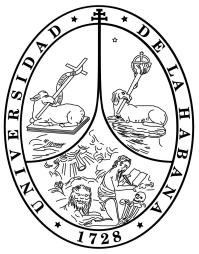

Informe escrito de Estadística
"Primera Fase"

Carlos Aryam Martínez Molina c.molina@estudiantes.matcom.uh.cu
Grupo C411

Eziel Ramos Piñón e.ramos@estudiantes.matcom.uh.cu
Grupo C411

Ariel Plasencia Díaz a.plasencia@estudiantes.matcom.uh.cu
Grupo C411

Índice

1 Introducción

El objetivo de este trabajo es, a partir de las 50 mejores canciones de la aplicación Spotify y de tres observaciones escogidas, obtener sus estadísticos descriptivos, graficar los resultados mediante un histograma y el gráfico de cajas y bigotes e interpretar los resultados alcanzados.

Además mostraremos cómo generar una población normal y seleccionar de ésta 8 muestras de distintos tamaños (con y sin reemplazamiento). Con todos los datos generados haremos un análisis minucioso en cuanto a distintas ramas de la estadística moderna como la estadística descriptiva, la estadística inferencial y la estadística matemática.

Para llevar a cabo lo explicado anteriormente nos apoyaremos en el lenguaje de programación R, el cual fue diseñado por Ross Ihaka y Robert Gentleman en 1933 y es muy utilizado en nuestros días para la investigación por la comunidad estadística, en el campo de la minería de datos, la investigación biomédica, la bioinformática y las matemáticas financieras.

2 Ejercicio 1

2.1 Problema

De acuerdo a su set de datos:

1. Utilice los estadísticos descriptivos estudiados en la conferencia 1 para describir el comportamiento de tres de sus variables. Seleccione las que sean más importantes y explique porqué seleccionó éstas.
2. Grafique los resultados.
3. Interprete los resultados en términos del problema.

2.2 Solución

De todas las observaciones dadas, nos llamó más la atención las muestras de *Popularidad*, *Acústica* y *Bailabilidad*, ya que, a nuestro juicio, corresponden tres factores esenciales para el éxito de una canción en nuestros días. Es necesario mencionar que todos estos cálculos y gráficos se llevaron a cabo con el intérprete RStudio, un entorno para programar en R.

2.2.1 Medidas de Tendencia Central

Observación	Media Aritmética	Moda	Mediana
Popularidad	87.50	88.00 y 89.00	88.00
Acústica	22.16	12.00	15.00
Bailabilidad	71.38	75.00	73.50

Interpretación de los datos:

1. Los resultados obtenidos revelan que los índices de *popularidad* de las canciones se encuentran alrededor de 88.
2. A pesar de que la media *acústica* es aproximadamente 22, el valor que más se repite en los datos para esta muestra es 12.
3. Los resultados nos muestran altos índices de danceabilidad, alrededor de 75, el cual se distintivo de canciones movidas.
4. Si tomamos una canción muy cercana a esta lista de Spotify (por ejemplo la número 53) podemos esperar un valor de popularidad cercano a 88.
5. Como valor más frecuente dentro de la muestra de *acústica* tenemos al 12.
6. Dado que la mediana asociada a la *bailabilidad* de una canción supera el valor de la media aritmética, puede asegurarse que más del 50 % de los datos están por encima de la media.
7. Debido a que los datos corresponden a canciones modernas se obtienen buenos resultados con respecto a la propagación del sonido, ya sea en lugares abiertos o cerrados.

