

UNIVERSIDAD SIMÓN BOLÍVAR

Departamento de Cómputo Científico y Estadística

CO3321 Estadística para Ingenieros

Abril - Julio 2013

## **PROYECTO 1**

**Profesora:**

Zoraida Martínez

**Realizado por:**

Coello Coraspe, Katyuska 07-40767

Heligon, Roberto 09-10395

Quintero, Natacha 08-10910

**GRUPO 31**

Julio de 2013.

## **INDICE**

	<b>Pág.</b>
Resumen	3
Introducción	3
Objetivos	4
General	4
Específicos	4
Metodología	4
Resultados y Análisis	5
Parte 1	5
Parte 2	10
Conclusiones y Recomendaciones	14
Bibliografía	15
Anexos	16

## **RESUMEN**

El presente informe describe un análisis descriptivo de datos asociados al número de hijos nacidos vivos de una mujer, dichos datos se dividen entre cuatro entidades principales. Se muestran los gráficos correspondientes a la aplicación de estadística descriptiva a los datos, al igual que el análisis de los mismos.

Además de dicho análisis basado en estadística descriptiva se tiene una variable aleatoria y dos estimadores para la misma, a los cual se le hacen análisis para conocer su sesgo, confiabilidad, entre otros.

## **INTRODUCCIÓN**

La estadística descriptiva el área de la estadística que se dedica a recolectar, ordenar, analizar y representar un conjunto de datos, con el fin de describir apropiadamente las características de este.

En estadística, un estimador es un estadístico usado para estimar un parámetro desconocido de la población. Para cada parámetro pueden existir varios estimadores diferentes. En general, escogeremos el estimador que posea mejores propiedades que los restantes, como insesgadez, eficiencia, convergencia y robustez.

Tomando éstos dos conceptos fundamentales de la estadística, estimadores dados y como base los datos que fueron enviados por los profesores de la materia por grupo, se realizarán a lo largo del informe un análisis estadístico de las variables asociadas al mismo, describiendo las variables y utilizando el paquete estadístico R para la elaboración del análisis descriptivo completo.

## **OBJETIVOS**

### **General**

Elaborar un análisis descriptivo completo para datos suministrados asociados al número de hijos nacidos vivos de una mujer y realizar un análisis basados en estimadores dados de una variable.

### **Específicos**

1. Usar la herramienta R para generar gráficos y estadísticos.
2. Elaborar un análisis descriptivo los número de hijos nacidos vivos de una mujer.
3. Hacer el análisis de los estimadores dados correspondientes a una variable aleatoria  $X$  con una distribución normal.
4. Elaborar una variable categórica que clasifique el tipo de sismo según su magnitud.

## **METODOLOGÍA**

Se utilizará el paquete analítico R para la realización de un script para la generación de gráficos y datos estadísticos que corresponden a la aplicación de estadística descriptiva a los datos de estudio.

Luego de obtener toda la información necesaria, se pasará a la parte de análisis y luego se generará el informe con el análisis y resultados.

## RESULTADOS Y ANÁLISIS

### Parte 1:

a) Tomando en cuenta que la Variable X es aleatoria con distribución normal, para el primer estimador tenemos:

\*\*\*\*\* PRIMER ESTIMADOR \*\*\*\*\*

Tenemos que el primer estimador es :  $\hat{u}_1 = 1/4*(X_1 + 2X_2 + X_3)$

$$E(\hat{u}_1) = E(1/4*(X_1 + 2X_2 + X_3))$$

$$E(\hat{u}_1) = 1/4*(E(X_1) + E(2X_2) + E(X_3))$$

$$E(\hat{u}_1) = 1/4*(E(X_1) + 2E(X_2) + E(X_3))$$

-  $E(X_1) = u$  (miu) por ser la esperanza de una distribución normal. Lo mismo para  $E(X_2)$  y  $E(X_3)$

$$E(\hat{u}_1) = 1/4*(u + 2u + u)$$

$$E(\hat{u}_1) = 1/4*(4u)$$

$$E(\hat{u}_1) = u$$

- Según la definición, Si  $E(\hat{u}_1) = u$  entonces el estimador es insesgado.

$$VAR(\hat{u}_1) = VAR((1/4*(X_1 + 2X_2 + X_3))$$

$$VAR(\hat{u}_1) = VAR(1/4*X_1 + (1/4)*2X_2 + 1/4*X_3)$$

$$VAR(\hat{u}_1) = (1/4)^2 * VAR(X_1) + (1/2)^2 * VAR(X_2) + (1/4)^2 * VAR(X_3)$$

- Sabemos que  $VAR(X_1) = 16$ ,  $VAR(X_2) = 16$ ,  $VAR(X_3) = 16$

$$VAR(\hat{u}_1) = (1/4)^2 * 16 + (1/2)^2 * 16 + (1/4)^2 * 16$$

$$VAR(\hat{u}_1) = (1/16) * 16 + (1/4) * 16 + (1/16) * 16$$

$$VAR(\hat{u}_1) = 1 + 4 + 1 = 6$$

SESGO:

$$B(\hat{u}_1) = E(\hat{u}_1) - u = u - u = 0$$

Para el segundo estimador se tiene que:

