciones hacen referencia a la modelización de situaciones en las que interesa determinar el número de hechos de cierto tipo que se pueden producir en un intervalo de tiempo o de espacio, bajo presupuestos de aleatoriedad. Su función de densidad está dada por:

$$f(x,\lambda) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}; x \in \{0,1,2,3,...\}$$

2.2.2. Distribución Logística.

La función de distribución de la logística se utiliza como modelo de crecimiento. Por ejemplo, con un nuevo producto a menudo encontramos que el crecimiento es inicialmente lento, entonces gana impulso, y finalmente se ralentiza cuando el mercado está saturado o se alcanza alguna forma de equilibrio. Esta distribución tiene uso en los modelos logit, ya que estos modelos trabajan con su función de distribución acumulada denominada función logistica, por la simpleza de esta para encontrar probabilidades. Su función de distribución acumulada está dada por:

$$F(x,a,b) = \frac{1}{1+e^{-(\frac{x-a}{b})}}$$

2.2.3. Prueba Cramer-Von Mises.

La distribución normal es utilizada en muchos campos ya que es el supuesto básico de algunas herramientas estadísticas tales como las pruebas t-student, Chi-Cuadrado, ANOVA, entre otros. Esto hace que muchos fenómenos puedan ser modelados mediante ella, lo cual promueve el desarrollo de distintas pruebas estadísticas que se encuentran en la literatura de las cuales se destacan las de Pearson (Pearson 1900), **Cramer-Von Mises (Cramer 1928 y Mises 1947)**, Kolmogorov-Smirnov (Lilliefors 1967), Anderson-Darling, (Anderson y Darling 1954) Shapiro-Wilk, (Shapiro y Wilk 1965) Shapiro-Francia, (Shapiro y Francia 1972) Jarque-Bera (Jarque y Bera 1987) y D'Agostino (D'Agostino 1972), de las cuales veremos la aplicación de una de ellas en el presente informe que tiene como fin verificar el Teorema Central del Limite y establecer bajo que condiciones especificas se presenta convergencia con la distribución Normal (μ,σ^2) a partir de la simulación de variables aleatorias.

Para los cuales esta prueba plantea las hipotesis:

$$H_0 : f(x, \theta) = f_0(x, \theta)$$

$$H_1 : f(x, \theta) \neq f_0(x, \theta)$$

Para la cual se tiene como estadístico de prueba a:

$$W = \frac{1}{12n} + \sum_{i=1}^{N} [P_i - \frac{2_i - 1}{2n}]$$

3. SIMULACIÓN DE VARIABLES ALEATORIAS Y PRUEBAS DE NORMALIDAD

3.1. Simulación de las observaciones:

3.1.1. Distribución Poisson:

Es importante recordar que los parámetros sobre los cuales se van a generar las observaciones son $(\lambda = 1)$, $(\lambda = 5)$, $(\lambda = 10)$, con 20000 simulaciones que representan la población para las cuales se sacaron muestras de tamaño $(n = 1)$, $(n = 10)$, $(n = 25)$, $(n = 50)$ datos. La generación de estos números se hizo por medio del comando rpois del software estadístico R.

3.1.2. Distribución Logística:

El uso de esta distribución en la problemática del articulo es una función derivada de ella, es decir, no se usa directamente la función de densidad logística para describir el fenómeno, se usa su función de distribución acumulada, denominada también función logística. Teniendo en cuenta lo anterior, se generaron datos con distribución Logística de parámetros $(a = 0, b = 1)$, $(a = 9, b = 4)$, $(a = 15, b = 6)$, con 20000 simulaciones que representan la población para las cuales se sacaron muestras de tamaño $(n = 1)$, $(n = 10)$, $(n = 25)$, $(n = 50)$ datos.