2.2.2 Medidas de Dispersión

Observación	Desviación Estándar	Varianza	Coeficiente de Variación
Popularidad	4.49148854921785	20.1734693877551	0.0513312977053469
Acústica	18.9955526481375	360.831020408163	0.857200029248082
Bailabilidad	11.929880167727	142.322040816327	0.167131972089199

El análisis de las **medidas de dispersión** nos permiten entender qué tan homogénea es la muestra y cuán veraces son nuestras **medidas de tendencia central**. Recordemos que:

Coeficiente de Variación	Interpretación
$CV \geq 26\,\%$	Muy heterogéneo
$16\,\% \leq CV < 26\,\%$	Heterogéneo
$11\,\% \leq CV < 16$	Homegéneo
$0\,\% \leq CV < 11$	Muy homogéneo

Teniendo en cuenta que los coeficientes de variación poseen valores de aproximadamente 5 %, 85 % y 16 %, para las muestras de *popularidad*, *acústica* y *bailabilidad* respectivamente, entonces podemos afirmar que los datos son muy homogéneos, muy heterogéneos o muy dispersos y homogéneos, en ese orden, para cada una de las observaciones mencionadas anteriormente.

2.2.3 Medidas de Posición

Las **medidas de posición** se utilizan para describir la posición de un dato específico con respecto al resto de los datos cuando están en orden por categorías. Cuartiles y percentiles son dos de las medidas más utilizadas.

Observación	Mínimo	Primer Cuartil	Segundo Cuartil	Tercer Cuartil	Cuarto Cuartil	Máximo
Popularidad	70.00	86.00	88.00	90.75	95.00	95.00
Acústica	1.00	8.25	15.00	33.75	75.00	75.00
Bailabilidad	29.00	67.00	73.50	79.50	90.00	90.00

Interpretación de los datos:

1. El 75 % de los valores de la muestra danceabilidad son menores que 79.5.
2. Se puede decir que el 25 % de los datos correspondientes a los valores de *popularidad* son menores que 86.
3. Los índices de *acústica* se encuentran entre 1 y 75, pero el 50 % de estos índices son menores que 15, por lo que los datos de la *acústica* se consideran bajos.
4. Tanto el 25 % como el 75 % de los datos correspondientes por la muestra de *bailabilidad* son mayores que 79.5 y que 67.0 respectivamente.
5. Podemos afirmar que no hay ninguna canción con valor de *popularidad* más alto que 95.
6. Tanto el 25 % como el 75 % de los datos señalados por la muestra de *acústica* son menores que 8.25 y que 33.75 respectivamente.
7. No existe ningún tema con nivel de danceabilidad superior a 90 ni inferior a 29.

2.2.4 Gráficos

A continuación, presentaremos tanto el histograma como el gráfico de cajas y bigotes de cada observación. Estos gráficos nos permitirán argumentar los intevalos donde menos y más valores encontramos, así como la posición de los datos y cuán alejada está la moda del valor central.

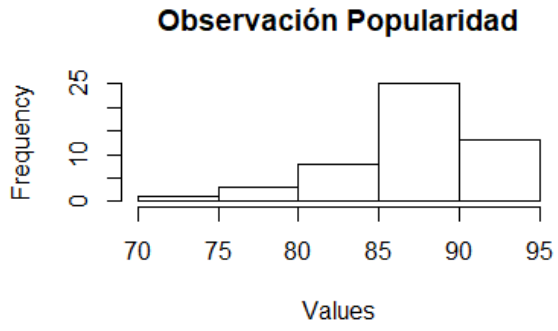

Figura 1: Histograma de Popularidad

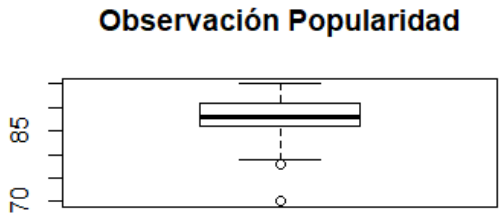

Figura 2: Gráfico de Cajas y Bigotes de Popularidad

Podemos decir que el rango donde menos índices de popularidad y donde más índices de popularidad encontramos es entre 70 y 75 y entre 85 y 90 respectivamente, por lo que la moda debe encontrarse en este último intervalo.