\*\*\*\*\* SEGUNDO ESTIMADOR \*\*\*\*\*

Tenemos que el segundo estimador es :  $\hat{u}_2 = 1/5*(X_1 + 2X_2 + X_3)$

$$E(\hat{u}_2) = 1/5*(E(X_1) + E(2X_2) + E(X_3))$$

$$E(\hat{u}_2) = 1/5*(E(X_1) + 2*E(X_2) + E(X_3))$$

-  $E(X_1) = u$  por ser la esperanza de una distribución normal. Los mismo para  $E(X_2)$  y  $E(X_3)$

$$E(\hat{u}_2) = 1/5*(u + 2*u + u) = 1/5*(4u)$$

- El estimador es sesgado.

$$VAR(\hat{u}_2) = VAR((1/5*(X_1 + 2*X_2 + X_3))$$

$$VAR(\hat{u}_2) = VAR(1/5*X_1 + (1/5)*2*X_2 + 1/5*X_3)$$

$$VAR(\hat{u}_2) = (1/5)^2 * VAR(X_1) + (2/5)^2 * VAR(X_2) + (1/5)^2 * VAR(X_3)$$

- Sabemos que  $VAR(X_1) = 16$  ,  $VAR(X_2) = 16$  ,  $VAR(X_3) = 16$

$$VAR(\hat{u}_2) = (1/5)^2 * 16 + (2/5)^2 * 16 + (1/5)^2 * 16$$

$$VAR(\hat{u}_2) = (1/25) * 16 + (4/25) * 16 + (1/25) * 16$$

$$VAR(\hat{u}_2) = ((1 + 4 + 1)/25) * 16$$

$$VAR(\hat{u}_2) = (6/25) * 16 \quad VAR(\hat{u}_2) = (96/25) = 3,84$$

SESGO :

$$B(\hat{u}_1) = E(\hat{u}_1) - u = 4/5u - u = 4/5u - 5/5u = ((4-5)/5)u = (-1/5)u$$

Además, tenemos el error cuadrático de ambos estimadores, que se calculan a continuación:

\*\*\*\*\* ERROR CUADRATICO DE AMBOS ESTIMADORES \*\*\*\*\*

$$ECM(\hat{u}_1) = VAR(\hat{u}_1) + [B(\hat{u}_1)]^2$$

$$ECM(\hat{u}_1) = 6 + 0 = 6$$

$$ECM(\hat{u}_2) = VAR(\hat{u}_2) + [B(\hat{u}_2)]^2$$

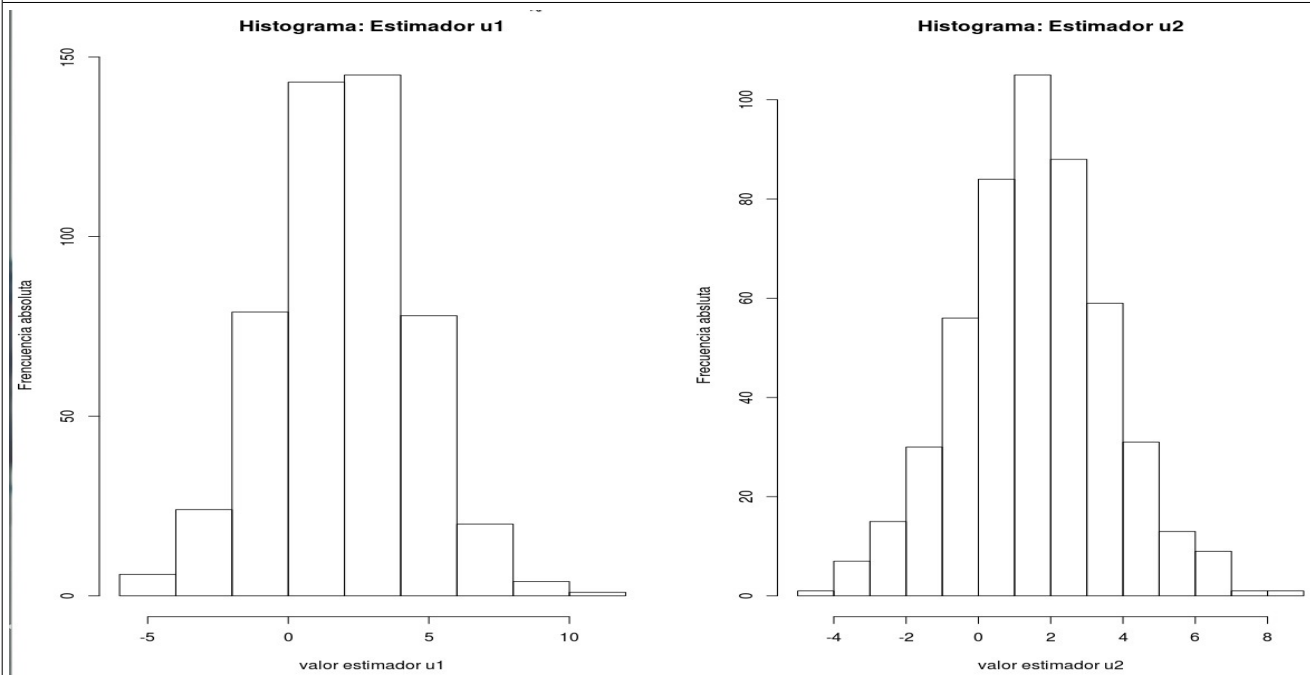
$$ECM(\hat{u}_2) = 3,84 + ((-1/5)u)^2$$

Es importante resaltar que si comparamos los errores cuadráticos de ambos estimadores, no podemos decir con certeza cuál de los dos es mejor, debido a que el segundo estimador depende del valor de la media, la cuál es desconocida.