3.2. Prueba Cramer-Von Mises.

Uno de los criterios mas utilizados para comparar funciones es el introducido por los investigadores Harald Cramér y Richard Edler Von Mises y conocido como criterio de Cramér - Von Mises (CM) siendo aplicado a problemas de bondad de ajuste de una distribución como el que nos compete trabajar en este informe a través de la función **cvm.test** que nos permite obtener los valores-p calculando así la proporción de rechazo de la prueba.

Una de las grandes ventajas del test de Cramér-Von Mises para n muestras independientes radica en que su distribución no depende de la distribución de procedencia de las muestras. Por este motivo, para ilustrar la calidad de las aproximaciones propuestas, no es necesario considerar distintos modelos, por lo cual para nuestro escenario de simulación hemos propuesto tamaños de muestras $(n = 1)$, $(n = 10)$, $(n = 25)$, $(n = 30)$, $(n = 50)$ a partir de 20000 simulaciones que serán las que permitan hacer las aproximaciones y saber así bajo que condiciones se alcanza la convergencia de nuestras distribuciones (Poisson, Logística) a una distribución Normal $(\mu, \frac{\sigma^2}{n})$.

4. Resultados

4.0.1. Distribucción Poisson:

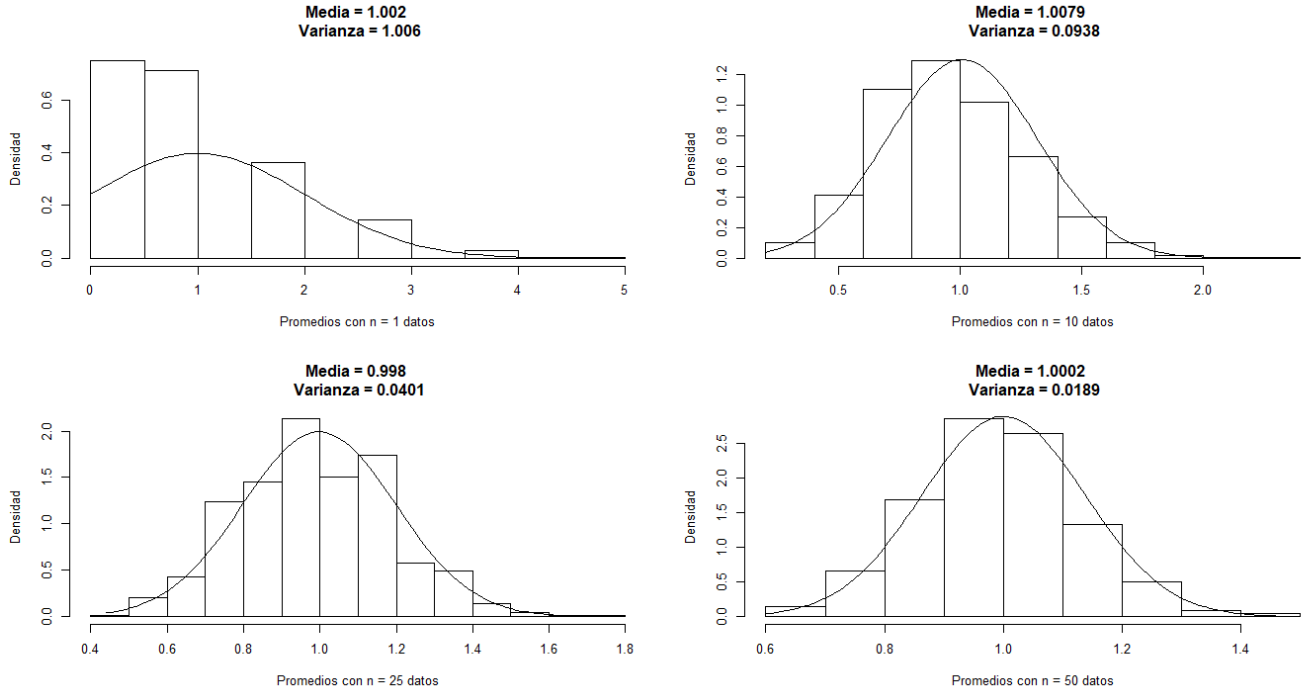

FIGURA 1: Simulaciones en R de la distribución Poisson($\lambda = 1$) con muestras de tamaño ($n = 1$), ($n = 10$), ($n = 25$), ($n = 50$))