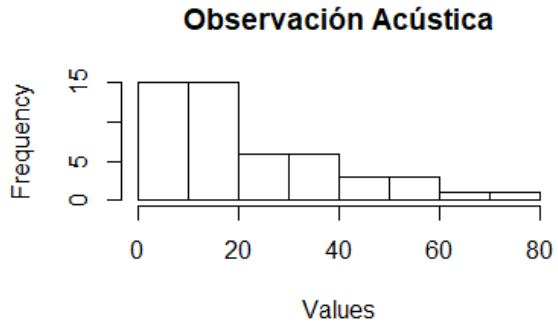

Figura 3: Histograma de Acústica

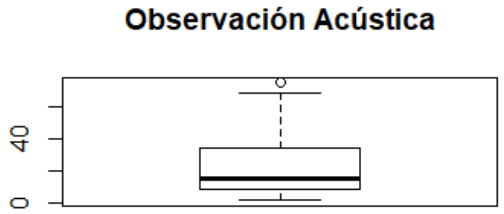

Figura 4: Gráfico de Cajas y Bigotes de Acústica

En este caso, cabe señalar que aunque los datos se encuentran en el intervalo [1, 75], existe una mayor concentración de éstos en su primera mitad, siendo el 25 % y el 50 % menores que 8.25 y que 15.00 respectivamente.

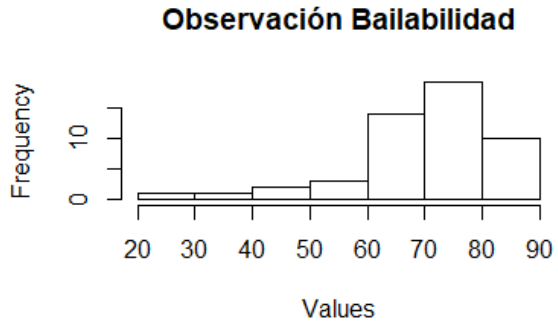

Figura 5: Histograma de Bailabilidad

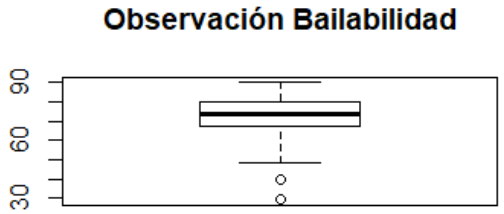

Figura 6: Gráfico de Cajas y Bigotes de Bailabilidad

En este último caso, podemos reafirmar que hay una mayor frecuencia en el intervalo [70, 80], pudiendo apreciar la poca diferencia que existe entre las **medidas de tendencia central**, valores que oscilan alrededor de 73.

2.3 Código

Solución en R

3 Ejercicio 2

3.1 Problema

Genere una población normal de tamaño 500, seleccione 8 muestras de varios tamaños (mucho mayor que 30, mayor que 30, exactamente 30 y exactamente 20), 4 muestras con reemplazo y 4 sin remplazo.

1. Calcule para cada una de las muestras los estadísticos descriptivos de la conferencia 1.
2. Calcúlelos para la población inicial y analice las diferencias.
3. Grafique los resultados.
4. Para cada muestra, calcule los intervalos de confianza para la media y para la varianza.
5. Analice las diferencias en los resultados de las muestras de tamaños similares.

3.2 Solución

3.2.1 Población

Para generar una población normal, hemos utilizado la función *rnorm()* del lenguaje de programación R, además hemos calculado los estadísticos descriptivos, así como ploteado sus gráficos, usando las mismas funciones que en el capítulo anterior. He aquí los resultados:

Media Aritmética:	$-0.00672856710740702 \approx -0.006729$
Mediana:	$0.00263899160027405 \approx 0.002639$
Desviación Estándar:	$0.951999278760831 \approx 0.951999$
Varianza:	$0.906302626761142 \approx 0.906303$
Coeficiente de Variación:	$-141.486183248858 \approx -141.486183$
Primer Cuartil:	-0.577333
Segundo Cuartil:	0.002639
Tercer Cuartil:	0.618719
Cuarto Cuartil:	3.324199
Mínimo:	-3.085717
Máximo:	3.324199