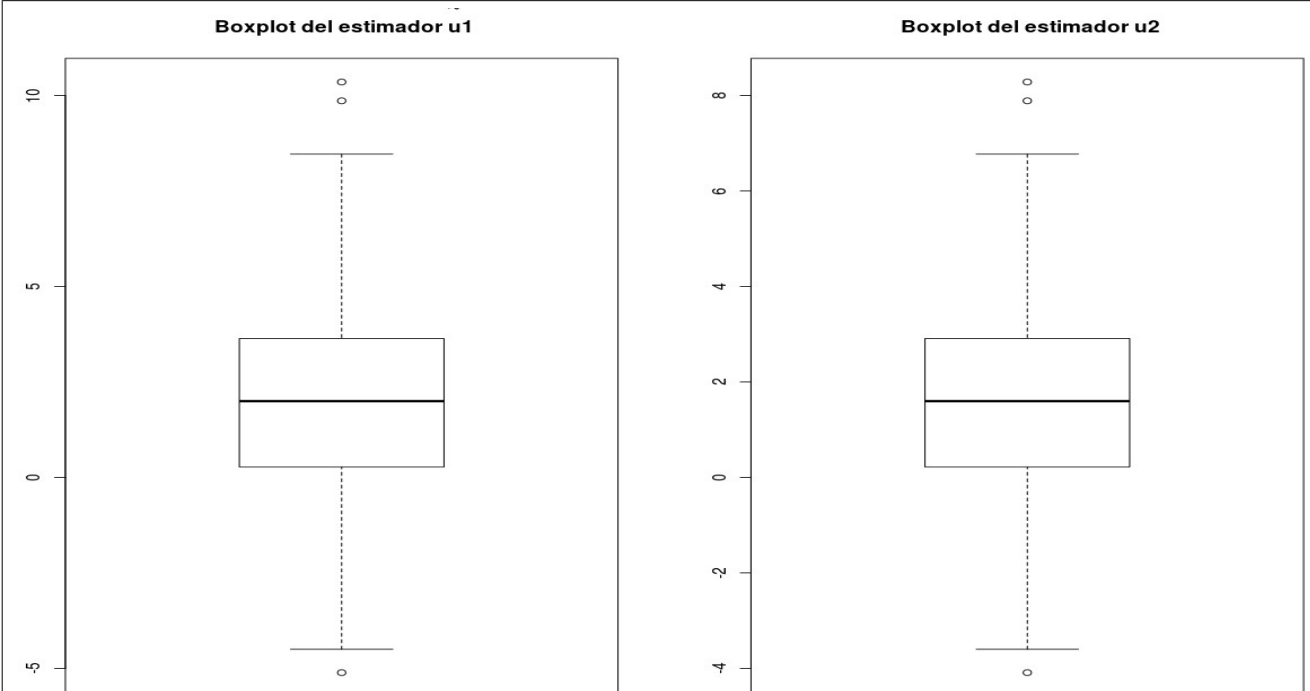
b)