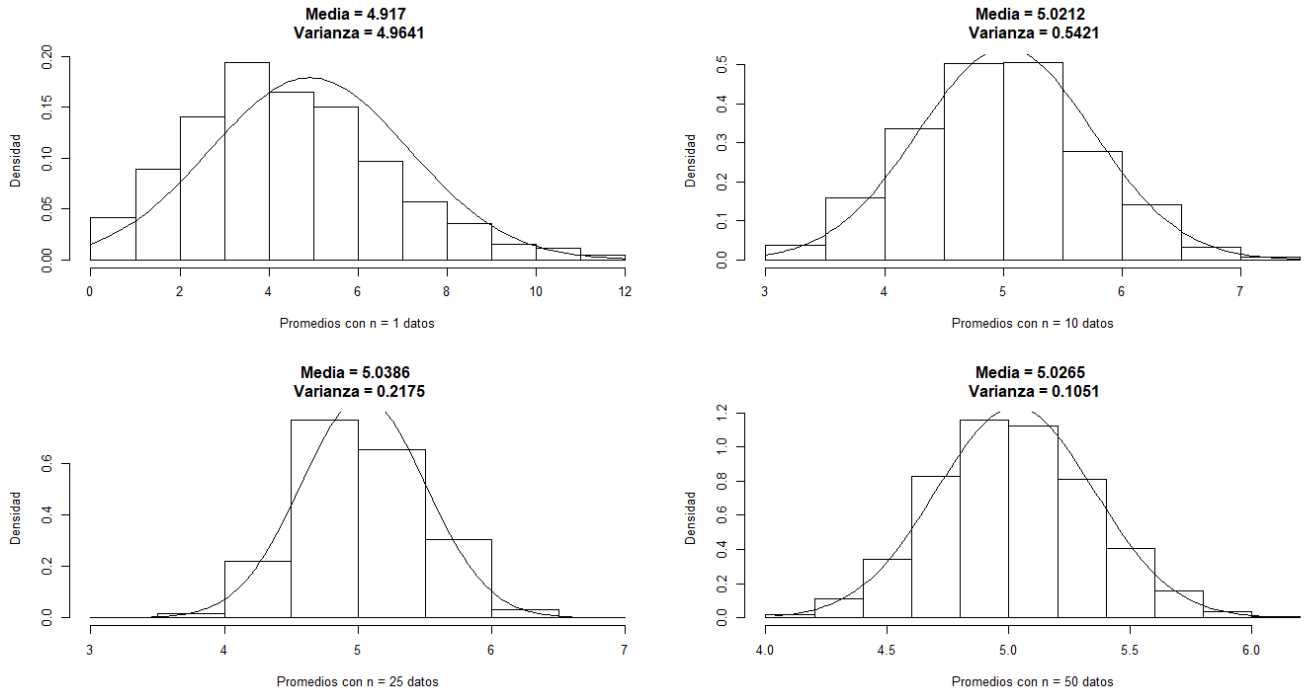

FIGURA 2: Simulaciones en R de la distribución Poisson($\lambda = 5$) con muestras de tamaño ($n = 1$), ($n = 10$), ($n = 25$), ($n = 50$)