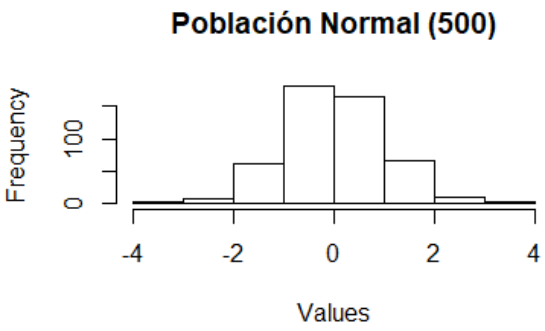

Figura 7: Histograma

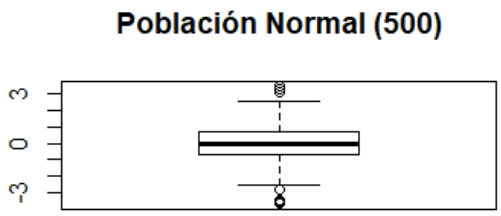

Figura 8: Gráfico de Cajas y Bigotes

3.2.2 Muestras Generadas

A partir de una población normal con tamaño 500, se generaron muestras con y sin reemplazamiento de distintos tamaños con el objetivo de analizar una serie de datos, así como una comparación entre ellos. La población ha sido generada con función de densidad $Z \sim N(\mu = 0,\ \sigma^2 = 1)$, esto trae como consecuencia que la varianza sea aproximadamente igual a 1 y que la media sea aproximadamente igual a 0, fenómeno que ocurre también en cada una de las muestras generadas. Para un mejor entendimiento solo mostraremos los estadísticos descriptivos de las muestras generadas sin reemplazamientos, sin embargo las muestras con reemplazamiento se presentarán en los anexos (epígrafe 5.1). A continuación un resumen de dichas muestras sin reemplazamiento agrupadas en dos tablas.

No.	Tamaño	Mínimo	Primer Cuartil	Mediana	Media	Tercer Cuartil	Máximo
1	300	-3.08572	-0.57201	-0.01988	-0.02301	0.60752	3.32420
2	100	-2.19822	-0.72940	-0.05036	-0.06884	0.60202	2.27425
3	30	-3.0857	-0.1425	0.1260	0.0438	0.5506	1.9628
4	20	-1.9281	-0.9524	-0.3005	-0.1849	0.1754	3.3242

No.	Tamaño	Varianza	Desviación Estándar	Coeficiente de Variación
1	300	0.859467883077966	0.927074906940084	-40.2822316750545
2	100	0.956403030614009	0.97795860373229	-14.2068164922199
3	30	0.962322309768774	0.980980280010141	22.3985039952095
4	20	1.29950932214534	1.13996022831735	-6.16692713827628

3.2.3 Diferencias

Respecto a la mediana y a la media, puede apreciarse que son ligeramente menores en las muestras que en la población con excepción de la muestra de tamaño 30, o sea, que el valor central de los datos se encuentran más a la izquierda en las muestras tomadas que en la población, pero a medida que aumentamos el tamaño de la muestra el valor se acerca cada vez más al de la población.

Respecto a las **medidas de posición**, a medida que aumenta el tamaño de la muestra disminuye el valor del primer cuartil, el cual se acerca cada vez más al valor del primer cuartil de la población, sin embargo el tamaño de la muestra y del tercer cuartil son inversamente proporcionales.

Podemos concluir que aunque las muestras fueron generadas aleatoriamente, a medida que se toma una muestra de mayor tamaño su distribución y estadísticos descriptivos se asemejan cada vez más al de la población inicial.

3.2.4 Intervalos de Confianza

El intervalo de confianza es un par de números entre los cuales se estima un cierto valor desconocido con una determinada probabilidad de acierto. Para el cálculo de los intervalos de confianza, nos hemos basado en las fórmulas vistas en conferencias. Los resultados son los siguientes:

Intervalos de confianza para la varianza ($\alpha = 0.95$)

Población normal de tamaño 500 :	[0.9039196 ; 0.9111291]
Muestra sin reemplazamiento de tamaño 300 :	[0.8569810 ; 0.8658247]
Muestra sin reemplazamiento de tamaño 100 :	[0.9543242 ; 0.9715368]
Muestra sin reemplazamiento de tamaño 30 :	[0.9686728 ; 1.0014260]
Muestra sin reemplazamiento de tamaño 20 :	[1.319027 ; 1.374631]