**Para la Media = 2**

**Histograma**



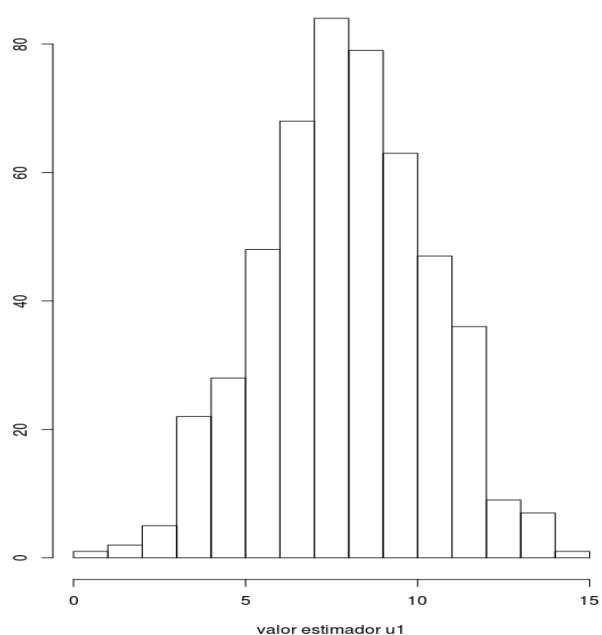
**Diagramas de Caja**



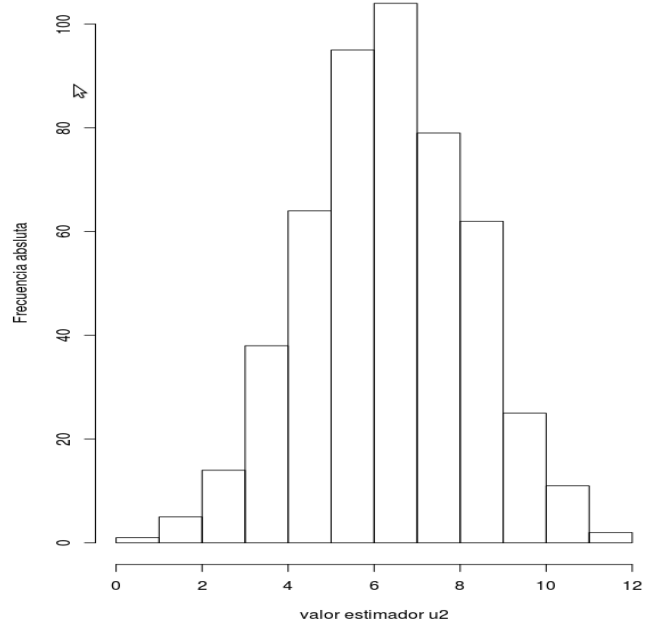
## Para la Media = 8

### Histogramas

Histograma: Estimador u1

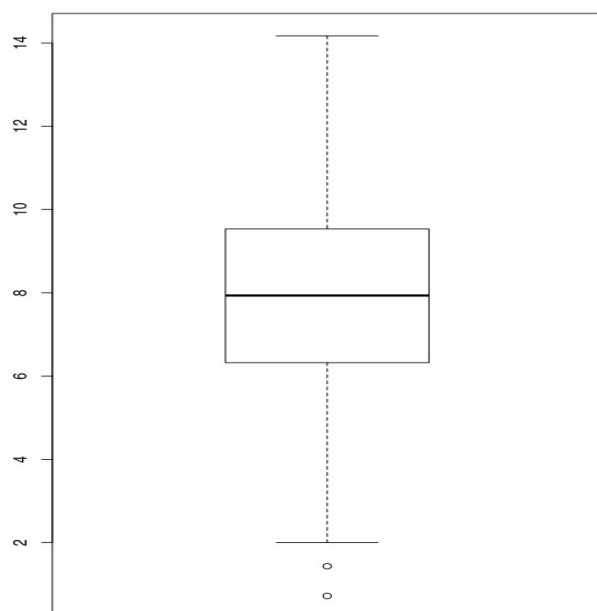


Histograma: Estimador u2

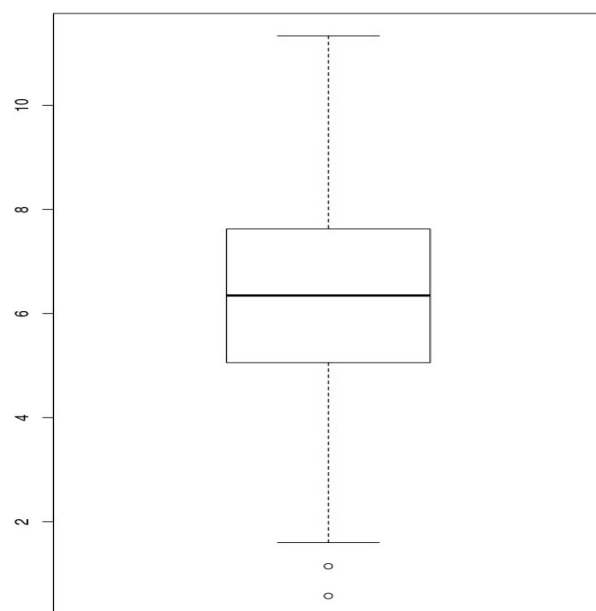