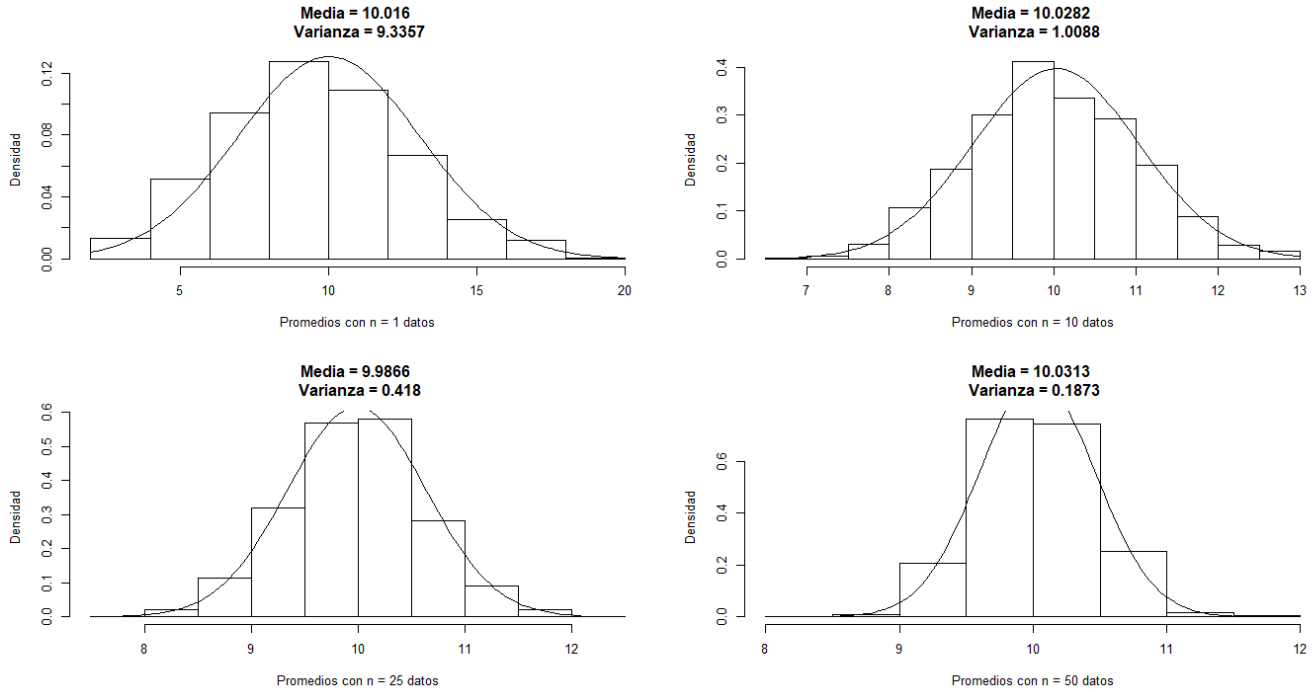

FIGURA 3: Simulaciones en R de la distribución Poisson($\lambda = 10$) con muestras de tamaño ($n = 1$), ($n = 10$), ($n = 25$), ($n = 50$)

4.0.2. Prueba Cramer-Von Mises Distribución Poisson.

	n=1		n=10		n=25		n=50	
Parametro	W	P-Valor	W	P-Valor	W	P-Valor	W	P-Valor
$\lambda = 1$	10.619	1.123e-10	0.95609	0.003111	0.423	0.06284	0.42211	0.06318
$\lambda = 5$	2.3964	1.622e-06	0.178	0.3145	0.08937	0.6396	0.08607	0.6576
$\lambda = 10$	1.0749	0.001638	0.12153	0.4893	0.06887	0.7586	0.062997	0.7952