Intervalos de confianza para la media ($\alpha = 0.95$)

Población normal de tamaño 500 :	[−0.07339627 ; 0.05993914]
Muestra sin reemplazamiento de tamaño 300 :	[−0.10463443 ; 0.05860545]
Muestra sin reemplazamiento de tamaño 100 :	[−0.22615158 ; 0.08847702]
Muestra sin reemplazamiento de tamaño 30 :	[−0.2547318 ; 0.3423252]
Muestra sin reemplazamiento de tamaño 20 :	[−0.6873005 ; 0.3175993]

Los intervalos anteriores se consideran intervalos pequeños, por lo que ofrece una estimación más precisa aunque aumenta la probabilidad del error.

3.3 Código

Solución en R

4 Ejercicio 3

4.1 Problema

¿Se puede afirmar que el pop, sus subgéneros y derivados son más populares que el resto de los géneros?

Sugerencia: Asuma que todas las observaciones provienen de una distribución normal.

4.2 Solución

Para esta solución, nos apoyaremos completamente en el lenguaje de programación R. El código fuente se mostrará en la sección 4.3, sin embargo señalaremos los resultados alcanzados.

Primeramente haremos una prueba de hipótesis para la varianza de igualdad contra diferencia, donde la hipótesis nula es la igualdad entre los valores de la varianza, y la hipótesis alternativa es la diferencia. Cabe destacar, que el p-valor es el menor nivel de significación bajo el cual la hipótesis nula es rechazada y que siempre trabajaremos con un intervalo de confianza de un 95 %. Los resultados obtenidos en esta primera prueba son:

1. El estadígrafo es $F = 3.0006$.
2. Los grados de libertad son 22 y 26.
3. El p-valor es de 0.00825.
4. El intervalo de confianza es $[1.337013;\ 6.927914]$.

Como se obtiene un p-valor de aproximadamente igual a 0.008, esto implica que se acepta la hipótesis alternativa, ya que el p-valor es menor que 0.05, por lo cual las varianzas son distintas. A continuación, realizaremos una prueba de hipótesis para la media con varianzas desconocidas y diferentes. Recordemos que para esta segunda prueba, nuestra hipótesis alternativa es que la media de *popularidad* de los *géneros* pop y sus derivados es mayor estricto que la media de *popularidad* de los otros *géneros*. Los resultados alcanzados son los siguientes:

1. El estadígrafo es $T = -2.6116$.
2. El grado de libertad es 33.949.
3. El p-valor es de 0.9933.
4. El intervalo de confianza es $[-5.372263;\ +\infty)$.
5. La media de la *popularidad* de los *géneros* pop y sus subgéneros es 85.73913 mientras que la de los restantes *géneros* musicales es 89.00.

Como el p-valor es aproximadamente igual a 0.993 y mayor que 0.05, entonces se rechaza la hipótesis nula y se acepta la hipótesis alternativa. Con este resultado podemos afirmar que la media de *popularidad* de los *géneros* pop y sus subgéneros es mayor que la media de *popularidad* de los otros *géneros* musicales.

4.3 Código

Solución en R

5 Anexos

5.1 Muestras con reemplazamiento

Muestra (X_1) que representa a una muestra con reemplazamiento de tamaño 300

Media Aritmética:	$-0.020089780595447 \approx -0.02009$
Mediana:	$-0.0233116479420424 \approx -0.02331$
Desviación Estándar:	$0.91792439692969 \approx 0.91792$
Varianza:	$0.842585198478734 \approx 0.84259$
Coeficiente de Variación:	$-45.6911110884767 \approx -45.69111$
Primer Cuartil:	-0.66440
Segundo Cuartil:	-0.02331
Tercer Cuartil:	0.57348
Cuarto *Cuartil* :	3.32420
Mínimo:	-2.19822
Máximo:	3.32420
Intervalo de Confianza para la Media:	[-0.10010645 ; 0.05992689]
Intervalo de Confianza para la Varianza:	[0.8401472 ; 0.8488171]