### Diagramas de Caja

Boxplot del estimador u1



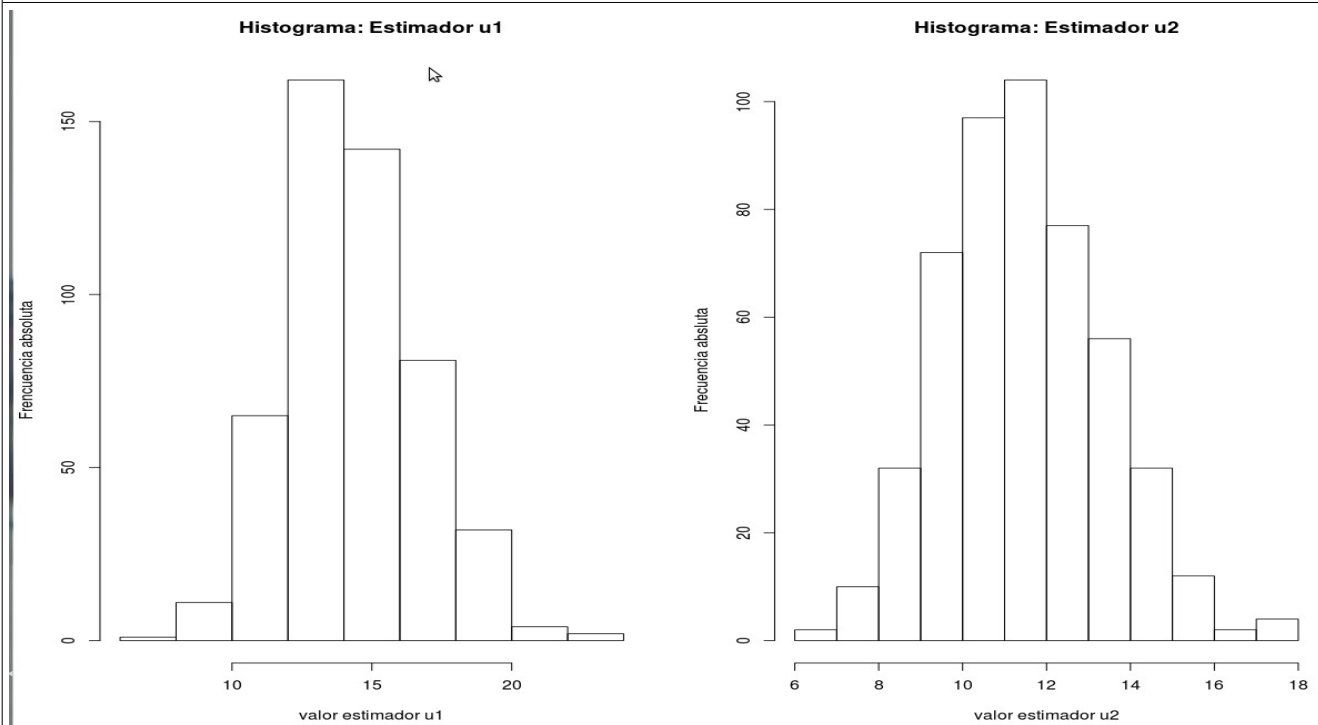
Boxplot del estimador u2



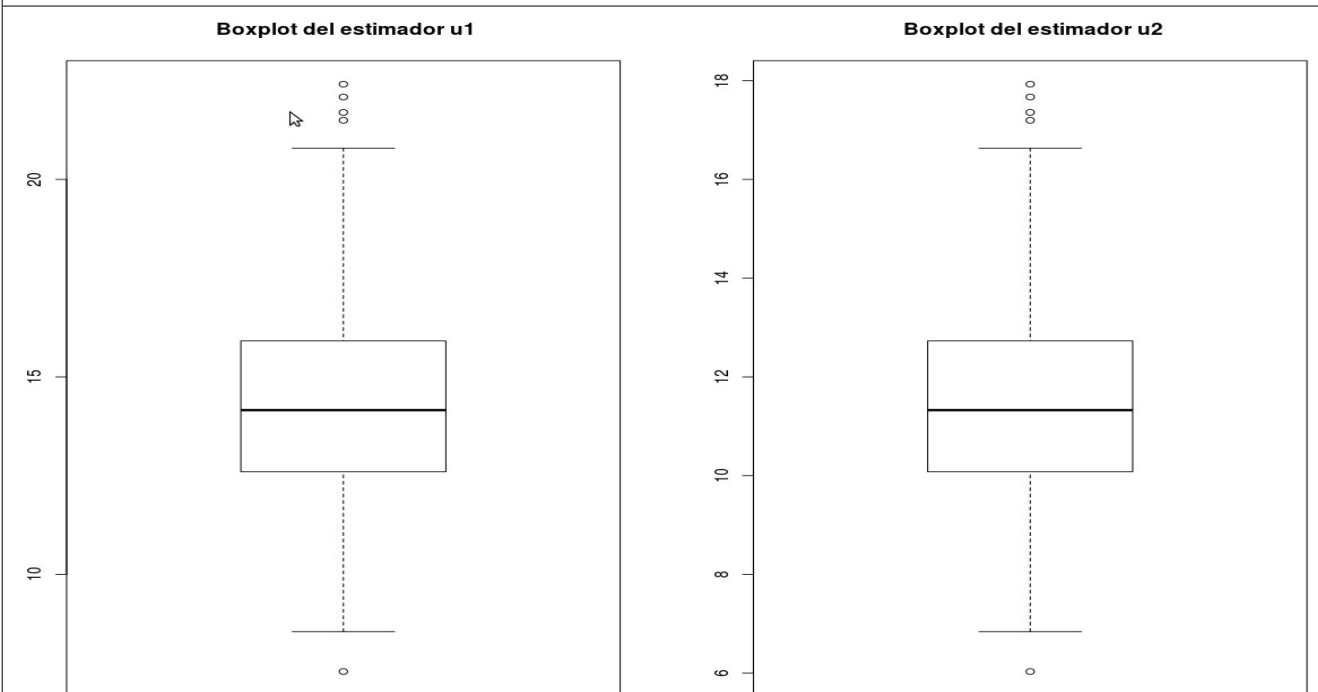


## Para la Media = 14

### Histogramas

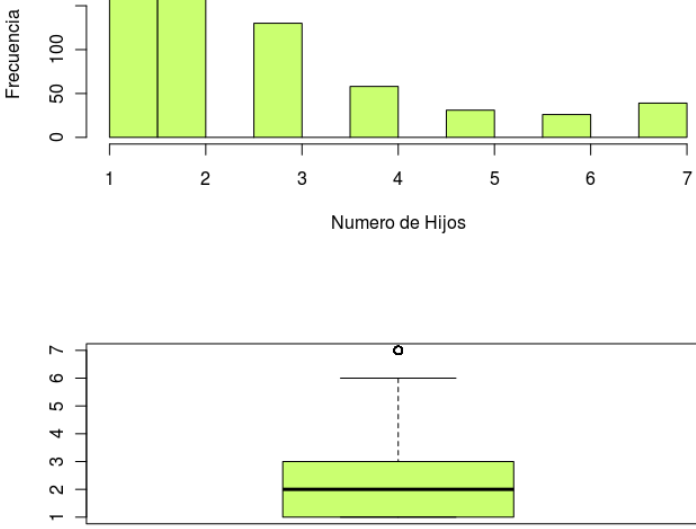
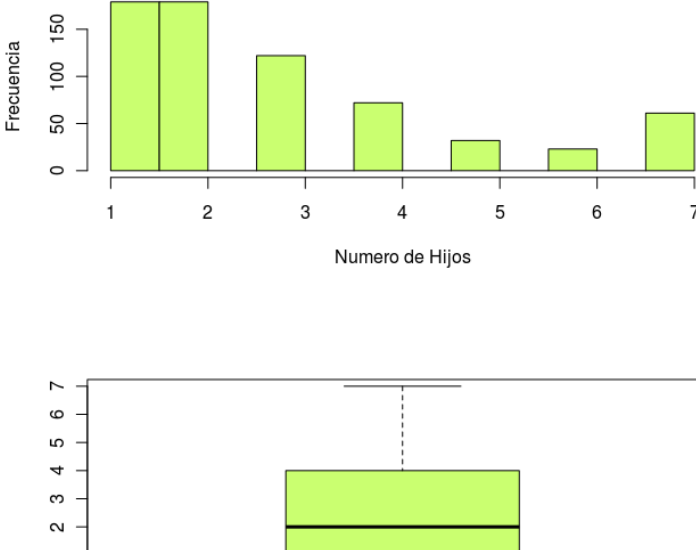