En las graficas presentadas en la **sección 4.0.1** se puede observar como los valores obtenidos para la distribución de poisson de parámetros ($\lambda = 1$), ($\lambda = 5$), ($\lambda = 10$) con tamaños de muestra ($n = 1$), ($n = 10$), ($n = 25$), ($n = 50$) por el generador de R, se aprecia que a medida que se aumenta el tamaño n de la muestra, el comportamiento de la distribución se traslada hasta llegar a asemejarse a la distribución normal por lo cuál se puede concluir que se cumple así con el teorema del limite central el cual afirma que cuando el tamaño de la muestra es lo suficientemente grande, la distribución de las medias sigue aproximadamente una distribución normal. La afirmación anterior se puede validar mediante la prueba de bondad de ajuste, para el caso particular que nos compete se uso la prueba de normalidad Cramer-Von mises que nos permite determinar la pertenencia de los datos simulados a la distribución normal, los valores-p correspondientes a la prueba se encuentran en la tabla **sección 4.0.2** en los cuales se observa como el valor-p aumenta a medida que aumenta el tamaño de la muestra, indicando que mientras mayor sea el tamaño de muestra, mejor va a ser la convergencia de la distribución de las medias a la distribución normal.

4.0.3. Distribucion Logistica:

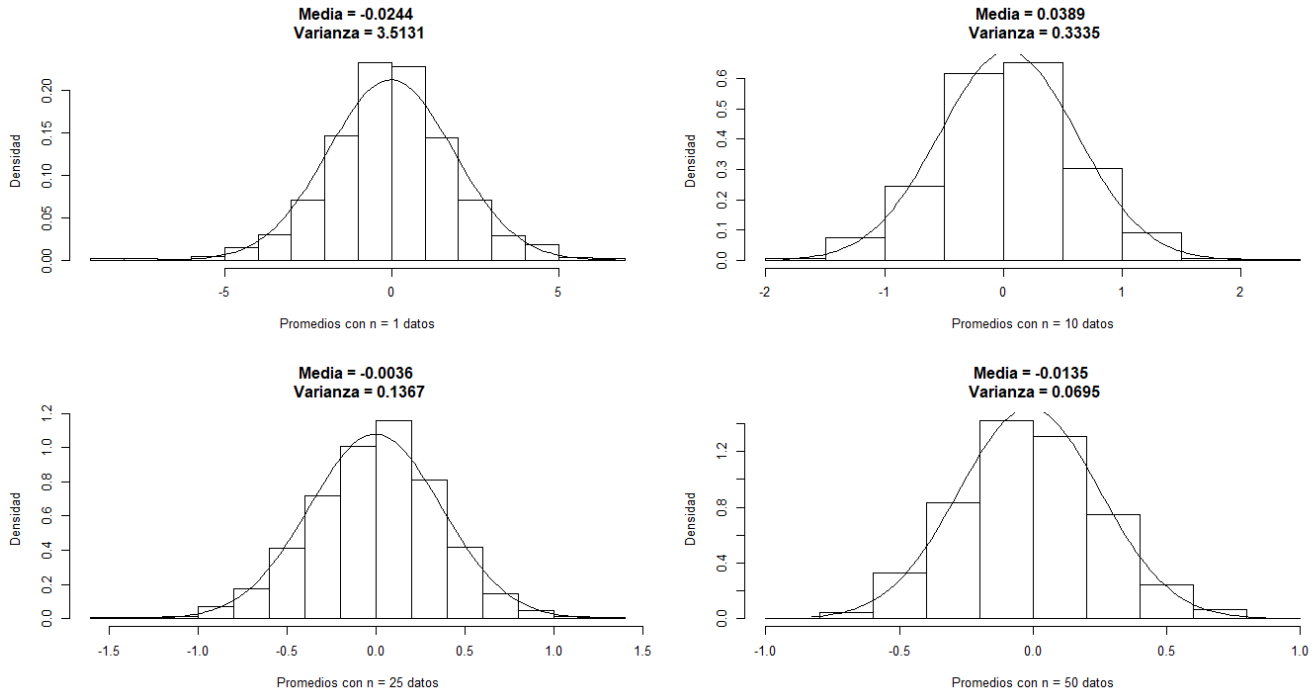

FIGURA 4: Simulaciones en R de la distribución Logística($a = 0, b = 1$) con muestras de tamaño ($n = 1$), ($n = 10$), ($n = 25$), ($n = 50$)