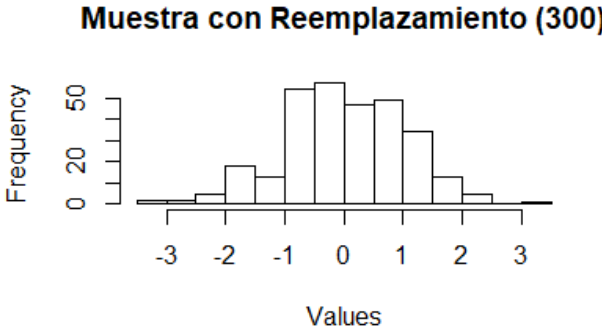

Figura 9: Histograma de X_1

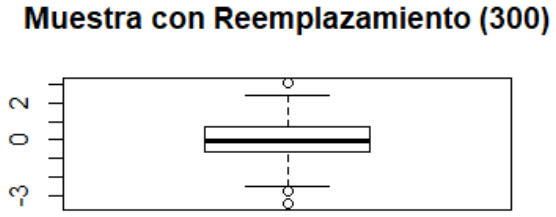

Figura 10: Gráfico de Cajas y Bigotes de X_1

Muestra (X_2) que representa a una muestra con reemplazamiento de tamaño 100

Media Aritmética:	$0.230302731187064 \approx 0.2303$
Mediana:	$0.265050630007453 \approx 0.2651$
Desviación Estándar:	$0.95714564446572 \approx 0.9571$
Varianza:	$0.916127784719699 \approx 0.9161$
Coeficiente de Variación:	$4.15603253826927 \approx 4.1560$
Primer Cuartil:	-0.3187
Segundo Cuartil:	0.2651
Tercer Cuartil:	0.8564
Cuarto Cuartil:	2.2742
Mínimo:	-2.0576
Máximo:	2.2742
Intervalo de Confianza para la Media:	[0.38099234 ; 0.07961312]
Intervalo de Confianza para la Varianza:	[0.9141365 ; 0.9306243]

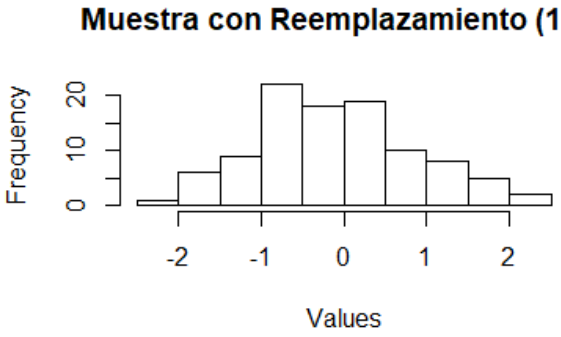

Figura 11: Histograma de X_2

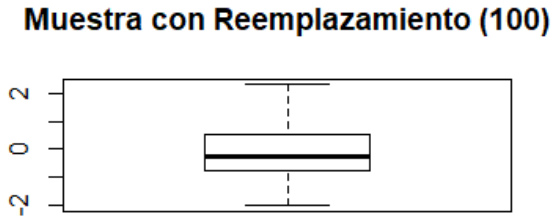

Figura 12: Gráfico de Cajas y Bigotes de X_2

Muestra (X_3) que representa a una muestra con reemplazamiento de tamaño 30

Media Aritmética:	0.209371679569799 ≈ 0.2094
Mediana:	0.274735007319923 ≈ 0.2747
Desviación Estándar:	1.11843429984452 ≈ 1.1184
Varianza:	1.2508952830687 ≈ 1.2509
Coeficiente de Variación:	5.34186047579401 ≈ 5.3419
Primer Cuartil:	−0.2869
Segundo Cuartil:	0.2747
Tercer Cuartil:	0.8848
Cuarto Cuartil:	2.3800
Mínimo:	−2.9223
Máximo:	2.3800
Intervalo de Confianza para la Media:	[−0.1786770 ; 0.5974203]
Intervalo de Confianza para la Varianza:	[1.259150 ; 1.301725]