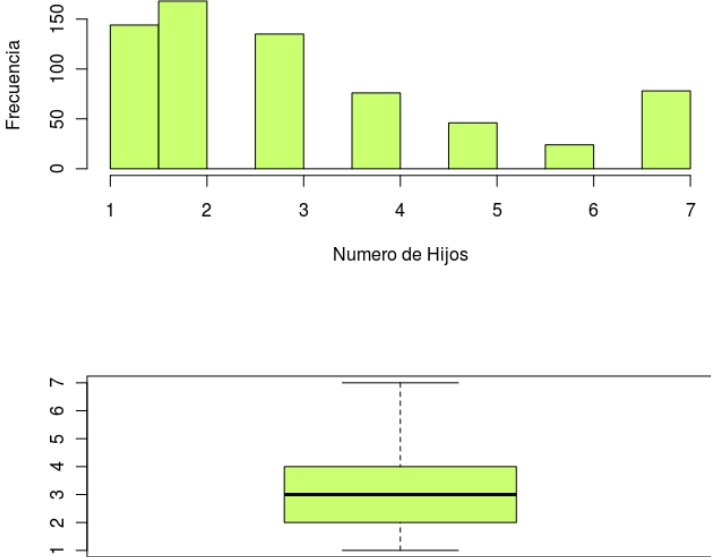
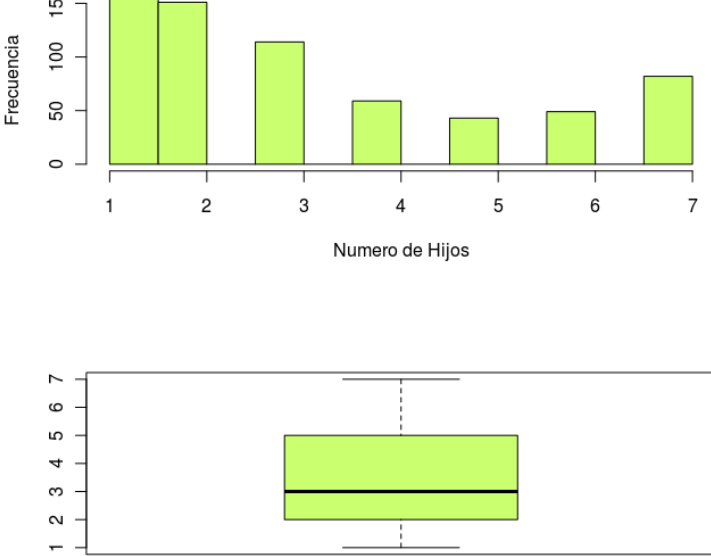
### Diagramas de Caja



## Parte 2:

a) A continuación tenemos el análisis descriptivo de la variable asociada al número de hijos nacidos vivos de una mujer por entidad federal de Venezuela:

Entidad	Análisis Descriptivo	Histograma y Gráfico de Caja
Distrito Capital	Media: 2.68292682926829 Mediana: 2 1er Cuartil: 1 3er Cuartil: 3 Desv. Estándar: 1.710 Coef. de Variación: 0.6374	 <p>El histograma para el Distrito Capital muestra la distribución de hijos nacidos vivos. El eje horizontal representa el 'Numero de Hijos' (1 a 7) y el eje vertical la 'Frecuencia' (0 a 100). Las barras son de color verde. El gráfico de caja correspondiente muestra la mediana en 2, el primer cuartil en 1, el tercer cuartil en 3, y un valor atípico en 7.</p>
Miranda	Media: 2.8682 Mediana: 2 1er Cuartil: 1 3er Cuartil: 4 Desv. Estándar: 1.84334 Coef. de Variación: 0.6426	 <p>El histograma para Miranda muestra la distribución de hijos nacidos vivos. El eje horizontal representa el 'Numero de Hijos' (1 a 7) y el eje vertical la 'Frecuencia' (0 a 150). Las barras son de color verde. El gráfico de caja correspondiente muestra la mediana en 2, el primer cuartil en 1, el tercer cuartil en 4, y valores atípicos en 7.</p>