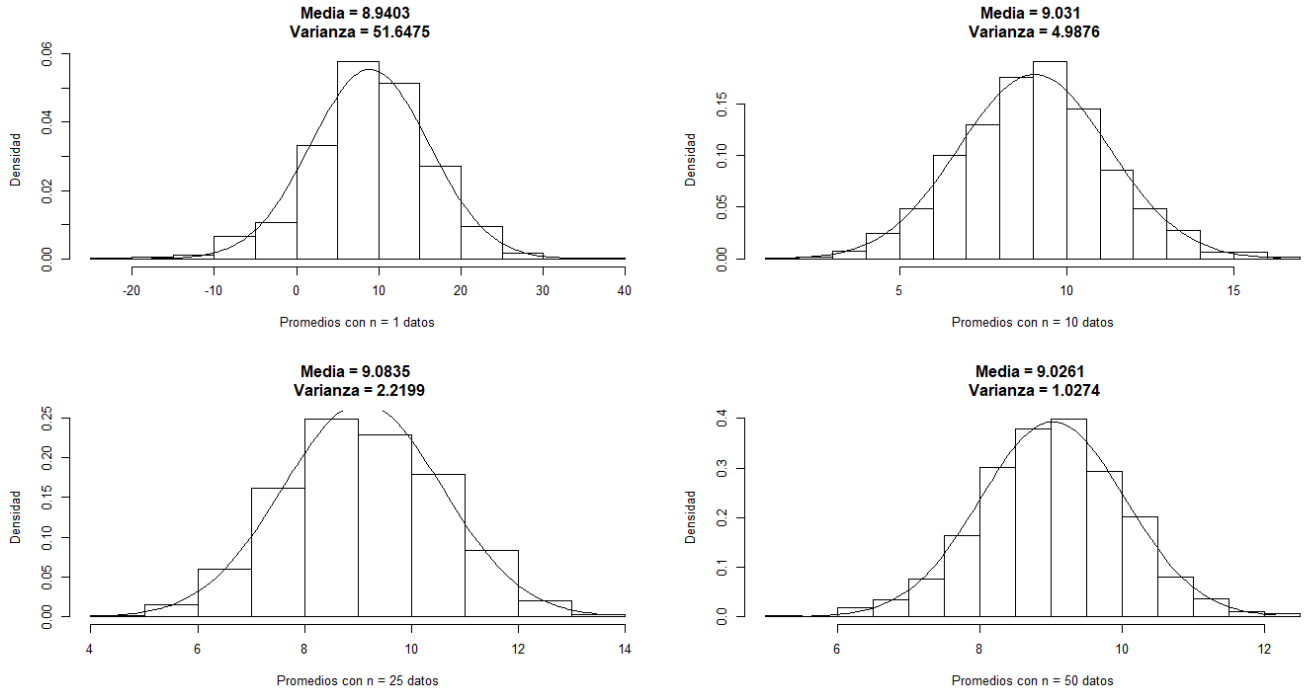

FIGURA 5: Simulaciones en R de la distribución Logística($a = 9, b = 4$) con muestras de tamaño ($n = 1$), ($n = 10$), ($n = 25$), ($n = 50$)