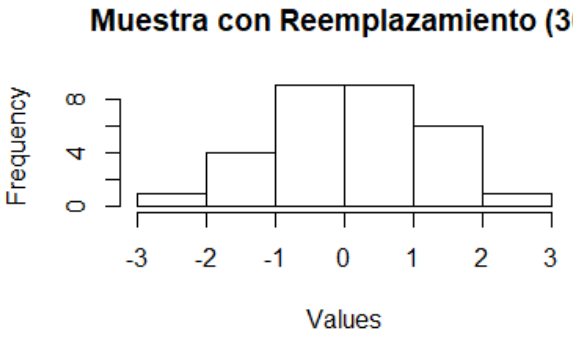

Figura 13: Histograma de X_3

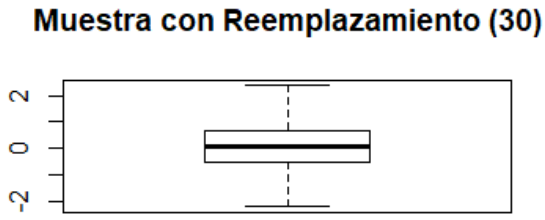

Figura 14: Gráfico de Cajas y Bigotes de X_3

Muestra (X_4) que representa a una muestra con reemplazamiento de tamaño 20

Media Aritmética:	$-0.184850607564658 \approx -0.1849$
Mediana:	$-0.300458810915177 \approx -0.3005$
Desviación Estándar:	$1.13996022831735 \approx 1.1400$
Varianza:	$1.29950932214534 \approx 1.2995$
Coeficiente de Variación:	$-6.16692713827628 \approx -6.1669$
Primer Cuartil:	-0.9524
Segundo Cuartil:	-0.3005
Tercer Cuartil:	0.1754
Cuarto Cuartil:	3.3242
Mínimo:	-1.9281
Máximo:	3.3242
Intervalo de Confianza para la Media:	$[\ -0.1287000\ ;\ 0.3853002\ ]$
Intervalo de Confianza para la Varianza:	$[\ 0.6746742\ ;\ 0.7031158\ ]$