Aragua	<p>Media:3.14</p> <p>Mediana: 3</p> <p>1er Cuartil: 2</p> <p>3er Cuartil: 4</p> <p>Desv. Estándar: 1.9139</p> <p>Coef. de Variación: 0.6089</p>	 <p>The figure for Aragua consists of two charts. The top chart is a histogram with the x-axis labeled 'Numero de Hijos' (Number of Children) ranging from 1 to 7, and the y-axis labeled 'Frecuencia' (Frequency) ranging from 0 to 150. The bars show frequencies of approximately 140 for 1 child, 160 for 2 children, 135 for 3 children, 75 for 4 children, 45 for 5 children, 25 for 6 children, and 80 for 7 children. The bottom chart is a box plot with the same x-axis. The box extends from the 1st quartile (2) to the 3rd quartile (4), with a median line at 3. Whiskers extend from 1 to 7.</p>
Zulia	<p>Media: 3.2366</p> <p>Mediana: 3</p> <p>1er Cuartil: 2</p> <p>3er Cuartil: 5</p> <p>Desv. Estándar: 2.04173</p> <p>Coef. de Variación: 0.6308</p>	 <p>The figure for Zulia consists of two charts. The top chart is a histogram with the x-axis labeled 'Numero de Hijos' (Number of Children) ranging from 1 to 7, and the y-axis labeled 'Frecuencia' (Frequency) ranging from 0 to 150. The bars show frequencies of approximately 150 for 1 child, 145 for 2 children, 115 for 3 children, 60 for 4 children, 45 for 5 children, 50 for 6 children, and 80 for 7 children. The bottom chart is a box plot with the same x-axis. The box extends from the 1st quartile (2) to the 3rd quartile (5), with a median line at 3. Whiskers extend from 1 to 7.</p>

Podemos observar luego de realizar un análisis descriptivo por sector que sólo en los datos de Distrito Capital de presentan datos atípicos, y a muy baja escala, por lo que no es un dato importante para nuestro análisis. La media de las 4 entidades se encuentra entre 2,5 y 3,5 y vemos que este número es bastante constante en todas, teniendo una desviación estándar de 2 hijos por mujer, la cuál puede ser un poco mayor debido a que cuando se coloca 7 en los datos, puede ser 7 o más.

Se tiene también que los coeficientes de variación son bastante bajos y que existe un sesgo hacia la derecha en todas las gráficas. También se puede observar que el número de hijos en todos los estados, sobre todo en Distrito Capital y Miranda tiende a ser en la mayoría 1 o 2 hijos, mientras que el resto de la población se distribuye más normalmente en las demás cantidades.

**b)** Para dicho cálculo, utilizaremos el Intervalo de confianza para medias para cada entidad federal y luego haremos la comparación correspondiente para determinar si al 88% de confianza las medias poblacionales son iguales. Para esto, utilizaremos las siguientes líneas en el código en R que acompaña al Anexo 2 (ver anexos):

```
p1=c(zulia)
```

```
p2=c(dc)
```

```
p3=c(aragua)
```

```
p4=c(miranda)
```

```
a= t.test(p1,p2,conf.level = 0.88, var.equal= TRUE)
```

Para realizar el estudio, compararemos una tabla a la vez con las otras 3, para ver si son iguales, a un intervalo de confianza del 88%, si en los 4 casos son aceptadas, entonces tenemos que se acepta la hipótesis que las medias son iguales. A continuación vemos el caso de Distrito Capital, que fue el caso por el que rechazamos a un nivel de 88% de confianza que la media sea igual, debido a que se obtuvo lo siguiente:

<b>Distrito Capital</b>	<b>Miranda</b>	<b>Aragua</b>	<b>Zulia</b>
-	T = -1.8956 df = 1322 p-value = 0.05822 confidence interval: -0.33744452 -0.03322876	T = -4.6145 df = 1325 p-value = 4.323e-06 confidence interval: -0.6152811 -0.3050053	T = -5.323 df = 1309 p-value = 1.200e-07 confidence interval: -0.7155520 -0.3918768

Podemos ver que los intervalos de confianza no son suficientes para decir que las medias no son iguales, esto además podemos comprobarlo con la variación en los p-valor de cada caso. Esto también podemos verlo en los gráficos mostrados en la parte a de esta pregunta.

**c)** Para dicho cálculo, utilizaremos la diferencia de proporciones para cada entidad federal y luego haremos la comparación correspondiente para determinar si al 88% de confianza el número de mujeres con más de 3 hijos es igual. Para esto, utilizaremos las siguientes líneas en el código en R que acompaña al Anexo 2 (ver anexos):

```
p1=c(zulia>3)
p2=c(dc>3)
p3=c(aragua>3)
p4=c(miranda>3)
```

```
a= popo.test(c(p1,672),c(p2,657),conf.level = 0.88, var.equal= TRUE)
```

Para realizar el estudio, primero cambiamos el número de mujeres en total, debido a que es variable entre los datos, compararemos una tabla a la vez con las otras 3, para ver si son iguales, a un intervalo de confianza del 88%, si en los 4 casos son aceptadas, entonces tenemos que se acepta la hipótesis que las medias son iguales.

Luego de realizar el análisis completo, podemos decir que se acepta la hipótesis, debido a que no hubo ningún caso en donde se diera el rechazo, por lo que el número de mujeres con más de 3 hijos es igual proporcionalmente entre las entidades.

## **CONCLUSIONES Y RECOMENDACIONES**

Luego de haber estudiado los resultados y analizar de forma estadística los mismos, podemos decir que, efectivamente, el número de hijos que nacen vivos por mujer son eventos aleatorios matemáticamente hablando, debido a que son eventos naturales no predecibles, por lo que podemos obtener valores que difieren uno con el otro, sin seguir ningún tipo de patrón.