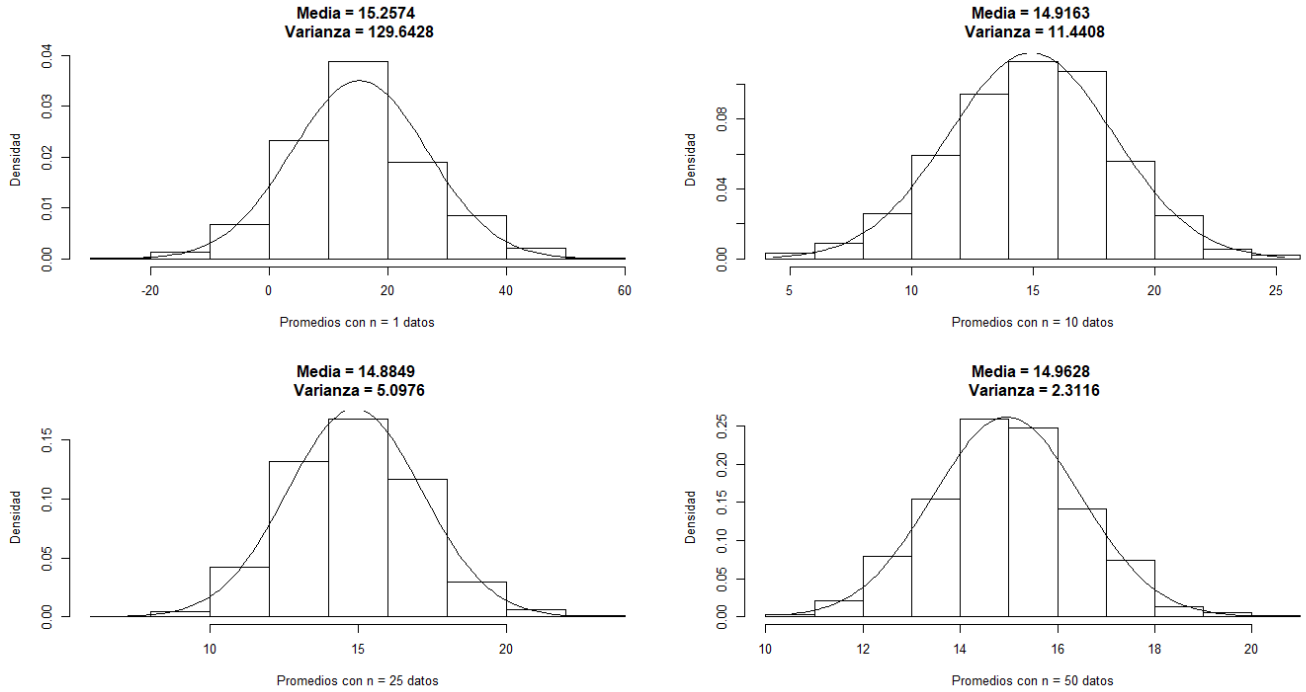

FIGURA 6: Simulaciones en R de la distribución Logística$(a = 15, b = 6)$ con muestras de tamaño $(n = 1)$, $(n = 10)$, $(n = 25)$, $(n = 50)$

4.0.4. Prueba Cramer-Von Mises Distribución Logística.

	n=1		n=10		n=25		n=50	
Parametro	W	P-Valor	W	P-Valor	W	P-Valor	W	P-Valor
$(a = 0, b = 1)$	0.18989	0.288	0.0804	0.6894	0.02928	0.9785	0.0477	0.8903
$(a = 9, b = 4)$	0.2185	0.2343	0.02528	0.989	0.07099	0.7456	0.026256	0.9868
$(a = 15, b = 6)$	0.55174	0.02948	0.03095	0.973	0.019361	0.9976	0.044503	0.9088

En las graficas presentadas en la **sección 4.0.3** se puede observar como los valores obtenidos para la distribución Logística de parámetros $(a = 0, b = 1)$, $(a = 9, b = 4)$, $(a = 15, b = 6)$, con tamaños de muestra $(n = 1)$, $(n = 10)$, $(n = 25)$, $(n = 50)$ por el generador de R, se aprecia que la distribución tiene ya un comportamiento semejante al de la distribución normal, siendo esta simetrica al rededor del parametro **a**. Mas sin embargo se aprecia como al aumentar el tamaño n de la muestra, el comportamiento de la distribución logística se hace mas a la distribución normal por lo cuál se puede concluir que se cumple así con el teorema del limite central el cual afirma que cuando el tamaño de la muestra es lo suficientemente grande, la distribución de las medias sigue aproximadamente una distribución normal. La afirmación anterior se puede validar mediante la prueba de bondad de ajuste, para el caso particular se uso la prueba de normalidad Cramer-Von mises que nos permite determinar la pertenencia de los datos simulados a la distribución normal, los valores-p correspondientes a la prueba se encuentran en la tabla **sección 4.0.4** en los cuales se observa como el valor-p aumenta a medida que aumenta el tamaño de la muestra, indicando que mientras mayor sea el tamaño de muestra, mejor va a ser la convergencia de la distribución de las medias a la distribución normal.