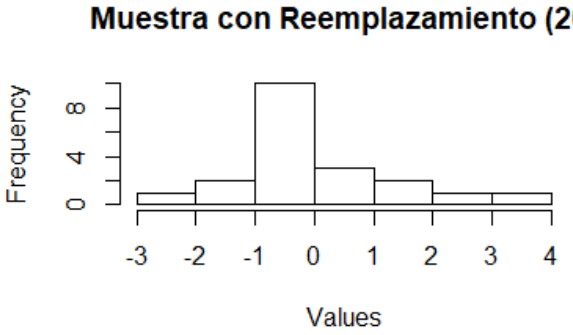

Figura 15: Histograma de X_4

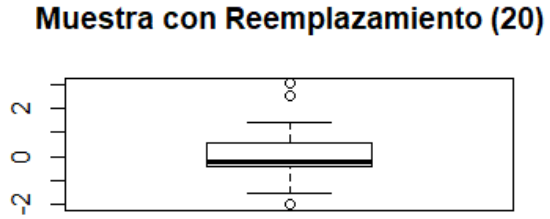

Figura 16: Gráfico de Cajas y Bigotes de X_4

5.2 Gráficos de las muestras sin reemplazamiento

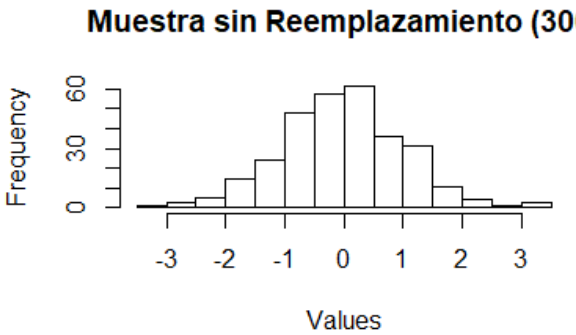

Figura 17: Histograma

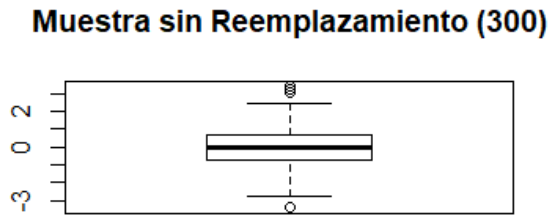

Figura 18: Gráfico de Cajas y Bigotes

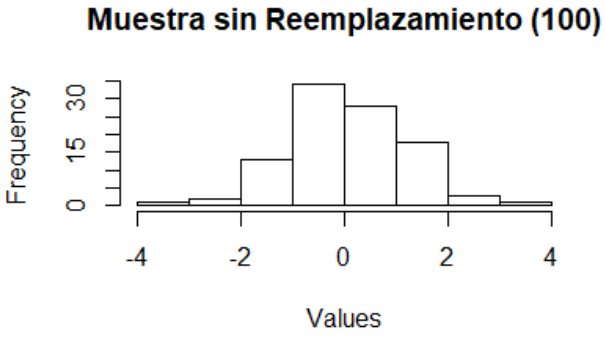

Figura 19: Histograma

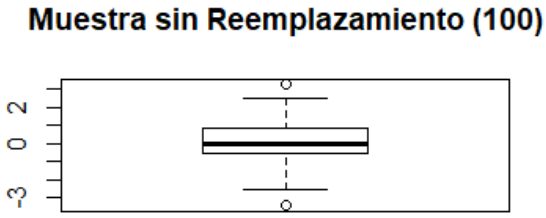

Figura 20: Gráfico de Cajas y Bigotes

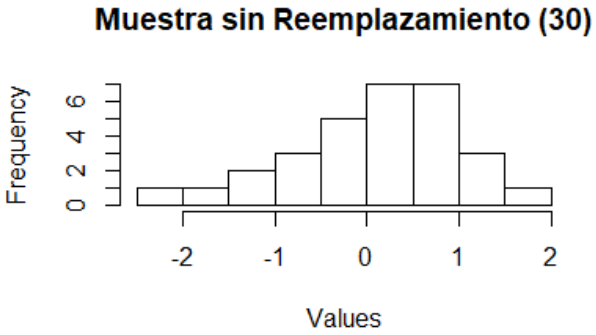

Figura 21: Histograma

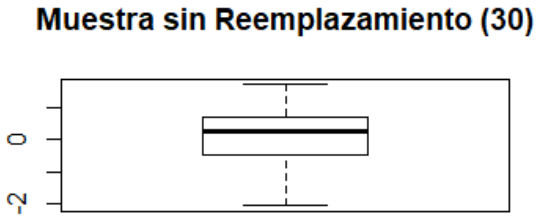

Figura 22: Gráfico de Cajas y Bigotes

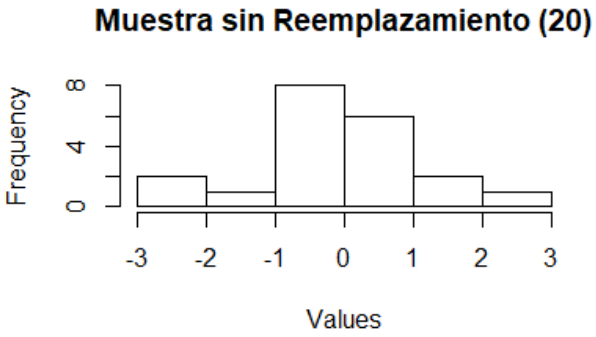

Figura 23: Histograma

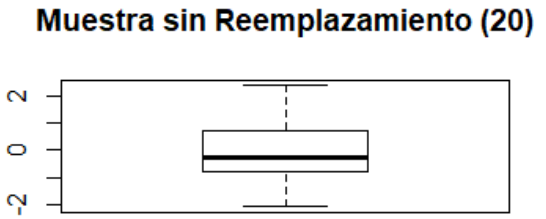

Figura 24: Gráfico de Cajas y Bigotes