Podemos concluir que el análisis estadístico nos muestra información bastante relevante que puede ser utilizada para futuros estudios en el área de partos en el país, tomando en cuenta que este análisis es básico, y aunque hay tendencia a generalizar a toda la población, las primeras conclusiones obtenidas tras el análisis descriptivo, es un estudio calculando una serie de medidas y estadísticos para ver en qué medida los datos se agrupan o dispersan.

También se concluye que tomando los conceptos de estimadores y estadística descriptiva pueden realizarse análisis estadísticos de las variables asociadas a los mismos, describiendo las variables y utilizando el paquete estadístico R para la elaboración del análisis descriptivo completo, siendo ésta herramienta de gran ayuda para la realización de informes de éste tipo.

Se recomienda a los profesores del curso proveer el manual de R y a los estudiantes leerlo antes de su utilización debido a que hay funciones estadísticas que son útiles que ya se encuentran desarrolladas dentro del mismo y pueden ser utilizadas sin problema alguno.

## **BIBLIOGRAFÍA**

- [1] Estadística Matemática con Aplicaciones (7ma ed.) México: CENGAGE Learning.  
Wackerly, D.; Mendenhall, W. y Scheaffer R. (2008).
- [2] Guía de la Prof. María Egleé. Guía para la realización de informes de CO3321.
- [3] [http://es.wikipedia.org/wiki/Estad%C3%ADstica\\_descriptiva](http://es.wikipedia.org/wiki/Estad%C3%ADstica_descriptiva)
- [4] <http://es.wikipedia.org/wiki/Estimador>

## ANEXOS

### Anexo 1: Código en R utilizado para la parte 1 del proyecto

```
desviacions=4
Muestras= 500
X=matrix(0,500,3)
u1= matrix(0,500,1)
u2= matrix(0,500,1)

# Primer caso : Estimadores con mu= 2

mu = 2
j= 1

while (j <=Muestras){

X[j,]= rnorm(3,mu,desviacions)
u1[j,]=(X[j,1] + 2*X[j,2] + X[j,3])/4
u2[j,]=(X[j,1] + 2*X[j,2] + X[j,3])/5
j = j +1

}

par(mfrow=c(1,2))
hist( u1,xlab="valor estimador u1",ylab="Frecuencia absoluta",main="Histograma: Estimador u1")
hist( u2,xlab="valor estimador u2",ylab="Frecuencia absoluta", main="Histograma: Estimador u2")
boxplot(u1, main =" Boxplot del estimador u1")
boxplot(u2, main =" Boxplot del estimador u2")

# Segundo caso : Estimadores con mu= 8

mu = 8
jj= 1

while (jj <=Muestras){

X[jj,]= rnorm(3,mu,desviacions)
u1[jj,]=(X[jj,1] + 2*X[jj,2] + X[jj,3])/4
u2[jj,]=(X[jj,1] + 2*X[jj,2] + X[jj,3])/5
jj = jj +1

}

par(mfrow=c(1,2))
hist( u1,xlab="falta",main="Histogramas")
hist( u2,xlab="falta",main="Histogramas")
boxplot(u1, main =" Boxplot del estimador u1")
boxplot(u2, main =" Boxplot del estimador u2")

# Tercer caso : Estimadores con mu= 14

mu = 14
jjj= 1

while (jjj <=Muestras){

X[jjj,]= rnorm(3,mu,desviacions)
u1[jjj,]=(X[jjj,1] + 2*X[jjj,2] + X[jjj,3])/4
u2[jjj,]=(X[jjj,1] + 2*X[jjj,2] + X[jjj,3])/5
jjj = jjj +1

}
```



```

par(mfrow=c(1,2))
hist( u1,xlab="falta",main="Histogramas")
hist( u2,xlab="falta",main="Histogramas")
boxplot(u1, main =" Boxplot del estimador u1")
boxplot(u2, main =" Boxplot del estimador u2")

```

## Anexo 2: Código en R utilizado para la parte 2a del proyecto

```

#- Lectura de un Archivo desde los archivos .txt -#
z = read.table("zulmus_31.txt",header=TRUE)
d = read.table("dcmus_31.txt",header=TRUE)
m = read.table("mirmus_31.txt",header=TRUE)
a = read.table("aramus_31.txt",header=TRUE)

```

```

zulia = z[,1]
dc = d[,1]
aragua = a[,1]
miranda = m[,1]

```

```

#- Analisis descriptivo -#
analisis = function (datosCompleto) {
  # Media
  media = mean(datosCompleto)
  # Mediana
  mediana = median(datosCompleto)
  # Primer Cuartil
  q1 = quantile(datosCompleto,0.25)
  # Tercer Cuartil
  q3 = quantile(datosCompleto,0.75)
  # Desviacion Estándar
  de = sd(datosCompleto)
  # Coeficiente de Variacion
  coefv = sd(datosCompleto)/mean(datosCompleto)
  # Se establece el lienzo de dibujo
  par(mfrow=c(2,1))
  # Histograma
  hist(datosCompleto, ylab="Frecuencia", col= "darkolivegreen1")
  # Gráfico de cajas
  boxplot(datosCompleto, col= "darkolivegreen1")
  # Parámetros de retorno
  return (c(c("Media:",media), c("Mediana:",mediana),c("1er Cuartil:",q1), c("3er Cuartil:",q3), c("Desv. Estándar:",de), c("Coef.
de Variación:",coefv)))
}

#Descomentar linea por linea para el analisis descriptivo
p=analisis(zulia)
#p=analisis(dc)
#p=analisis(miranda)
#p=analisis(aragua)

```