- Es posible afirmar que para n de tamaño mayores a 30 $(n > 30)$ se cumple el Teorema.
- La distribución de probabilidad de X afecta la convergencia.
- Los parámetros de la distribución de probabilidad de X afecta la convergencia.

5. Referencias

- $https://www.vitutor.com/estadistica/inferencia/intervalos.html$
- $http://www.ub.edu/stat/GrupsInnovacio/Statmedia/demo/Temas/Capitulo5/B0C5m1t4.htm$
- $http://e-stadistica.bio.ucm.es/glosario2/teorlimitecentral.html$
- $https://www.researchgate.net/publication/262701897_Sobre_el_Estadistico_de_Cramer-Von_Mises$
- $http://www.scielo.sa.cr/pdf/rmta/v19n1/a07v19n1.pdf$

6. Scripts

```
#----------------------------------#
# Kevin Steven Garcia Chica
# Cesar A. Saavedra
# Simulacion Estadistica Trabajo 2
#----------------------------------#
install.packages("goftest")
library("goftest")
#funci?n:
Fun<-function(x,n,k){
N=length(x)
medias=rep(0,k)
for (i in 1:k) {
O=order(runif(N))
medias[i]=mean(x[O[1:n]])
}
m=round(mean(medias),4)
s2=round((var(medias)*(k-1)/k*10000)/10000,4)
hist(medias,freq=FALSE,ylab="Densidad",xlab = paste("Promedios con n =",n,"datos"),
main = paste("Media =",m,"\n Varianza =",s2))

curve(exp(-((x-mean(medias))/sd(medias))^2/2)/(sd(medias)*sqrt(2*pi))
,min(medias),max(medias),add=T)
cvm.test(medias,"pnorm")
}

#Distribuci?n Logistica:
x<-rlogis(20000,0,1)  #Par?metros 0,1
x11()
par(mfrow=c(2,2))
Fun(x,1,5000)
Fun(x,10,5000)
Fun(x,25,5000)
Fun(x,50,5000)


x1<-rlogis(20000,9,4)  #Par?metros 9,4
x11()
par(mfrow=c(2,2))
Fun(x1,1,5000)
Fun(x1,10,5000)
Fun(x1,25,5000)
Fun(x1,50,5000)

x2<-rlogis(20000,15,6)  #Par?metros 15,6
x11()
par(mfrow=c(2,2))
Fun(x2,1,5000)
Fun(x2,10,5000)
Fun(x2,25,5000)
Fun(x2,50,5000)

#Distribuci?n Poisson:
x3<-rpois(20000,1)  #Par?metro lambda=1
x11()
par(mfrow=c(2,2))
Fun(x3,1,5000)
Fun(x3,5,5000)
Fun(x3,10,5000)
```

```
Fun(x3,30,5000)

x4<-rpois(20000,5)  #Par?metros lambda=5
x11()
par(mfrow=c(2,2))
Fun(x4,1,5000)
Fun(x4,5,5000)
Fun(x4,10,5000)
Fun(x4,30,5000)

x5<-rpois(20000,10)  #Par?metros lambda=10
x11()
par(mfrow=c(2,2))
Fun(x5,1,5000)
Fun(x5,5,5000)
Fun(x5,10,5000)
Fun(x5,30,5000)
```